Departamento de Biología Molecular
Facultad de Ciencias
Universidad Autónoma de Madrid

Emergent patterns in protein, microbial and mutualistic systems

TESIS DOCTORAL

Alberto Pascual García

Madrid, Mayo de 2015

La presente edición de este documento ha sido revisada por Silvia Pascual García (*xxilv@hotmail.com*). La maquetación se ha realizado con TEXIS v.1.0+, proyecto desarrollado por Marco Antonio y Pedro Pablo Gómez-Martín (*http://gaia.fdi.ucm.es/research/texis*). La cubierta ha sido diseñada por Alfonso Núñez Salgado (*alfonso.n.s@gmail.com*).

El autor agradece a todos ellos sus aportaciones.

Departamento de Biología Molecular
Facultad de Ciencias
Universidad Autónoma de Madrid

Emergent patterns in protein, microbial and mutualistic systems

Memoria presentada por:
Alberto Pascual García
para optar al grado de Doctor en Ciencias por la Universidad Autónoma de Madrid.
Director: Dr. Ugo Bastolla
Tutor: Prof. José María Carazo.

Madrid, Mayo de 2015

Este trabajo ha sido realizado en el laboratorio del Dr. Ugo Bastolla en el Centro de Biología Molecular "Severo Ochoa"(CSIC-UAM) en Madrid (España), en el laboratorio del Prof. Julián Echave en la Universidad Nacional de San Martín de Buenos Aires (Argentina) y en el laboratorio del Prof. Andrés Moya en el Centro Superior de Investigación de Salud Pública de Valencia (España).

A mis padres

A la memoria de Ángel R. Ortiz,
fundador de la Unidad de Bioinformática del CBMSO
y codirector de esta tesis en sus estadíos iniciales.

Summary

In this thesis we analyse emergent patterns in complex biological systems. We say that these patterns emerge, given that they result from behaviours of the system that are difficult to explain starting from a microscopic description. These behaviours are strongly dependent on the interactions between elements, and thus our research focuses on the identification and evaluation of interaction networks. In particular, we have analysed interactions that may reflect the response of the system to long term conditions, whose analysis may be compatible with an evolutionary interpretation.

The methodological and conceptual framework needed for the development of our research is complex. This is the reason why the first part of the thesis is devoted to clarify the epistemological approximation we have followed. In subsequent chapters, we present our research results, which have been developed around three systems with notable differences among them.

The first system considers a representative subset of all the protein structures known up to date. We develop a method that objectively demonstrates the existence of structural protein classes known as *folds*, defining conserved interaction patterns between amino-acids. We go deeper into the evolutionary interpretation of this result investigating the role of protein function in the structural conservation and divergence.

Second, we analyse high-throughput sequencing experiments collecting the presence of bacterial taxa in different environments. From this data we infer aggregation and segregation patterns suggesting that bacterial mutualistic interactions are very relevant, and whose functional role is explored in more detail analysing the bacterial assembly process in a group of infants during their development.

Last, we have considered mutualistic communities of plants and pollinators. We predict the structural stability of this system defining two magnitudes: the effective interspecific competition and the propagation of perturbations. These magnitudes rationalize the relative effect of competition *versus* mutualism –and, in particular, of the different mutualistic networks– in the structural stability, which we show has a main role for sustaining biodiversity.

Resumen

En esta tesis analizamos patrones emergentes en sistemas biológicos complejos. Estos patrones los calificamos como emergentes porque son el resultado de comportamientos del sistema difíciles de caracterizar partiendo de una descripción microscópica. Dichos comportamientos son fuertemente dependientes de las interacciones entre elementos, por lo que nos centramos en la identificación y evaluación de redes de interacción. En particular, hemos analizado interacciones que esperamos que reflejen la respuesta del sistema a condiciones relevantes en escalas de tiempo largas, cuyo análisis puede ser compatible con una interpretación evolutiva.

El marco metodológico y conceptual necesario para el desarrollo de nuestra investigación es complejo. Por ello, la primera parte de la tesis está orientada a clarificar la aproximación epistemológica que hemos seguido. En los siguientes capítulos presentamos el resultado de nuestra investigación, desarrollada alrededor de tres sistemas con notables diferencias entre ellos.

El primer sistema considera un conjunto representativo de todas las estructuras de proteínas conocidas hasta la fecha. Desarrollamos un método que demuestra objetivamente la existencia de clases estructurales de proteínas conocidas como *folds*, que definen patrones de interacción entre aminoácidos. Profundizamos en la interpretación evolutiva del resultado investigando el rol de la función de proteínas en la conservación o divergencia estructural.

En segunda lugar analizamos experimentos de secuenciación masiva que recogen la presencia de taxones bacterianos en distintos ambientes. De estos datos inferimos patrones de agregación y segregación que sugieren que las interacciones mutualistas entre bacterias son muy relevantes, y cuyo rol funcional es explorado en más detalle analizando el proceso de ensamblaje bacteriano en un grupo de bebés durante su desarrollo.

Por último, hemos considerado comunidades mutualistas de plantas y polinizadores. Predecimos la estabilidad estructural de este sistema definiendo dos magnitudes: la competición efectiva interespecífica y la propagación de las perturbaciones. Estas magnitudes permiten racionalizar el efecto relativo de la competición *versus* el mutualismo –y, en particular, de las distintas redes mutualistas– en la estabilidad estructural, cuyo papel mostramos que es esencial en el sostenimiento de la biodiversidad.

Outline of the thesis

This thesis is presented as a *compendium* of articles that investigate with computational methods three, apparently, very different biological systems, namely natural protein structures, bacterial communities, and mutualistic ecosystems of plants and pollinators. The commonalities between these different fields are discussed in a fourth line of research, devoted to the epistemological basis of the modellization of complex biological systems.

In this way, each section in which the thesis is structured, namely Introduction, Objectives, Results, Discussion and Conclusions, refers to all four lines of research.

The Introduction is divided in three parts. First, we present a preliminary part devoted to a general introduction of concepts and methods arising in the modelling of complex biological systems. In this part, the examples shown are not necessarily related to the particular systems that we have investigated. In the second part we specifically introduce the main background of the articles that are presented in the Results block. Finally, we close the Introduction by describing the publications and manuscripts developed during this thesis.

The Results section is divided according to the four lines of research presented. Each line summarizes the problems addressed in the corresponding articles included in the thesis, and may be viewed as an extension of the Introduction.

Whereas in most of the thesis the different systems are discussed separately, in the Discussion block we discuss similarities and differences between the works presented in this thesis from a methodological and epistemological point of view.

Finally, both the Objectives and Conclusions sections are subdivided into points devoted to each of the research lines.

Contents

Part I

Introduction

The soul of wit may become the very body of untruth. However elegant and memorable, brevity can never, in the nature of things, do justice to all the facts of a complex situation. On such a theme one can be brief only by omission and simplification. Omission and simplification help us to understand – but help us, in many cases, to understand the wrong thing; for our comprehension may be only of the abbreviator's neatly formulated notions, not of the vast, ramifying reality from which these notions have been so arbitrarily abstracted.

Aldous Huxley

Modelization: general aspects

Complex biological systems

The emergence and evolution of life is a question that has always fascinated mankind. Nature continuously presents us new forms and processes challenging our understanding. Furthermore, nowadays we live an exciting scientific moment where the development of new experimental settings gives us access to more detailed features of living systems. These new methods provide information never thought before, such as the direct measurement of kynetic properties in single molecule experiments, or the genetic sampling of unculturable bacteria in arbitrary environments, provided by next generation sequencing.

The availability of new data also allows for the development of new models, thus attracting the attention of scientists coming from other disciplines, as it was predicted by Schrödinger [Schrödinger (1992)]. There is growing interest in the development of mathematical and computational models aiming at a more formal description of biology. Such a description is desirable, because it facilitates the generation of models that may provide new predictions, thus reducing the number of experiments needed for hypothesis testing, while building a more solid knowledge structure upon accepted models.

Nevertheless, there are also new difficulties to be faced. Data obtained from new experimental setups require the investment of important efforts to identify possible sources of innacuracy or biases, that may be very relevant in the interpretation of results. In addition, the large amount of new data generated not always goes hand in hand with the development of new methods for efficient processing and storing, an important problem nowadays.

All these new data allow depicting a more comprehensive view of biological systems, but their integration into suitable models represent another important challenge arising from the high complexity of biological phenomena. In biological systems we observe a large number of entities interacting non linearly, which are further organized in levels going over different spatio-temporal scales. This kind of systems are difficult to model from the very first steps, where we need to determine which is our system, until the more advanced stages of research, where we aim to reproduce –the many times

complex– collective behaviours arising from their dynamical performance.

In addition, in living systems we observe the coexistence of (at least) two different spatio-temporal scales. On the one hand, the scale where the physico-chemical processes required to maintain the system out of equilibrium take place, –hereafter the physical scale–. On the other hand, the scale where the evolutionary events become fixed in the population –thus changing the system itself–, what we will call the evolutionary scale. This is why even if we have an accurate characterization of the system's behaviour in the physical scale, we do not reach a comprehensive understanding of the observed phenomena until we also explain the evolutionary process that has led to these observations and further provide predictions. This is the reason why it is claimed that *nothing in biology makes sense except in the light of evolution* [Dobzhansky (1973)]. In summary, even if we deal with a large and detailed amount of new data, building models of complex biological systems appear as a difficult challenge.

In this thesis we develop mathematical and computational methods to analyse patterns observed in complex biological systems. Given the difficulties highlighted, in the following section we briefly introduce some general considerations around modelling in these systems, such as the separation between microscopic and macroscopic descriptions or a more precise definition of emergent patterns. Next, we introduce the modelling approximation that has been explored in the different works. Briefly, this proposal focuses on the role of interactions between the entities of the system, and in particular in the identification of evolutionary conserved interactions. We finish this section introducing three particular systems around which our research has been developed.

Microscopic description

As we anticipated, in this thesis we have investigated some properties observed in complex biological systems. Before explaining which are the properties we are interested in, we start discussing how these systems are characterized.

The characterization of complex systems typically starts building a microscopic description. This means that we associate to each component o a set of variables $\{x^a\}$ that are sufficient to determine its state: $o = \{x^a\}$ (see Fig. [1] (a), a figure which we will systematically refer to along this introduction). It is important to emphasize that microscopic does not mean atomic, but it rather refers to the largest subdivision of the system that we consider in our description. For instance, a component may be an amino-acid if the system is a protein, a cell if the system is a tissue, or a car if the system is a traffic network.

The microscopic characterization is already a difficult task because it

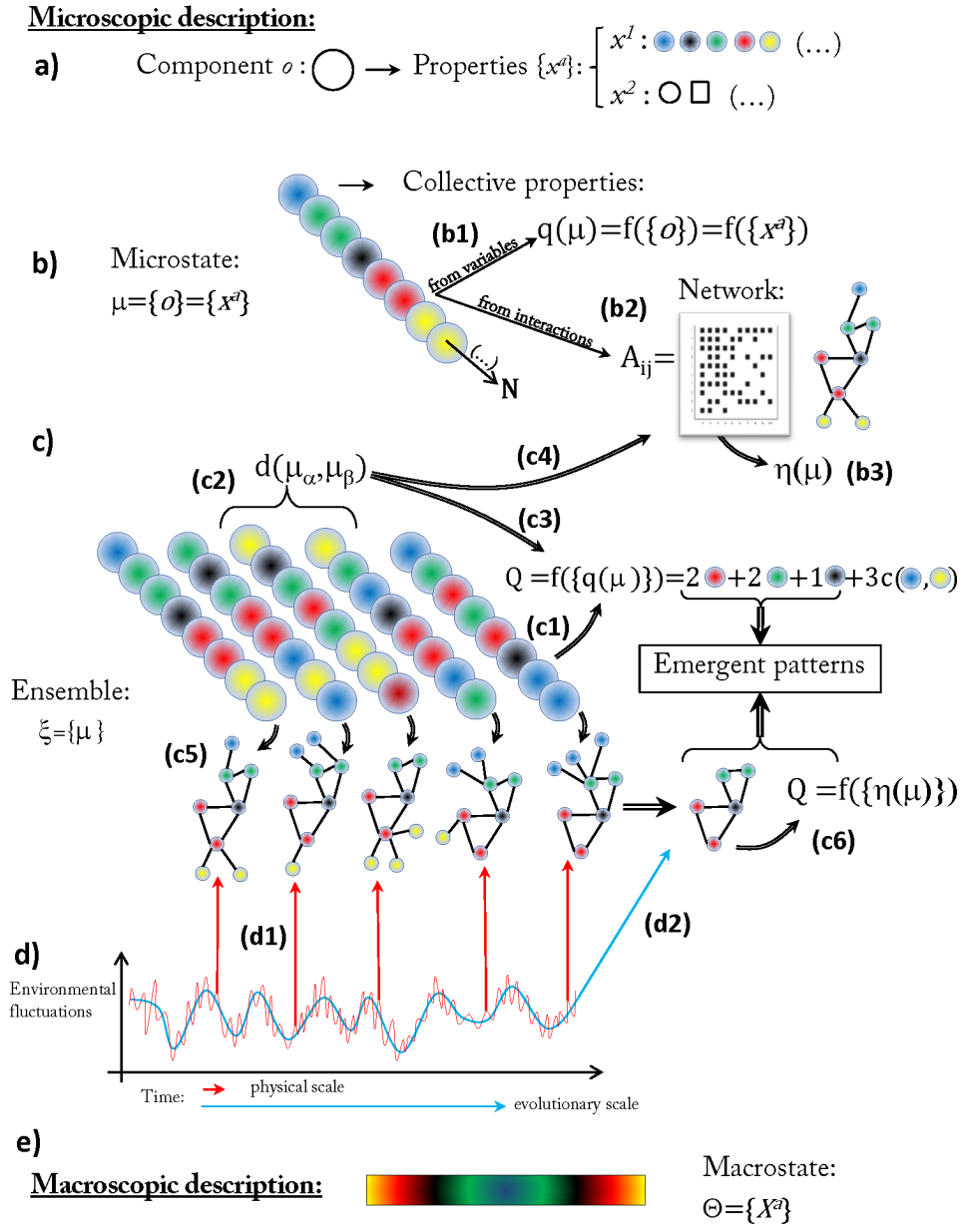


Figure 1: Scheme of the modelling process. The different steps are labelled from the microscopic description (a) towards the macroscopic description (e), and the different substeps are further labelled with numbers. See main text for details.

determines the system itself, and the system definition is a subtle question in the analysis of biological systems –although not exclusive of these sytems, see for instance [Georgescu-Roegen (1971)]–.

The reasons for these difficulties are, first, that living beings are open systems that exchange energy and matter with the environment. Thus it is not easy to individuate the boundaries of the system when there is not a clear physical separation with the environment –such as a membrane in the cells– [Maturana et al. (1975)].

In addition, what we consider the environment of the system –containing both biotic and abiotic entities–, is sensitive to the system's behaviour. Thus, even if the physical boundaries of the system are well established, when the environment shows a relevant response according to the system's dynamics, this may suggest that our description of the system should be expanded to explicitly include some parts of the environment.

Finally, it is not possible to incorporate in the model the large number of entities and processes typically present in the system, thus requiring to select those relevant for our research interests. It is unavoidable to face an epistemological challenge where elements belonging either to the environment or to the system are implicitly incorporated in the model or neglected. This is a critical task that should be addressed with caution, trying to avoid any arbitrary choice that could artificially provide spurious results while maintaining the necessary components to reproduce the observed behaviour.

Collective behaviour and macroscopic description

As soon as we characterize the microscopic components, we aim to look for a characterization of the system as a whole. We will call microstate μ each realization of the microscopic description of the whole system in terms of particular values of the variables $\{\tilde{x}_i^a\}$ at a given time t, i.e. $\mu = \{\tilde{x}_i^a\}$ (Fig. [1] (b)). The whole set of microstates that the system can visit is what we call the phase space of the system $\Omega = \{\mu\}$, and we will call to any subset of the phase space an *ensemble* of microstates $\xi = \{\mu\} \subseteq \Omega$ (Fig. [1] (c)). As we will see, the analysis of those ensembles of microstates visited when we observe a behaviour that we consider interesting is an important task to build a suitable model.

Collective behaviours and the associated patterns can be many times characterized at a macroscopic scale. A macroscopic description differs from the microscopic description in the precision of our observations (see Fig. [1] (e)). The macroscopic description is built with macroscopic variables X^a having a lower resolution than the microscopic variables. Frequently we do not require a knowledge of the values of the microscopic variables for their characterization. For instance, we can observe the behaviour of a hurricane, without knowing in detail all the positions and momenta of the particles

constituting this pattern. But if we have an incomplete (statistical) microscopic description of the system –modelled with a probability distribution of the phase space $P(\mu)$–, we can quantify macroscopic variables X^a by averaging the correspondent microscopic variable weigthed by this distribution, $< x^a(\mu)P(\mu) >$. The number of microscopic components should be large enough to achieve a macroscopic characterization, given that the error in its determination scales as $1/\sqrt{N}$, being N the number of components. Once we have determined the macroscopic variables, we will call macrostate Θ each macroscopic description of the system in terms of particular values $\{\tilde{X}^a\}$ of the macroscopic variables $\Theta = \{\tilde{X}^a\}$. Each macrostate is observed when the system significantly visits certain ensemble of microstates and, even if the microscopic variables change, the macroscopic values remain constant. In this way, focusing on the analysis of ensembles of microstates we may understand collective behaviours and, in turn, the associated macroscopic properties.

To finish this section, we note that what is considered a microstate and a macrostate may change if the scale of observation also changes. This is an important question for us because, as we said, the physical and the evolutionary scales coexist in the systems we are interested on. And it may happen that we focus on the description of a system which macroscopic description in the physical scale becomes a microscopic description on the evolutionary scale.

For instance, if we consider a population of individuals, the whole population would constitute a macroscopic description that we will model with a microscopic description of the dynamics of the individuals. Each configuration of individuals will be seen as a microstate in the physical scale –in some descriptions is rather each genealogy of individuals [Demetrius (2013)]– . But if we now aim to model some property concerning the dynamics of the whole population on an evolutionary scale, we will probably need to neglect detailed individual information, considering species instead of individuals. In this case, each snapshot of the genetic pool of population species would constitute a microstate in an evolutionary phase space.

As we will see, the examples we discuss in this thesis pertain to this class of systems, where the patterns over which we focus may have an evolutionary role. We will analyse the systems on the evolutionary space and we will attempt to understand which role they play in the system's performance within the physical scale.

Emergent properties and emergent behaviour

A basic tenet in scientific method is that macroscopic properties can be described starting from microscopic descriptions. For instance, even if we deal with an incomplete (statistical) microscopic description, the field of

Statistical Physics has shown considerable success explaining a wide variety of macroscopic phenomena.

Nevertheless, complex systems are constituted by a myriad of entities interacting non-linearly. And, apart from the difficulties in their characterization above mentioned, they frequently depict behaviours that are not reducible to the analysis of their components in isolation [Minati et al. (2006)]. This means that, when we analyse the properties of these behaviours, we may not be able to establish a linkage between some of them and the microscopic description.

Among this kind of collective behaviours, we find phenomena such as magnetism, patterns observed in dissipative systems like hurricanes or convenction cells or, in biological systems, patterns on animal skins or flocking behaviour. The apparent discontinuity between this particular kind of macroscopic properties and the microscopic description has led to coin the adjective *emergent* for these behaviours. Moreover, we will say the associated patterns are *emergent patterns*.

Emergent collective behaviours are surrounded by an aura of mysticism probably because they seem to be contrary to the above tenet, we highlight here some reasons. First, these behaviours lead to properties that refer to the whole, not to the components. Although this is not an exception for emergent behaviours, it has been popularly highlighted as a particularity claiming, for instance, *that the whole is more than the sum of the parts*. Second, these behaviours are in some sense unexpected. We will see below that we will work with this idea to identify significant patterns in our research. And third (and in our view the most important feature of these properties), given the nature of complex systems (in terms of number of components and non linear interactions) these behaviours are difficult to explain starting from a microscopic description. In this sense, we say that the collective pattern (behaviour), is hardly *traceable* from the microscopic description of the components.

Although it is difficult to explain these properties, it does not mean that it is not possible to provide an explanation for the emergence of these patterns, but rather that it is a difficult task. Indeed, once the existence of these properties is recognized, the attention is shifted towards the kind of explanations that these properties admit, i.e. in their epistemological accessibility.

In this sense, we adhere to the notion of emergence provided by Bedau, that has been called *weak emergence* [Bedau (1997)]. He argues that this term should be coined for those patterns that can be reproduced with a computational model –which is used as a proxy of the epistemological accessibility–.

Nevertheless, we would like to highlight another interesting notion that has been considered fundamental as opposed to epistemological, which is called *strong emergence* [Bar-Yam (2004)]. In Physics it is accepted that

knowing the positions and velocities of particles is sufficient to determine the pairwise interactions. This assumption is frequently found in Physics-inspired models of collective behaviour, where individual motion results from averaging responses to each neighbour considered separately. Nevertheless, Bar-Yam argues that this assertion would not hold if the system is embedded in responsive media, such as the motions of impurities embedded in a solid, or further in any process where global optimization (instead of local) is involved.

In this way, if there is a constraint in the system acting on all the components simultaneously and it is strong enough –i.e. it is a global constraint–, it is not possible to determine the state of the system considering only pairwise interactions. In some sense, the parts are determined *downward* from the state of the whole.

We believe that this idea is suggestive in our context, because biological systems are continously interacting with the environment. Although we do not expect that, in general, the influence be so strong as to determine the state of the system –otherwise, we could reconsider our system's definition–, it certainly exists. Thus, following this reasoning we may expect that external constraints affect the system globally, and this may be reflected in the global organization of the system.

Moreover, this may be particularly interesting in evolutionary processes where we will find fluctuations present in larger time scales, because these are global constraints that will probably shape the organizational long term outcome of the system. In particular, the concept of stability is widely discussed in this thesis, and it is intringuing whether the optimization of the stability of the whole, such as a population of species, is compatible with a notion of selection acting *only* over individuals locally.

Indeed, a subtle question readily arises here. As we said, a macroscopic description in the physical scale may become a microstate in a description at an evolutionary scale. Consider for instance that, after analysing the stabilty of certain ecosystem in the physical scale, we observe that some particular configurations are observed in the evolutionary scale, thus improving the stability performance against evolutionary changes. Under this scenario, which is the macroscopic emergent pattern on the evolutionary scale? The simple answer is *life itself* –which is not a trivial answer, given that it accounts for important questions such as the determinants of biodiversity or the distribution of the biomass–.

Nevertheless, we may aim to find a more specific answer, and this task typically requires to invoke any notion of biological function and its evolutionary role [Corning (2002)]. However, this kind of arguments are not exent of controversy, because the definition of function necessarily requires to specify the object of selection with respect to which *correct* function is defined. In this sense, in the above example one may wonder whether an ecosystem should be then considered an object of selection. Given that this question has

no obvious answer and that we do not make any kind of *a priori* assumption about the system's function in the development of our work, we will leave this question for further discussion at the end of the thesis.

Searching for collective properties: The role of interactions

A strategy to characterize an emergent behaviour starts from the analysis of the associated ensemble of microstates. When we observe an emergent behaviour, there are internal or external forces acting on the system. These forces limit the number of microstates actually visited, and thus the observed phase space Ω^{obs} is a subset of the phase space observed if they were absent, i.e. $\Omega^{obs} \subset \Omega^{free}$. We say that the system's collective behaviour is *constrained* by these forces, and we will use constraint and force equally hereafter.

These constraints lead to the restriction (even loss) of some degrees of freedom in the system, thus biasing (vanishing) some values in the observed system. Focusing on the ensemble of observed microstates $\xi^{obs} = \{\mu^{obs}\}$, we may look for a collective property $q = f(\mu)$ (see Fig. [1] (b1)) such that it is found in every observed microstate $\{q \ / \ q \gg 0; \ \forall \mu \in (\xi = \Omega^{obs})\}$. Furthermore, if such a property is found, we can define a distribution of q that characterizes the whole ensemble $Q = f(\{q(\mu)\})$ (Fig. [1] (c1)). For instance, if we analyse flocking behaviour, a component in this system would be a single bird and a microstate would be a snapshot in the flight of the flock. If we look for a property q found in every microstate, we may find a certain distribution in the angular positions between birds, which is characteristic when flocking behaviour is observed [Bialek et al. (2012)]. Another example may arise if we measure the wavelength in the sand dunes of a desert, and we see that it follows a certain distribution every time we observe ripples.

Nevertheless, the computation of wavelengths when ripples are observed seems to be a natural choice, but there are other behaviours where it is not so clear which is the measure we should define to characterize the microstates. In these cases, the comparison of microstates is a common procedure to extract common properties.

For instance, if we deal with an ensemble of protein sequences $\{\mu\}$, we can look for the pairwise similarity $S = align(\mu_\alpha, \mu_\beta)$ with a protein alignment algorithm such as BLAST [Altschul et al. (1990)]. From these alignments, we may find common amino-acids that are conserved in every sequence, which are a consequence of the evolutionary behaviour (see Fig. [1] (c2 and c3) and Fig. [2] for an specific example with protein sequences).

Other example arises if we consider that a microstate $\{\mu\}$ now reports the abundances of different bacterial taxa in a given environment. Considering an ensemble of microstates, we can compare the relative co-occurrence of taxa

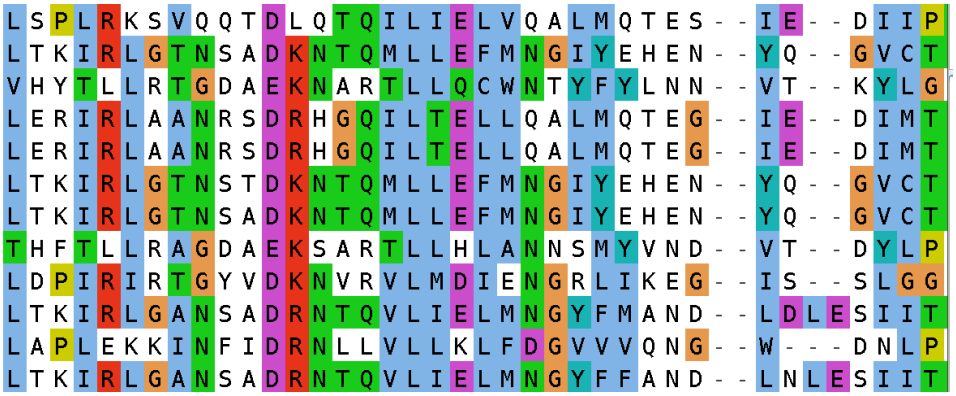


Figure 2: Illustration of a multiple protein sequence alignment. These algorithms attempt to maximize the number of common positions in the set, reflecting the existence of constraints to conserve certain amino-acids in these positions. Different properties of the amino-acids are highlighted: hydrophobic residues are in blue, polar in green, basic in red, acidic in violet and Glycines (G) and Prolines (P) are further differentiated. With this information, we aim to understand which physical properties are favoured by natural selection on this set.

with a correlation function $corr(\mu_\alpha, \mu_\beta)$ that computes the interdependence between the variables (the taxa) among microstates (environments).

Irrespective on whether we compute a similarity measure, a distance, or if we look for correlations, what we attempt to identify is any bias in the values of the variables for the observed microstates. This bias may be the result of external or internal constraints, and in this thesis we will focus on those restrictions that may be compatible with the existence of interactions between the components.

The rational behind this choice is that, when there is an interaction, there is a transmission of information –in its more general sense, for instance an element of biomass or a synaptic transmission– that may influence the value of one or several variables characterizing the state of the components involved in the interaction. This *systematic* influence is what leads to the observed bias.

A significant similarity between two protein secuences may be viewed as an evolutionary interaction where the transmission of information took place when the gene was duplicated. And a positive correlation in the abundances of two bacterial taxa, may be viewed as a positive (synergistic) interaction (see Fig. [1] (c4)).

In the examples shown above, we invested our effort in the search of putative interactions. Nevertheless, in relatively recent years there have been

increasing attention in the experimental characterization of systems *directly* in terms of their interactions. The representation of the interactions between a component o_i and a component o_j is made with a matrix A_{ij}, typically called *adjacency matrix*, whose elements describe the strength and sign of the interactions (see Fig. [1] (b2)).

This approximation is the one followed in the application of Complex Networks Theory for the analysis of complex systems [Boccaletti et al. (2006)]. Under this representation, microstates are described as networks, where the components are the nodes –which may be further characterized by their variables and values– and the links provide information about their interactions.

This representation is powerful because it already provides, in every single microstate, the information we obtain from the analysis of the microscopic properties with the whole ensemble if we do not know the underlying interactions. But, in addition, the development of the Complex Network Theory has led to the proposal of different properties η that can be measured on networks, $\eta(\mu)$ (given that a microstate μ is now represented with a network), providing a more refined representation of the constrains acting on the fluxes of information of the system (see Fig. [1] (b3)).

In this way, if we deal with a network for every microstate (see Fig. [1] (c5)), we may wonder which network properties are found in the whole ensemble, what may confirm that we are actually dealing with certain constraints influencing the collective behaviour (see Fig. [1] (c6)). This finding will facilitate the development of mechanistic models aiming to simulate the observed behaviour, as we will discuss below.

The role of evolution

Under the above description, if we characterize a system only considering its interactions, the phase space of the system would consist of the ensemble of all possible configurations of the interactions –configurations that are also called topologies of the network–. Nevertheless, when the system is large, the number of possible networks is huge –indeed, the number of total configurations scales as 2^{N^2}, being N the number of components–, and an exhaustive exploration of the whole phase space is unfeasible.

A strategy to reduce this search, consists of searching interactions that are conserved in the evolutionary space, reflecting the existence of environmental or evolutionary constraints present in evolutionary time scales (see Fig. [1] (e)). In some sense, these interactions would represent a scaffold (Fig. [1] (e2)) around which other configurations arising from processes taking place in shorter time scales would be found (Fig. [1] (e1)).

To identify these interactions, we focus again on the comparison of different systems –that may be found in very different environments–, sharing

any evolutionary relationship. Such comparison allows us to reduce a large number of possible configurations to a small number. Particularly relevant are global network properties such as the connectance or the assortativity [Boccaletti et al. (2006)] that may reflect the presence of global constraints and, as we said when strong emergence was introduced, may have interesting interpretations from an evolutionary perspective.

For instance, following the example of flocking behaviour, we would consider different snapshots of the flight of different species in their respective habitats. We would then build an appropriate representation of their interactions –for instance, a representation based on the visual capability of the birds with respect to their neighbours–. And we would finally look for common interaction patterns irrespective of the species and habitat. If we find a common pattern in the interactions, it may reflect a flight mechanism evolutionary selected, whereas differences on the patterns would rather reflect specific features in the flight of each species, adaptations to its particular environment or fluctuations occuring during the experiments, among other reasons.

Addressing significance

Once we focus on the kind of strategy we will follow, we wonder how we manage the data in order to look for properties that may reflect the effects of interactions, and how we address the significance of these patterns.

When we discussed emergent properties, we said that the emergent behaviour was a consequence of the existence of constraints, and in particular of interactions between components. If we characterize the observed data with a distribution $P^{obs}(\mu)$, what we aim to understand is which distribution should be expected if the constraints investigated were absent, $P^{free}(\mu)$. The behaviour of the (more) *free* system provides a baseline for our expectancy on the viable values of the microscopic variables. Indeed, the microstates belonging to the observed ensemble $\mu^{obs} = \{\mu \ / \ \mu \in \xi = \Omega^{obs}\}$, typically have a low probability of being observed if these forces are absent, $P^{free}(\mu^{obs}) << 1$, which explains its interest. Therefore, when the constraints act on the system, the expected behaviour of the variables becomes biased, leading to a new –more ordered– behaviour, and eventually to a *pattern* formation.

In this way, if we have identified any property in every observed microstate –which may be built either from the observation of the variables q or of the networks η–, that we believe is informative of the existence of an unknown constraint, we should demonstrate that the probability of observing values above certain threshold c fullfils $P^{obs}(q) \gg P^{free}(q)$, with $q > c$.

Therefore, addressing the significance of any explanatory argument pointing towards the existence of constraints, requires to build the reference distribution P^{free}. In the following, we briefly explain two different procedures

to build this distribution, depending on the nature of our data.

Pattern-based (null) models

When we have little information about the system, we must limit ourselves to perform a statistical analysis. A powerful approximation arises from pattern-based analysis (also called null modelling [Gotelli and Ulrich (2012)]). In our context, given that we are focusing on the existence of interactions, these kind of models typically contain at least two features. First, it is assumed that the components are independent, i.e. they do not interact. Second, we should explicitly include in the model every condition affecting the system that may explain the observed pattern, i.e. any known constraint. For instance, if we observe that an increase (decrease) in the growth of a given bacteria systematically corresponds with an increase (decrease) of other bacterial taxa, we may think that there exists an ecological interaction. But there is a simpler explanation, namely that there exists any environmental condition that simultaneously benefits both species. If this is the case, we should include this condition as a constraint in the model.

Therefore, the procedure we consider when we deal with pattern-based models reads as follows (see the scheme found in Fig. [3]). We start with some experimental data (on the scheme, the presence of species in different environments Fig. [3] (a)), over which we aim to compute a measure (b) that may reflect the existence of interactions, such as a similarity in the environmental preferences (c). Considering known constraints (in the example, the presence of two different environments (d)), we build a probabilistic model (e) considering that the species do not interact (in the example, a generalized linear model (GLM) that will be explained in the results section). From this model, random realizations of the pattern observed are generated conditioned to the existence of these constraints (f). The fact that we aim to explain the observed pattern considering a random process under given constraints, constitutes a null hypothesis that should be rejected to accept the possibility that other processes –such as ecological interactions–, may have any role in the observed outcome. This is the reason why we call this kind of models null models.

As soon as we obtain random realizations, we can compute the same measure over the random ensemble (see Fig. [3] (g)), obtaining the probability that any specific value of the measure is obtained in the null model (h). If the observed value is significantly high (i), we can reject the hypothesis that the observation is due to the constraints present in the null model, accepting the possibility that other constraints (such as interactions) are the main drivers of the observed pattern.

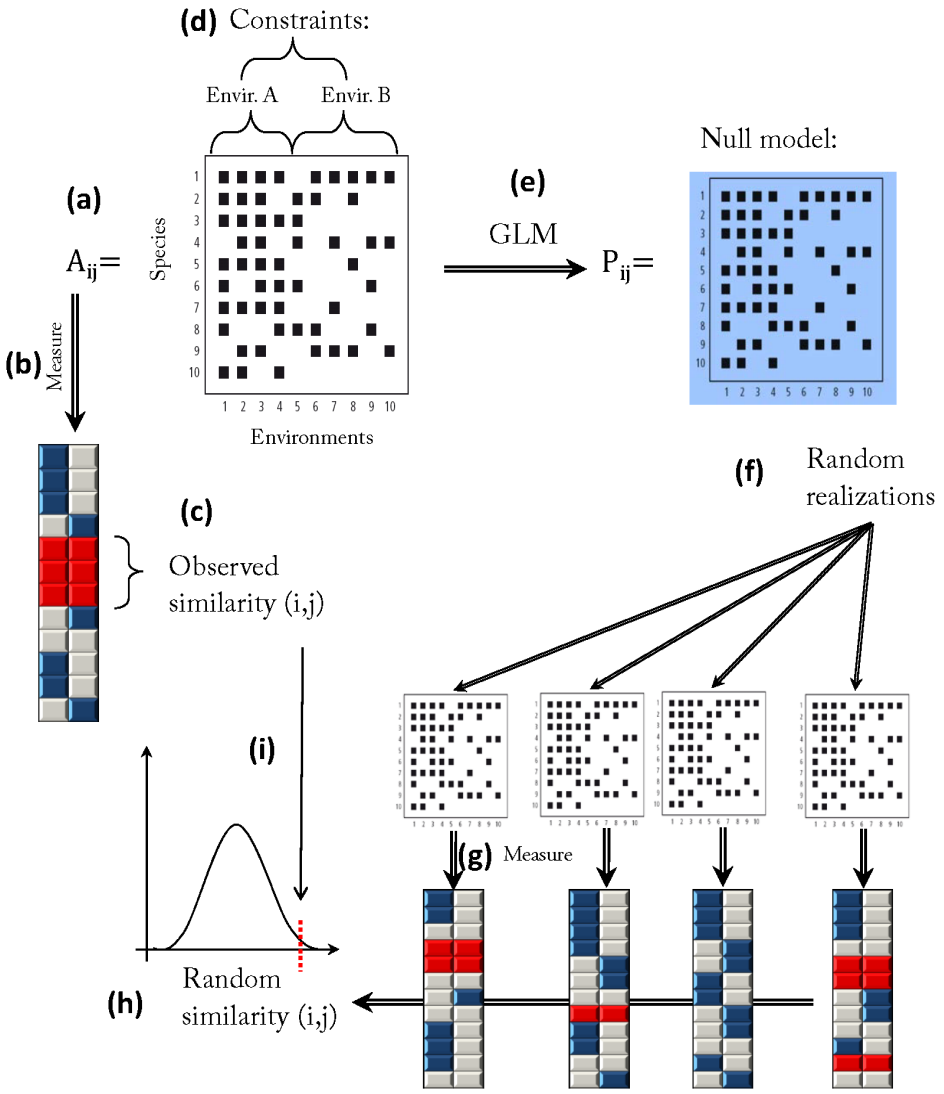


Figure 3: Scheme of the modelling process in pattern-based (null) models (see main text for details).

Mechanistic models: explicit simulation

The second approximation that we consider arises when we have enough information about the system to build a mechanistic model, meaning that we have any hint on the mechanism by which the variables of the different components are modified. In Physics, this is the case for instance when we know the Hamiltonian of the system in Classical or Quantum Mechanics

In biological systems, we are still far from reaching well founded (commonly acepted) mathematical models such as the one provided by Statistical Mechanics in Physics, to explain macroscopic phenomena from microscopic *first principles* (although there are interesting approximations, see for instance [Sella and Hirsh (2005); Capitán and Cuesta (2011); Demetrius (2013)]).

But there are also well known approximations to macroscopic observations that led to a considerable amount of theoretical development. For instance, in Ecology, we find stochastic models, such as neutral models [Gotelli and McGill (2006)], and deterministic models, such as the classical Lotka-Volterra models [Volterra (1926)]. And, in Molecular Biology, techniques such as molecular dynamics are widely used [Karplus et al. (1990)], where both classical and quantic computations are implemented.

The idea under the mechanistic approximation is to model a dynamical behaviour of the system that leads to the observed pattern, considering the minimum amount of assumptions. For instance, a useful –although oversimplified– classification of ecological interactions is built considering the effect that a given interaction between a pair of individuals has on each of them. In this classification, only three effects are considered, either positive $(+)$, negative $(-)$, or neutral (0). In this way, we will say that the interaction is mutualistic if the effect is positive for both individuals, competitive if it is negative for both, and so on (see Fig. [4]).

This simplified picture can be further refined incorporating functional responses from the observation of the system, leading to more realistic models. For instance, mutualistic interactions may lead to divergences in the abundances of the species, but this effect can be seen as unrealistic if there is a saturation in the effect that these interactions have on each specie involved. If it is the case, we can circumvent this problem including a non linear term on the model reflecting saturation effects.

Irrespective of the sophistication of the model, we will follow a similar approximation to the one followed for pattern-based models. We aim to individuate which of the components of the model is more relevant to explain the observations (again, in our case, we focus on the role of interactions). In this way, once we propose a model reaching a reasonable fit with respect to the observed behaviour (i.e. we reproduce the behaviour within some error ϵ^{mod}), we should test whether our hypothesis, namely, the relevance of any component over others, holds.

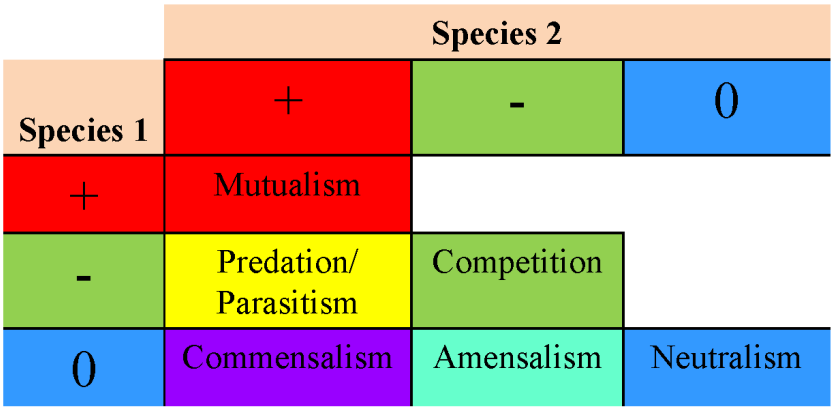


Figure 4: Simplified classification of ecological interactions regarding the effect that the interaction has on each specie, which may be positive (+), negative (−), or neutral (0).

To assess this question, we *intervene* on the model, what means that we build another model (that could be considered again a null model) where the component under analysis is absent, and we compare whether the error of the null model ϵ^{null} is significantly larger than the alternative model, thus $\epsilon^{null} \gg \epsilon^{mod}$. This kind of procedure highlighting the importance of the intervention [Boschetti (2011)] to measure the causal effect of a component is in agreement with the paradigm of Granger causality [Seth (2010)].

For instance, an example that will be discussed in this thesis is the importance of mutualistic interactions to sustain biodiversity in ecosystems. The null model in this case would be a system where mutualistic interactions are absent (i.e. there are only competitive interactions), and we will test whether this system sustains a larger biodiversity than the alternative model, which contains mutualistic interactions.

Another example may be whether a particular configuration of mutualistic interactions (parametrized by any network measure η) is also positively related to biodiversity. A null model will consider that the property η is not significant –where significance can be assessed with a pattern-based model discussed in the previous section–, and again we compare the ability of the null model to sustain biodiversity with respect to the alternative model, where η is significant. If we observe that the performance of the alternative model is significantly better than the one shown by the null model, we can accept the possibility that η has a relevant role in the observation of a higher biodiversity.

In summary, in this thesis we have followed the approximations provided by both pattern-based and mechanistic models to analyse different emergent

patterns. The experiments proposed within these frameworks have been designed for being entirely compatible with the scientific method: we propose hypothesis based on the analysis of experimental data and we reject the hypothesis with numerical experiments, what allows us to identify causal relationships. Nevertheless, and as we will discuss in more detail at the end of the thesis, the power of the approximations followed must be evaluated on the basis of the ability of the different models for further providing *predictions* that can be verified with independent experiments.

On the following section we will explain how the different notions we have introduced for the modelization of complex biological systems are materialized in particular examples.

Introduction to specific systems

Following, we will introduce four sections around which the results presented are organized (see Fig. [5]). The first section is devoted to epistemological questions related with the modelization of complex biological systems. Although this work is still in progress, we find appropriate to introduce it in the first place. The reason is that it has been developed on parallel to this thesis and it follows a similar structure and contents. In this way, the introduction and results of the manuscript could be considered a formal extension of the present introduction, where some of the concepts discussed –such as the system determination, the classification schemes or the definition of emergence– are further clarified.

After this general epistemological introduction, we present the analysis of three biological systems with notable differences between them, around which we identify emergent patterns. In this introduction, we present a brief summary around the research we developed in each of these systems, and we will focus on the differences and similarities between them in the discussion section.

The first system deals with the set of all known protein structures. The second considers the composition of bacterial taxa observed in different environmental samples, and the third will focus on matrices describing the interactions in mutualistic ecosystems of plants and animals (either pollinators or seed dispersals).

Epistemology of complex biological systems

As we anticipated, dealing with complex biological systems first requires to clarify different notions concerning the way in which we approximate to the knowledge of these systems, and how the information acquired is incorporated into a suitable model. Indeed, there is an increasing interest for epistemological questions underlying areas of knowledge such as Systems Biology [Mazzocchi (2008); Regenmortel (2004)], whose deficiencies have led to strong criticism arising from authoritative voices –see for instance the in-

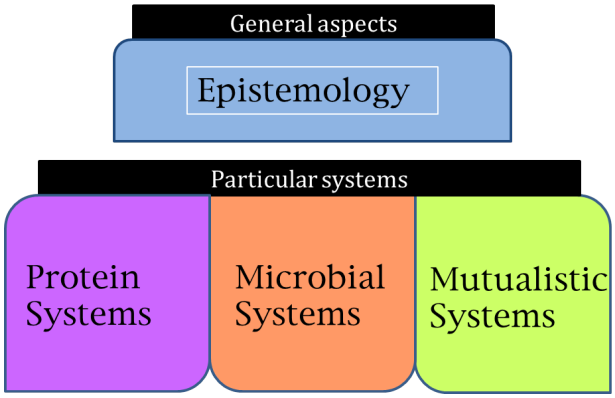


Figure 5: Organization of sections in the thesis

terview to Brenner in the Spanish journal cicNetwork (Num. 11; May 2012), entitled *Systems Biology is a loss of time*–. These questions have been analysed and clarified in the first manuscript we introduce [Pascual-García (2015)], and in a preliminary work [Pascual-García (2012)].

In this work, we wonder if there is any generic procedure to build the conceptual setting needed to work with complex systems. We reasoned that, if the way we acquire new knowledge from these systems –i.e. our epistemological approach– is similar irrespective of the system analysed, we may be able to identify the causes that may generate the appearence of typical problems found in the investigation of complex systems.

The first question that we address is the definition of system, where we particularly focus on the difficulties in the determination of system boundaries. As we anticipated in the previous introduction, the determination of a microsocpic description is a difficult task in biological systems, because we deal with open systems and responsive media. This problem can be related with the determination of classification schemes, which is a cornerstone in the development of any science [Bohm (1971)], and it has been classically a matter of controversy in Biology in particular [Dougherty and Braga-Neto (2006)].

Next, we explore the definition of emergence, emergent property and emergent behaviour. As we already pointed out in the introduction to modelling techniques, the main source of ambiguity when we talk of emergence seems to arise from the difficult characterization of an emergent behaviour from microscopic properties. Therefore, if we characterize an ensemble of microstates with a standard microscopic conceptual setting –from which we build up new concepts that refer to the whole ensemble–, we should be able to detect any particular feature in the final conceptual structure such that we can affirm that a collective property has emerged. In particular, we will look

for differences between weak and strong emergence, focusing on the number and scope of the constraints present in the system. We are aim to explore the positioning of some scientists who state that strong emergence is not epistemologically accessible. If this is the case, we should observe differences between emergent patterns accounting for this inaccessibility.

To finish, if the analysis is indeed generic from an epistemological perspective, we discuss whether is possible to derive an epistemological program, and which additional difficulties we face when we work with biological systems, given that we deal with both evolutionary and physical processes.

Protein systems

The first system we modeled in this thesis corresponds to a representative set of all protein structure domains known up to date. Protein structures are determined through either experiments, with X-ray diffraction techniques or nuclear magnetic resonance spectroscopy (NMR), or they are predicted with computational methods. All these structures are publicly available at the Protein Data Bank [Berman et al. (2000)], and they contain around 10^5 entries up to date. NMR provides an ensemble of structures, thus being appropriate for studies interested in a dynamical view of the protein structure. On the other hand, X-ray techniques achieve a higher resolution, where a a single structure is solved. In our work, we only considered proteins experimentally solved with the latter method. In addition, a structural domain is colloquially defined as the minimum unit in which a protein structure may be divided into, such that it can autonomously evolve and function. It is noteworthy to say that the automatic determination of structural domains is an unsolved problem, because there is not a consensus defintion [Holland et al. (2006)].

The first microscopic emergent observation in this system arises after observing that there exist some structures more similar to others and that it seems to be possible to classify them into well defined clusters. Note that the number of possible amino-acid sequences is 20^N, being N the length of a generic sequence, and that the number of known sequences has an order of 10^6 (around 130 orders of magnitude lower for $N = 100$). Such a huge reduction reflects the constraints that natural selection imposes in the thermodynamical and kynetic requirements of protein structures for fast and stable foldability and proper function. The fact that we can still reduce our representation of proteins to few thousand clusters is interesting, because it provides further insights into the interplay between the evolutionary process and the physical and biological features selected (see Fig. [6]).

One of the pioneriing observations of this pattern was made analysing 31 structural domains [Chothia and Michael (1976)] that were classified in four classes, regarding their content in secondary structures. Secondary structures

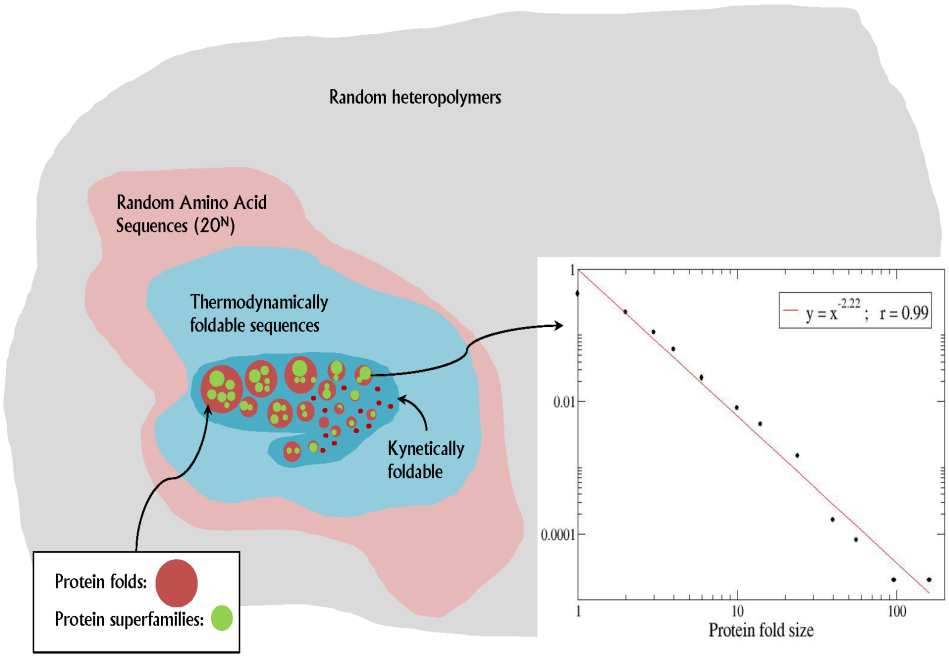


Figure 6: Schematic representation of proteins in nature embedded in the ensemble of random heteropolymers (left). The areas are not representative of the marked differences in the orders of magnitude between groups. Thermodynamic and kynetic requirements heavily reduce the number of sequences compatible with biological function. Within the area representing natural proteins, we further illustrate the existence of protein folds and protein superfamilies, whose distribution in size follows a power law (right).

are three dimensional forms of local segments in protein domains mainly determined by hydrogen bonds, and are classified as $\alpha-$helices and β sheets.

To get some insight into the strength of the constraints that hydrophobicity impose on protein structures, it is useful to observe a representation in the so-called Ramachandran plots (see Fig. [7]). In this kind of plots, dihedral angles Φ and Ψ of the backbone are represented. If interactions were absent, we would expect to find arbitrary values in the plot, whereas we observe that the points are constrained in some regions, which reflect the existence of secondary structures.

In this way, the classification in classes proposed by Chothia and Michael was: i) (All α) having only $\alpha-$helix secondary structure; ii) (All *beta*) having mainly β sheets; iii) $(\alpha + \beta)$ which are proteins containing both types of secondary structures, but fragments of each type tend to appear segregated from those of the other type, and iv) (α/β) containing also both types, but in

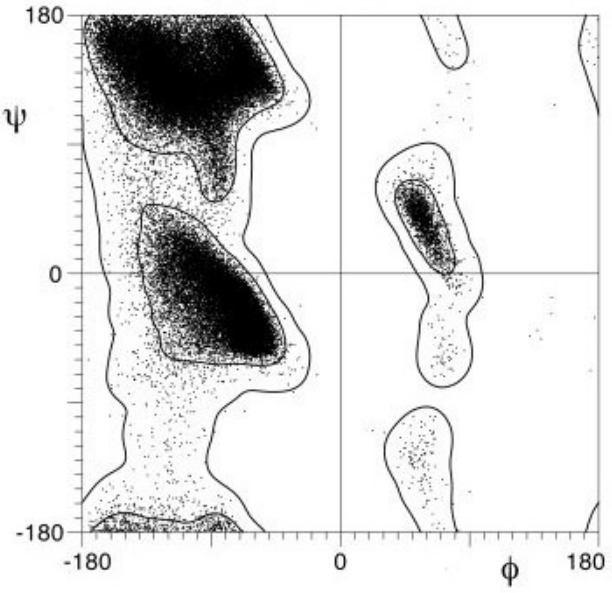


Figure 7: Ramachandran plot for the residues of 500 high resolution structures, excluding Glycines and Prolines. The populated upper left region describes residues located in β sheets, whereas those in the middle correspond to $\alpha-$helices. Reproduced with permission from [Lovell et al. (2003)].

this case they are found mixed, sometimes alternating, along the structure.

The growth in the number of entries motivated the development of protein classification databases with manual (SCOP [Murzin et al. (1995)]), semi automatic (CATH [Orengo et al. (1997)]), or fully automatic (FSSP [Holm and Sander (1997)]) classification procedures. The increasing number of entries made insufficient the starting classification scheme in classes, that has been further subdivided into different hierarchical levels. The most relevant levels are those immediately below the class, namely the *fold* level (in SCOP, architecture in CATH) and the *superfamily* level –where folds are further splitted considering any evidence of homology–. In Fig. [8] we show an example of three domains classified in CATH within the Globins superfamily.

Two weaknesses in this scheme are readily identified. First, the very definition of fold is ambiguous. For instance, the canonical definition of fold provided in SCOP states that two protein domains belong to the same fold if they share *the same major number and direction of secondary structures with a same connectivity*. This definition is rather loose, because it still needs to specify which secondary structures are considered major and which are considered embellishments.

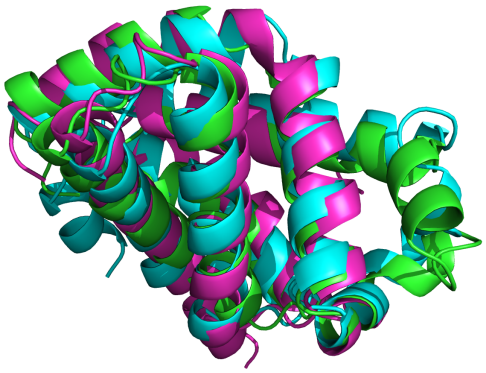


Figure 8: Example of structural domains belonging to the globins CATH superfamily. The PDB identifiers are $2yrsD$ in green, $7hbiA$ in cyan and $3vhbB$ in magenta. The structures are aligned with MAMMOTH multiple [Lupyan et al. (2005)] and represented with Pymol [DeLano (2002)].

Second, including an evolutionary definition at the superfamily level in hierarchy handicaps a purely structural definition of protein domains. Although structure is conserved in evolution very often, it is not always true, and thus including this information can generate inconsistencies. For instance, let us consider a protein which has converged during the evolutionary process towards certain function which is different from the function of its ancestor (see for instance [Bork et al. (1993)]). If we take into account its structural features, it may be similar to structures belonging to a fold whose members have no recognisable homology with the protein. And, on the other hand, it may be significantly dissimilar to proteins sharing recognisable homology. To classify a protein following the current hierarchical scheme implies that either this protein is classified in the fold where we observe structural similarities –and then it should be defined a superfamily with a single member–, or it is classified in a superfamily, where it has homology with their members but their structures are very different. None of these solutions is optimal, and this is the reason why an objective classification based on purely structural arguments is desiderable.

To address the viability of a purely structural classification, we should first wonder whether a classification scheme is justified. We observe that the existence of a global classification of protein structures is supported by at least two arguments. First, the molecular clock hypothesis [Bromham and Penny (2003)] recognized that, after gene duplication, protein sequences accumulate amino-acid substitutions almost linearly in time. Indeed, it has been observed the existence of protein gene families [Chothia (1992)] and further the existence of a characteristic distribution of clusters in these fa-

milies [Koonin et al. (2002)]. Second, it was shown in a seminal paper by Chothia and Lesk [Chothia and Lesk (1986)], that protein structures also diverge linearly in time, and thus it may be expected that the space of protein structure similarites ressembles the properties found in the space of protein sequence similarities.

Nevertheless, there is another observation pointing towards the impossibility of establishing a protein classification scheme. It has been found that there exists a significant number of local similarities between proteins without any recognizable homology. This observation is also justified considering other evolutionary events apart from gene duplications and substitutions, such as gene insertions or deletions. These events, although less frequent [Lupas et al. (2001)], may have dramatic effects on protein structures [Grishin (2001)], increasing the number of local similarities shared between proteins belonging to different putative folds. If this kind of local similarities were pervasive in the space of structures, any attempt of classification would become frustrated, because it may be impossible to objectively determine a similarity threshold over which well-defined clusters are found. This possibility would lead to a scenario where we can visit the whole space of protein structures *jumping* from protein to protein following local similarities [Skolnick et al. (2009)].

From a practical perspective, the importance of this question relies on the determination of protein function. On the one hand, protein structure facilitates the exploration of protein function, as it provides a more detailed picture of the biochemical and dynamical features of the protein, allowing for the application of computational methods which explicitly incorporate explicitly these properties, such as molecular dynamics. On the other hand, the number of protein structures known is almost two orders of magnitude lower than the number of sequences. Therefore, investigating the function of a given sequence using its structure requires, most of the times, solving a computational model of the structure. As we said, a useful strategy to explore the space of conformations of microstates (in this case, the space of all structural conformations given a sequence) is unfeasible, and evolutionary arguments can help us to reduce the number of conformations to be evaluated. Thus, a successful approach for modeling protein structures consists on the search of homologs whose structure is already solved, and then using these proteins as templates –together with an appropriate energy function– to build the model. In this sense, understanding whether the protein structure space is classifiable –and, in general, the search of structural similarities— allow us to understand the structural properties relevant for the inference of protein function, and to delineate new modelling approximations for structural prediction.

To address the problem of classification, we have followed the modelling approximation we explained in the introduction. First, we consider a

representation where the protein structure is modelled as a system of N components, in this case amino-acids, that are interacting physically, what it is called a contact. In practice, these contacts are determined considering a threshold in the distance between the amino-acids, below which it is considered that an interaction exists. Second, and as we will explain in more detail in the results section, we perform a comparison between protein structures in order to reveal those interactions that are conserved during the evolutionary process, thus reflecting different constraints acting on the evolution of structures.

These constraints may reflect necessary requirements for fast folding, thermodynamic stability, appropriate function performance, or mutational robustness, among others. However, we expect different effects on the structure from the different evolutionary forces acting on the protein. For example, whereas selection for thermodynamic stability may be expected to affect globally the protein structure, the specificity for protein function may be selected over a much smaller structural region, such as an active site.

Therefore, after computing the structural similarities, we aim to explore if it is possible to objectively determine a similarity threshold where protein structures are classified in the basis of either global similarity, local similarity, or if it should be considered continuous. This information, together with the information provided by protein sequences and protein function, may help us to understand which are the main constraints acting on the evolution of protein structures

Microbial systems

The next system we analyse consists on data from communities of microbes, found in different environments, obtained from experiments in which the genetic material found in these samples is analysed. These data have been obtained through Next Generation Sequencing (NGS) techniques, that allow to identify the genetic content in samples without the need of growing them, and quicker and cheaper than tradictional sequencing methods.

The number of NGS techniques and applications is diverse, and we will not enter into detail here, but we will present some notions about this type of data. The general idea underlying these techniques to increase its speed is that of parallel sequencing. The genes found in these samples are amplified in short pieces which are partially or totally sequenced (called reads 35-400bp length), allowing for the fast generation of a large amount of data. However, there are some shortcomings in the data obtained from these techniques that must be taken into account.

For instance, sequencing a high number of fragments is necessary because it intends to capture all the variability found among members of the same specie. In this sense, their accuracy depends on the number of reads

of the experiment, or in biological issues such as the existence of homopolymer sequences, that are the subject of frequent errors in some sequencing platforms.

Other sources of bias or innacuracies in the interpretation of results, arise for instance from primers choice –a strand of nucleic acid that serves as a starting point for DNA synthesis– which may generate biases in the diversity found when primers for specific taxa are used. Another drawback is that a definition of bacterial species is very difficult [Cohan (2002)], and an operative definition of species is invoked instead (Operational Taxonomic Units (OTUs)). This definition is typically made on the basis of the 16S rRNA gene, which is highly conserved among species but, at the same time, contains hypervariable regions that are characteristic for each specie. Nevertheless, the actual NGS technologies do not allow to sequence these regions together at once in one read for each DNA molecule from each individual, so it may happen that we deal with OTUs with different gene composition despite having similar 16S rRNA. This fact may affect the interpretation of results from an ecological perspective, given that members of the same OTU may have notable differences at a metabolic level.

Despite these difficulties, NGS allows for the identification of unculturable microbial taxa, revealing a microbial diversity much higher than the one already known with culturable strategies, opening new opportunities for the analysis of ecological questions in the bacterial world.

The mechanisms shaping species diversity have been classically a matter of intense research, being an important question the particular role of ecological interactions. A critical concept relating biodiversity and interactions was the competitive exclusion principle, stated for the first time by Gause in 1934 [Gause (1934)]. The competitive exclusion principle was briefly restated in four words by Garret Hardin in 1960 as: *complete competitors cannot coexist* [Hardin (1960)]. With this definition, Hardin emphasizes that exclusion is a strict consequence of a condition whose ambiguity is clearly expressed by the adjective *complete*. As requirements for completeness are relaxed in simplified model scenarios, the importance of competition can be grossly overestimated. A step forward to reach a more comprehensive view of biodiversity was given by Hutchinson in his seminal paper *Homage to Santa Rosalia* [Hutchinson (1959)]. The multidimensional definition of niche proposed by Hutchinson was a starting point to comprehend the extent of the competitive exclusion principle, whose complexity still pervades modern Ecology.

Hutchinson's question is probably one of the most interesting challenges in the microbial world. Microorganisms represent an amount of biomass at least as big as that of plants and an amazing diversity, and they have a key role in the evolution of the biosphere. Microbial communities evolved jointly with multicellular eukaryotes for more than one billion years and this

coevolution has influenced, and probably shaped, the evolution of vertebrates [Ley et al. (2008)]. Therefore, understanding the ecological features underlying prokaryotic diversity can open new perspectives on animal evolution, and may have important biomedical applications.

Nowadays, an increasing number of available data coming from high-throughput experiments have focused ecologistséfforts in the search of ecological trends, and there is an increasing evidence pointing to a qualitative similar picture between the patterns found in macro and microorganism [Horner-Devine et al. (2004)]. Some progress have been made identifying important trends such as taxa-area and distance decay relationships [Green and Bohannan (2006)], or the influence of environmental and geographic variables as depth [Rocap et al. (2003)] or salinity [Crump et al. (2004)]. These progress have been possible in part thanks to the development of prokaryotic Biogeography [Ramette and Tiedje (2006)].

In classical Biogeography, departing from matrices where the presence or absence of species in different spatial locations are represented, one aims to infer the role of community ecology processes. High throughoutput genomic data open new opportunities to tackle this question for microbes, as geographically distant samples sharing similar physico-chemical properties can be analyzed together with species associations [Tamames et al. (2010)], which can help us distinguish between contemporary or historical mechanisms [Martiny et al. (2006)].

It is important at this point to remember the debate between deterministic and chance-based explanations of observed diversity distribution arising from the approximation provided by Biogeography [O'Malley (2007)]. This debate started after observations made by Diamond [Diamond (1975)], who reasoned that the different interactions taking place between members sharing the same niche should lead to patterns distinguishable to those obtained by chance. These patterns would be a consequence of an assembly process, and he proposed four rules to explain them. These ideas were vigorously attacked by Connor and Simberloff [Connor and Simberloff (1979)], who argued that these rules were either tautologies or untestable rules. This discussion leaded to apparently irreconcilable positions about the actual value of these patterns, an ironically summarized debate by Lewin with the sentence: *Santa Rosalia was a goat* [Lewin (1983)]. Even if some controversies still persist, this discussion boosted the theoretical development of null models.

However, in order to consider bacterial data, we must face new challenges that should be carefully considered both in the development of null models and in the interpretation of results. Appart from the sources of innacuracy pointed out for NGS data, evaluating all the possible pairwise interactions implies to consider $n(n-1)/2$ possible associations, being n the number of species. In data coming from genomic experiments, it is possible to handle thousands of species, which will lead to consider millions of putative interac-

tions with matrices typically very sparse, what may generate, for instance, strong biases if classical measures of ecological ressemblance are used and the effect of unseen species is not considered [Chao et al. (2006)].

Concerning the interpretation of results, there are many microbial ecological mechanisms still poorly understood. It is known that bacteria interchange genes through Horizontal Gene Transfer [Koonin et al. (2001)], and they are able to establish syntrophic relationships through metabolic coupling under certain conditions [Morris et al. (2013)]. There are also complex ecological processes, such as Quorum Sensing [Miller and Bassler (2001)], which is a cell-to-cell communication process regulating complex behaviors such as biofilm formation or antibiotic virulence. Whereas the classical view of the competitive exclusion principle predicts that closely related species tend to compete more strongly –given that they similarly exploit the resources in the environment– the bacterial processes we mentioned may lead to expect the opposite observation. The reason is that closely related species share molecular mechanisms that are more compatible for the exchange of genes and metabolites than divergent species.

Thus, new ecological perspectives are required. For example, one general mechanism for the establishment of such cooperative interactions between sister bacterial populations is the recently proposed Black Queen Hypothesis [Morris et al. (2012)]. According to this hypothesis, the selective loss of genes involved in costly leaky functions creates dependencies between subpopulations that retain and lose these genes. This model has been proposed as a general mechanism for the establishment of cooperative bacterial communities [Sachs and Hollowell (2012)], and its paradigmatic example is the evolution of gene gain and loss in the marine cyanobacterium Prochlorococcus [Kettler et al. (2007)].

In summary, in this thesis we aim to explore the diversity and distribution of species in high-throughput samples obtained from different environments. We consider that each sample is a single microstate in the evolutionary phase space, and its observation at first sight already reveals an uneven distribution of the taxa and their abundances. We want to explore whether there is any systematic observation arising when we consider a whole ensemble of microstates, namely whether there exists any non-trivial pattern in the distribution of taxa in a large set of samples. Another interesting observation over which we focus arises from the pervasive cosmopolitanism of microbes –i.e. the ability of some taxa for living in a wide variety of environments [Tamames et al. (2010)]–, which is hardly explained in terms of the competitive exclusion principle.

In the results section we will further explain the assumptions and technical development of the experiments.

Mutualistic systems

The last system we discuss in this thesis consists of populations of plants and animals (either pollinators or seed dispersals). In these systems there is interspecific competition between species present in the same pool (plants or animals), and mutualistic relationships between species belonging to different pools. The experimental data upon which we build our models are matrices describing the mutualistic interactions observed between plants and animals.

The ecological concepts presented in the previous section are also pertinent here, given that we are also interested in the analysis of the diversity of mutualistic communities and the role of ecological interactions. Nevertheless, in this system we have access to matrices describing the interactions of mutualistic species found in very diverse environments all over the world (please, visit *www.web-of-life.es*), which may be interpreted as an ensemble of microstates in the evolutionary space. In this way, we will explore whether it is possible to relate the configurations found in these matrices with the biodiversity of ecosystems.

Given that we can explicitly consider interactions, we can assess an important relation between the interactions in a given ecosystem and its biodiversity, which is established invoking stability arguments –and that has led historically to the stability-complexity debate [Pimm (1984); Haydon (1994); Ives and Carpenter (2007)]–. This relation arises from the influence that a particular configuration of the species interactions has on the ability of the system to maintain positive biomasses –i.e. to avoid species extinctions–, when it is affected by perturbations.

In Community Ecology there has been historically much interest on dynamical models such as Lotka-Volterra equations [Volterra (1926)]. In these models the variables are the biomass abundances of the species, and we are interested in understanding if there exists any stationary state for the abundances such that every specie has positive biomasses, as a function of sets of parameters such as the intrinsic growth rates and the type and strength of the ecological interactions. In other words, which are the model parameters that guarantee that all the species coexist.

Indeed, for these systems it is frequently assumed the existence of stable fixed points. A stable fixed point is a solution of a system of first order ordinary differential equations $x*$ whose main feature is that, if the system's values are fixed at that solution and we modify these values with a small enough perturbation, the system will return again to the same solution. If we can determine that a fixed point exists, we next wonder which is the magnitude of the perturbations that the system can assume such that the system still returns to the original fixed point. This kind of analysis is what is called dynamical stability analysis, and there are particular types of stabilities that can be defined. For instance, given a system modelled with a set

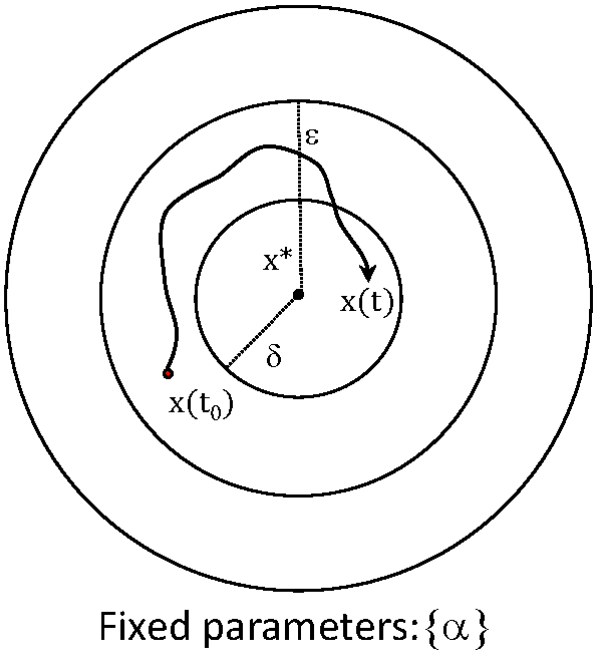


Figure 9: Schematic representation of Lyapunov stability. The fixed point x^* of the dynamical variables is said to be *Lyapunov stable* if, for fixed parameters $\{\alpha\}$, for every positive ϵ we can find a number δ such that every time that we perturb the dynamical variables at time t_0 by less than delta, $|x(t_0)-x^*| < \delta$, the evolution of the system remains close to x^*, $|x(t)-x^*| < \epsilon$ for $t > t_0$.

of parameters $\{\alpha\}$ whose variables at the fixed point are x^* (see Fig. [9]), we say that the system is *Lyapunov stable* in a region Ω_x ($x^* \in \Omega_x \subseteq \Omega$) if and only if [Justus (2008)]:

$$(\forall \epsilon > 0)(\exists \delta > 0) \ / \ |x(t_0) - x^*| < \delta \Rightarrow (\forall t \geq t_0)(|x(t) - x^*| < \epsilon)$$

If the perturbation that the solution can assume is small, we will say that the system is *locally stable*, and it is an interesting concept because small perturbations are analytically easy to deal with. On the other hand when variations can be arbitrarily large, we will say that the fixed point is *globally stable*.

The interest on this kind of models has been justified claiming that more stable systems should be observed more often [Haydon (1994)], and that their analysis would enable to disentangle the main determinants of natural selection, in turn allowing for the incorporation of more complex evolutionary

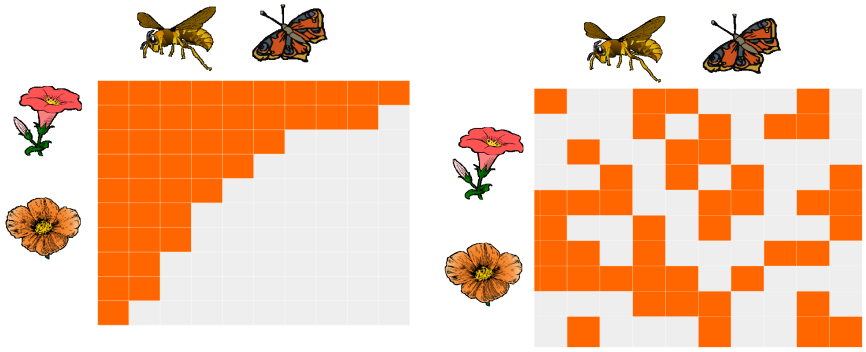


Figure 10: Illustrations of plant-pollinator interactions matrices for a completely nested pattern (left) and a random configuration (right).

processes [Vellend (2010)].

One of the most important controversies existing in the literature around the stability-complexity analysis has its origin on the different results found by MacArthur [MacArthur (1955)] and May [May (1972)]. On the one hand, MacArthur pointed out that a perturbation in the abundances of a given specie would be best corrected –thus diminishing the negative effects in the rest of the system– when the number of species and interactions increase. He argued that an increasingly large number of available paths for the fluxes of biomass in a trophic chain diminishes the probability to observe that any specie becomes isolated from resources when any path is affected by a perturbation. On the other hand, May proposed to assess the stability of a system computing the probability of finding a stable fixed point when the interaction parameters where randomly drawn, and reporting a negative trend when the number of species and the connectance of the network increased.

Following this latter line of reasoning, models of mutualism have been classically considered less stable. The reason is that non-saturating mutualistic interactions may lead to the divergence of the abundances of some species, and in turn to the extinction of others. Unfortunately, these difficulties and an oversimplified view of these –apparently negative– results have been reflected in a reduced interest for modelling mutualistic models in the literature (reviewed in [Pascual-García (2009)]).

Indeed, it was only after the discovery of a nested pattern of mutualistic interactions [Bascompte et al. (2003)], that renewed interest on mutualistic models was generated. In a nested pattern, more specialist species interact only with proper subsets of those species interacting with the more generalists, leading to a characteristic pattern illustrated in Fig. [10].

In this way –and similarly to the models that will be discussed in the results section of microbial ecology–, the earlier stages of research found in

literature around this pattern were focused on the development of pattern-based models. These models considered as a null hypothesis that the nested pattern could be obtained with a random process subject to some constraints, typically the conservation of row and column totals (exactly or on average) reflecting the species effects. The incorporation of more complex constraints in the generation of the matrices (such as correlations between rows and columns totals) aimed to evaluate the relevance of the nestedness pattern [Jonhson et al. (2013)]. However, these findings do not represent a qualitative advance unless the ecological meaning of the mathematical constraints incorporated in the generation of the matrices is also clarified.

Therefore, a more direct approach to the relevance of mutualistic interactions and their configurations arises from the development of mechanistic models. There are several possibilities to model these systems, being the classical Lotka-Volterra models a common choice. Models are built considering systems of equations where the abundances of species are the variables to be determined, and the growth rates and interactions parameters that must be fixed.

A step forward in our understanding of this kind of dynamical systems has been the focus in other kind of stability, the *structural stability*. This quest for structural stability, although common in other fields of computational biology is not so common in theoretical ecology, except for some recent exceptions [Bastolla et al. (2005, 2009); Rohr et al. (2014); Pascual-García and Bastolla (2015)]. In the analysis of structural stability, the focus is on the stability of the system with respect to changes in the parameters rather than changes in the dynamical variables. In the approach that we propose, stable fixed points are obtained as a starting point in the model, and then the model parameters are perturbed in order to see whether we still reach another stable fixed point compatible with the existence of positive biomasses, i.e. thus sustaining the observed biodiversity (see Fig. [11]).

This shift from dynamical stability to structural stability is justified thinking that demographic variations are detected through changes in the population values. But it is difficult to think in external processes directly modifying the abundances (examples may be the introduction of a disease or a sudden environmental fluctuation). Evolutionary or environmental changes would rather affect the parameters of the model –namely their growth rates or the strength of the interactions– and these changes affect, *in turn*, the abundances of the populations.

Therefore –and as we will explain in more detail in the results section–, in our work we developed a comprehensive analysis of a model of mutualistic ecosystems to explore the relative role of competitive *versus* mutualistic interactions in terms of structural stability. In addition, we will evaluate the effects that particular configurations of mutualistic interactions have on the stability and biodiversity of ecosystems.

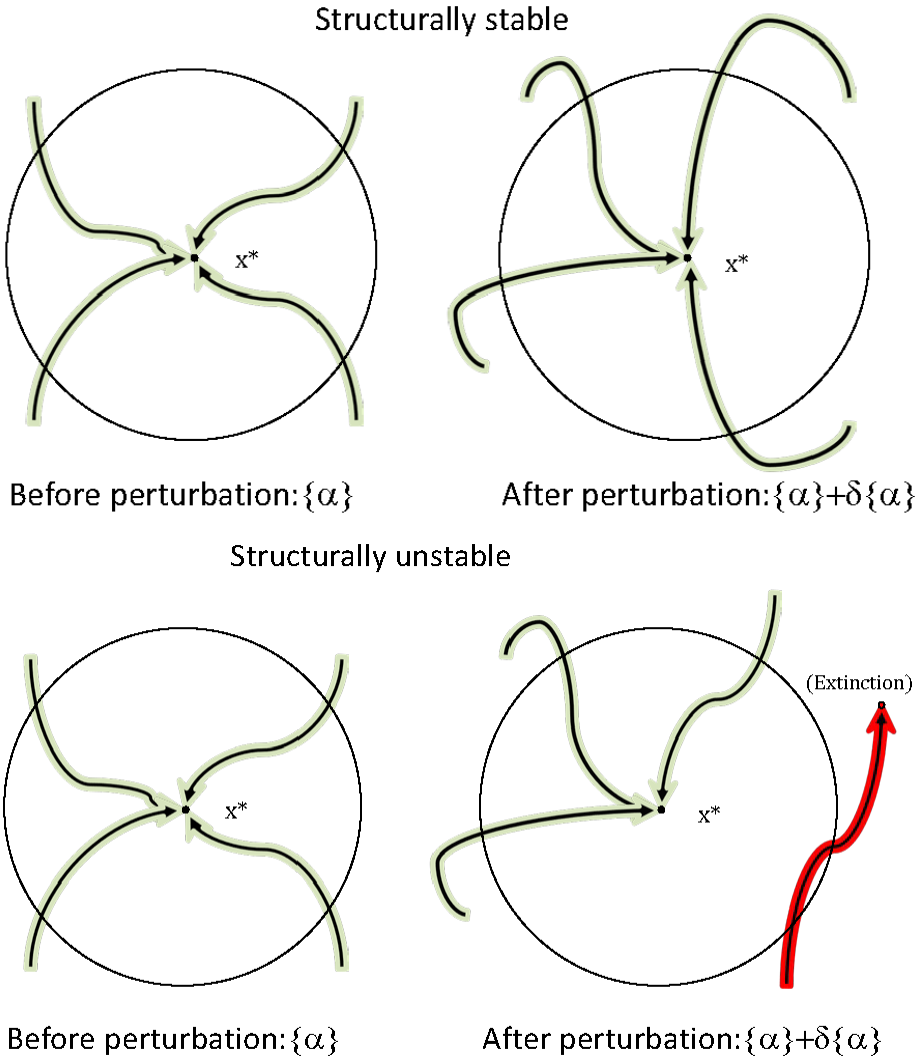


Figure 11: Illustrations of a structurally stable (above) and a structurally unstable (below) system. Starting from the fixed point x^*, the parameters are perturbed $\alpha \to \alpha + \delta\alpha$, and we observe whether we obtain an equivalent topology of the phase space –in the example, another fixed point where all the species have positive biomasses–. Below we observe that, after the perturbation, the system attains another fixed point which is not compatible with positive biomasses for all the species, and there is a specie that takes nonzero biomasses, thus becoming extincted.

Outline of articles

The work developed in this thesis has led to different research articles, book chapters, and a contribution to proceedings. These works are related with protein, microbial and mutualistic systems, and to epistemological questions. In the following results, we have selected the most representative works for the thread of this thesis.

In the following, we show a list with the articles splitted in blocks corresponding to the different results sections, and ordered chronologically within each section. We have considered one article in the epistemology section, and two for each particular system analysed. We add a short label for each article selected which is further highlighted in bold face. This label is also shown in the header of those pages where the articles are embedded, in this way the reader can easily identify to which section and article are referring the contents depicted.

Epistemology of complex biological systems

1. Pascual-García, A. On the epistemology of complex networks theory (2012). *AIFBI Proceedings*

2. **[EPIS-1]** Pascual-García, A. Epistemology of complex biological systems: insights into dimensionality reduction, constraints identification and emergence from a topological approach, (2015). *In preparation.*

Protein systems

1. **[PROT-1]** Pascual-García, A., Abia, D., Ortiz, Á. R. and Bastolla, U. Cross-over between discrete and continuous protein structure space: Insights into automatic classification and networks of protein structures, (2009). *PLOS Computational Biology, vol. 5(3), page e1000331. ISSN 1553-7358.*

2. **[PROT-2]** Pascual-García, A., Abia, D., Méndez, R., Nido, G. S. and Bastolla, U. Quantifying the evolutionary divergence of

protein structures: The role of function change and function conservation, (2010). *Proteins: Structure, Function, and Bioinformatics*, vol. 78(1), pp. 181-196, 2010. ISSN 1097-0134

3. SÁNCHEZ-NIDO, G., MÉNDEZ, R., PASCUAL-GARCÍA, A., ABIA, D. AND BASTOLLA, U. Protein disorder in the centrosome correlates with complexity in cell types number. (2012) *Molecular BioSystems*, 8(1), 353-367.

4. PASCUAL-GARCÍA, A. Alineamiento de estructura de proteínas (2014). In *Bioinformática con Ñ*.
Coor. Sebastián, A and Pascual-García, A. ISBN 978-84-617-1976-X.

5. PASCUAL-GARCÍA, A. Evolución de estructura de proteínas (2014). In *Bioinformática con Ñ*.
Coor. Sebastián, A and Pascual-García, A. ISBN 978-84-617-1976-X.

6. PASCUAL-GARCÍA, A. Modelos simplificados de plegamiento de estructura de proteínas (2014). In *Bioinformática con Ñ*.
Coor. Sebastián, A and Pascual-García, A. ISBN 978-84-617-1976-X.

7. SÁNCHEZ-NIDO, G., ROMANO, L., BASTOLLA, U., PASCUAL-GARCÍA, A. Structural bioinformatics within a snake puzzle. (2015). *In preparation.*

Microbial systems

1. **[MIC-1]** PASCUAL-GARCÍA, A., TAMAMES, J. AND BASTOLLA, U. Bacteria dialog with santa rosalia: Are aggregations of cosmopolitan bacteria mainly explained by habitat filtering or by ecological interactions? (2014) *BMC microbiology*, vol. 14(1), pp. 1-16.

2. **[MIC-2]** VALLÉS, Y., ARTACHO, A., PASCUAL-GARCÍA, A., FERRÚS, M. L., GOSALBES, M. J., ABELLÁN, J. J. AND FRANCINO, M. P. Microbial succession in the gut: directional trends of taxonomic and functional change in a birth cohort of Spanish infants. (2014) *PLoS genetics*, vol. 10(6), page e1004406.

Mutualistic systems

1. **[MUT-1]** BASTOLLA, U., FORTUNA, M., PASCUAL-GARCÍA, A., FERRERA, A., LUQUE, B. AND BASCOMPTE, J. The architecture of mutualistic networks minimizes competition and increases biodiversity. (2009) *Nature*, vol. 458(7241),pp. 1018-1020.

2. PASCUAL-GARCÍA, A. Explorando el rol de la Competición, el Mutualismo y la Arquitectura en Redes Ecológicas: Qué podemos decir sobre

la Biodiversidad? (2009). In *Evolución y Adaptación: 150 años después del origen de las especies.*
Coor. Dopazo, Hernán and Navarro, Arcadi. Ed. Sociedad Española de Biología Evolutiva, ISBN 978-84-92910-06-9

3. Pascual-García, A., Ferrera, A. and Bastolla, U. Does mutualism hinder biodiversity? (2014) *arXiv preprint* arXiv:1409.1683.

4. Pascual-García, A., Ferrera, A. and Bastolla, U. Effective competition determines the structural stability of model ecosystems. (2015) *Under revision*

5. Ferrera, A., Pascual-García, A., and Bastolla, U. Effective competition determines the global stability of model ecosystems. (2015) *Under revision*

6. **[MUT-2]** Pascual-García, A. and Bastolla, U. The complexity-stability relation of mutualistic systems reconciles MacArthur and May, (2015). *Under revision*

Part II

Objectives

We must walk consciously only part way toward our goal and then leap in the dark to our success.

Henry David Thoreau

Objectives

1. For protein systems, we aim to investigate whether the space of protein structures is classifiable. In other words, if there is an objective similarity threshold above which it is possible to define protein folds, i.e. equivalence classes of protein structures, and its interpretation in terms of evolutionary events.

2. We will also explore which is the relative role of protein function in the evolutionary conservation and divergence of protein structures, and its relation with protein sequence divergence.

3. For microbial systems, we aim to test whether it is possible to infer ecological interactions from high-throughput sequencing data, further exploring its relative role with respect to other selective processes such as habitat filtering, considering a large ensemble of samples from a wide diversity of environments.

4. In addition, we will attempt to provide a more specific functional role for the putative interactions, analysing the assemblage of bacteria in the gut of a group of infants in their early stage of life, comparing them with the communities found in their mothers.

5. For mutualistic systems of plants and pollinators, we aim to analyse the relative role of competitive versus mutualistic interactions in the structural stability of the model ecosystem and its relation with the ability of the system to support biodiversity.

6. A second objective for these systems will focus on the effect that different constraints in the configuration of mutualistic interactions have in the structural stability of the system.

7. The last objective is to incorporate the methods and concepts that have been applied in the study of the different systems within a common epistemological framework, in order to clarify some problematic concepts arising in the analysis of complex biological systems and to further delineate an appropriated modelling approach.

Part III

Results

Buenos días Alberto, he estado pensando (...)
es que tengo la impresión de que seguimos sin entenderlo del todo (...)
y creo que podríamos hacer un último experimento...

Ugo Bastolla

(...) Così tra questa
immensità s' annega il pensier mio:
e il naufragar m'è dolce in questo mare.

Giacomo Leopardi

Chapter 1

Epistemology of complex biological systems

Observation always involves theory.
Edwin Hubbel

Summary

In this chapter we present the results found for epistemological questions related with the analysis of experimental data and modelization of complex biological systems.

In this work, we proposed a novel approximation to these problems through the incorporation of a mathematical formalism [Sambin (2003)] that allows for reaching more clear definitions of fuzzy concepts such as the concept of emergence, whose different notions have been presented in the introduction. Our motivation to address this task has been that, during the development of the thesis, we have worked with such a kind of concepts, and thus we have felt that an epistemological clarification was required. In this way the results found –although obtained at the end of the thesis–, refer to the specific problems at which we worried. This fact, together with the more precise definitions obtained for the different concepts, are the reasons that induced us to present these results in a first place.

In brief, this epistemological formalism establishes a correspondence between objects of observation and concepts through a function called the extension of the concept, i.e. a function that given a concept returns the set of objects it refers to. It has been demonstrated that, through this map, the space of subsets of concepts induces a topology (in a set theoretical sense) in the space of objects [Boniolo and Valentini (2008)]. Therefore, the induced topology allows us to obtain the tools for dealing with concepts that are difficult to sharply define otherwise.

For instance, each concept has associated in the space of objects and open set. Therefore, we can look for the closure of this set and, the intersection between the closure and its complementary set determines the border of the set. If there are objects within this border, it means that the meaning of the starting concept extends somehow further than the limits of the open set it defines. This is the rational followed to say that the concept is vague and, interestingly, the formalism allows us to quantify the vagueness of the concept following this definition [Boniolo and Valentini (2008)].

For instance, if we determine the organizational systems that the adjective *democratic* maps on the basis of the concepts contained in its definition, some of these concepts will be present in other organizations that are not considered *fully* democratic. Thus this concept is not *sharply separated* from other concepts, containing an amount of vagueness that will lead to discrepancies. Similarly, we have used this definition of (extensional) vagueness to investigate some of the concepts we have discussed in this thesis.

First, we provided a careful definition of system compatible with the notation used in this topological formalism. From this starting point, we observe that both the determination of system boundaries and that of an optimal classification, are conceptually related problems. In both cases it is required to objectively determine a criterion to separate the system (or each cluster) from the environment (or all the other clusters). Difficulties on finding such a criterion are related with the existence of non-empty borders for the sets of objects associated to those concepts involved in the definition of the system (or the classification). We analyse in more detail the classification problem taking as an example the classification of protein structures, which is one of the main tasks developed in this thesis. The example analysed is very simple, and thus it facilitates the introduction to the articles discussed in the proteins chapter.

We have shown that the extension map between concepts and objects can be related with an important scientific activity: the identification of the information shared –and the information transmitted– between the components of a system. We have shown that this map stands on the definition of similarities and distances, which in turn are the basis of dimensionality reduction techniques.

Next we show that this procedure can be used to compare microstates in complex systems, being this the means by which we can decipher new patterns and constraints in the dynamics. Making use of a suitable toy model we are able to define concepts such as the traceability of the system –i.e. a minimum epistemic condition to consider that there is a correspondence between the microscopic and the macroscopic description–, or the definition of emergence itself. We provide a definition of emergence strength which is related with the information needed to describe the system relative to the constraints acting on it.

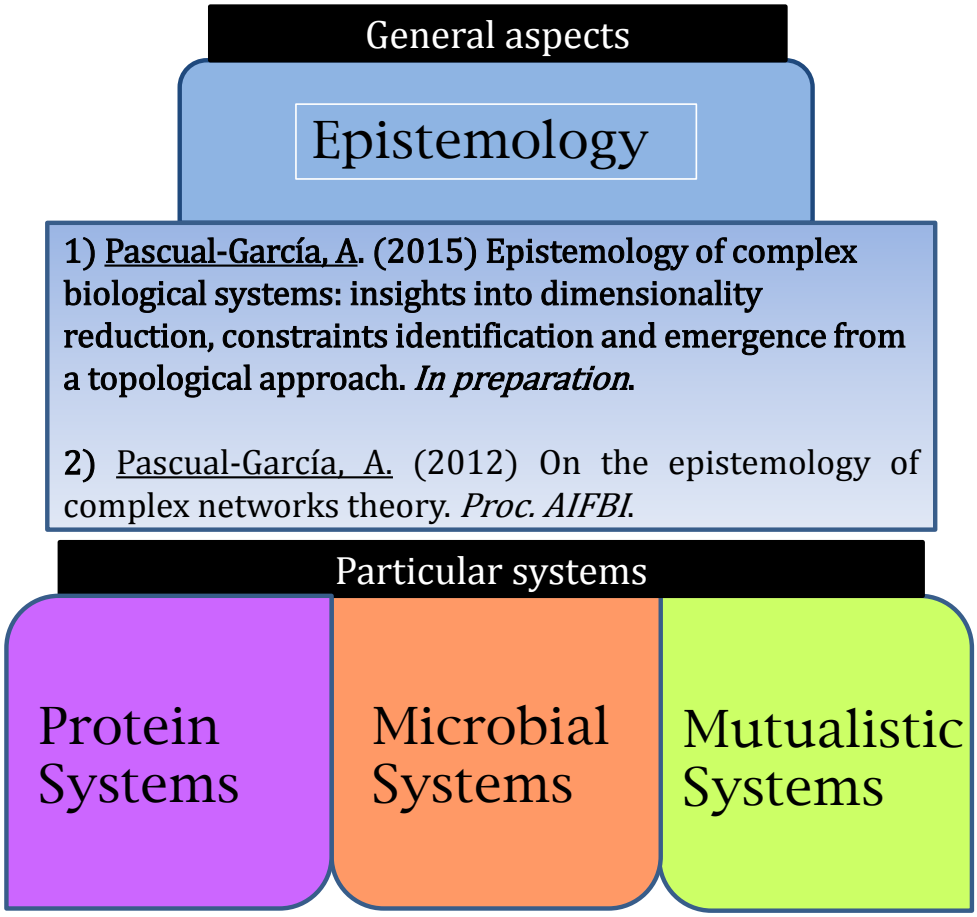
Given that, with our generic formalism, we can fairly compare different systems, we propose to measure the information content of the system and its constraints in terms of logical propositions. In this way, we consider that an epistemologically accessible emergent pattern should require less information to describe the constraints acting on the system than to describe the system itself, otherwise we would not be able to simulate the system behaviour following a bottom-up modeling approach. This is consistent with the notion of weak emergence proposed by Bedau [Bedau (1997)]. On the other hand, strong emergence would be a limiting case where we need the same number of propositions –and containing the same information– than those needed to describe the system, which would reflect the fact that the system's components are constrained as a whole and that the observed behaviour is hardly traceable from bottom-up approximations [Bar-Yam (2004)].

Nevertheless, we doubt on the practical viability of our proposal for real systems, and then we consider another approximation for measuring the emergence strength that considers the epistemological consequences derived from the intervention of the observer in the output of the model [Boschetti (2011)], a notion entirely compatible with Granger causality [Seth (2010)]. This is a methodological procedure which establishes the causality between variables in a dynamical model through the systematic perturbation of single variables, and a posterior measurement of the effects of these perturbations over the other variables.

These effects that can be quantified through the relative performance of the system with respect to the unperturbed behaviour. Therefore, if we deal with a model explaining an emergent behaviour, we can measure its emergence strength quantifying how the description is lost when we intervene in the system, systematically neglecting variables.

We discuss that this procedure is very general, and it may be applied to different methodological procedures such as pattern-based (null) models (corresponding to methods presented for microbial systems) or the development of mechanistic models (developed for mutualistic systems).

1.1. Article [EPIS-1]



Epistemology of complex biological systems: insights into dimensionality reduction, constraints identification and emergence from a topological approach

Alberto Pascual-García

April 4th, 2015

Centro de Biología Molecular "Severo Ochoa" (CSIC-UAM).
c. Nicolás Cabrera 1, campus-UAM, E-28049. Madrid (Spain).
Email. alberto.pascual.garcia@gmail.com

Abstract

Understanding complex processes requires to face essential epistemological problems arising at the different scientific stages, from the identification of relevant patterns to the development of predictive models. In this paper, we introduce a novel perspective that considers a framework founded on logical principles, what allow us to build a generic definition of complex system. Following this conceptual setting, we aim to identify the origin of difficulties found in the process of modelization attribuible to epistemological ambiguities or innacuracies in the concepts defined.

We start defining the conceptual setting that describes a complex system from experimental measurements. The extension of these concepts over the physical space of objects induce a concrete topology, an approximation that was previously used to propose a formal definition of *vagueness*. This formal setting allows for discussing important questions such as the reduction of dimensionality or the determination of system boundaries, where the difficulties we show are related with the existence of extensional vagueness. We next address the problem of the definition of emergent behaviour, and we find that it is possible to solve previous controversies about the definition of *weak* and *strong emergence* focusing in the identification of constraints in the system. Furthermore, we provide an operative definition based on the concept of *intervention*, compatible with the scientific method and consistent with the notion of Granger causality.

We hope that this novel approach estimulates both the application of this framework to new epistemological problems, and the interest of complex systems scientist in the epistemological basis of their research.

1 Introduction

"What urges you on and arouses your ardour, you wisest of men, do you call it "will to truth"? Will to the conceivability of all being: that is what I call your will! You first want to make all being conceivable: for, with a healthy mistrust, you doubt whether it

is in fact conceivable. But it must bend and accommodate itself to you! Thus will your will have it. It must become smooth and subject to the mind as the mind's mirror and reflection."

Friedrich Nietzsche [1]

Scientific modeling is probably one of the best examples of a human activity fitting the words of Zaratustra: it requires the generation of conceptual representations of processes which depend, many times, on uncomfortable features such as measurement inaccuracy, constituents interdependence or the dynamical nature of the process observed. We further aim to incorporate these representations within a mathematical or computational framework, which is nothing but a comfortable place where we reaffirm our confidence in the acquired knowledge. Indeed, building a formal framework provides a favorable environment for reaching new analytical and computational results, thus accelerating the outcome of new predictions while building a solid structure for those processes already accepted. Nevertheless, in comparison with the amount of results generated, there is relatively little attention given to the epistemological basis of scientific modeling and its formalization. In this sense, modern scientific modelling –and, by extension, epistemology–, faces with complex systems one of its most important challenges.

Complex systems are composed by a large number of entities and processes driven by nonlinear interactions between their components, which behaviour cannot be decomposed into the sum of the individual behaviours of the components in isolation. This definition can be applied to systems coming from a wide range of disciplines, being nowadays a central concept in interdisciplinary research. And we observe that, the most characteristic property of these systems, is that they show a particular kind of collective behaviour. Among this kind of collective behaviours we find phenomena such as magnetism, patterns observed in dissipative systems like hurricanes or convenction cells or, in biological systems, patterns on animal skins or flocking behaviour. Looking at these examples it seems that the particularity relies in an apparent discontinuity between the macroscopic properties and the microscopic description, which has led to coin the adjective *emergent* for these behaviours and to the correspondent patterns.

Since a basic tenet in scientific method is that macroscopic properties can be described starting from microscopic descriptions a critical question arises here, namely how is it possible to obtain a satisfactory conceptual representation of emergent behaviours when the definition of emergence apparently implies a discontinuity between the microscopic and the macroscopic representation, –thus it is hardly traceable from the analysis of lower level constituents–.

Here we will analyse the epistemological basis of different steps in the analysis and formalization of complex systems. Questions such as the system definition, the identification of patterns and dynamical constraints or the definition of emergent behaviours, will be analysed within a topological framework through simple examples. We reasoned that, if it is possible to provide a topological description of the different stages in complex systems modelization, we should be able to identify the origin of the different problems faced in this process. This means that we assume that complex systems are epistemologically accessible and the associated problems scientifically tractable. In particular, the framework we use establishes a topological linkage between the knowledge subject's conceptual setting and the objects observed. Analysing with simple examples how this conceptual setting is built for complex systems, we aim to identify the causes of the arising difficulties. For instance, if we find out a new property that we identify as emergent, we wonder whether it is observable any characteristic logical feature justifying the use of this adjective. Therefore, the novel introduction of this framework will allow us to highlight the difficulties we face in the modeling process, providing a more expressive perspective of the different problems.

The manuscript is organized as follows. First we introduce the phylosophical context of our proposal. Then we introduce the definitions and operations needed to understand how this proposal is formally materialized within a topological framework. Three different results are then discussed following this proposal. We will analyse in the first place the classification schemes in Biology –that can be exposed more generally in terms of the dimensionality reduction techniques– providing as an example the classification of protein structures. Making use of a toy model, we next discuss a critical step in complex systems modelling, namely the identification of constraints, what will lead us to introduce the concept of traceability. Using these new tools we will propose a formal definition of emergent properties further identifying the source of vagueness arising from their observation. Our extensive analysis allow us to establish correspondences with existing methods, suggesting an epistemological program that we summarize in the end, where we also introduce as an example the analysis of the nestedness pattern in mutualistic model ecosystems.

2 Methods

2.1 Phylosophical context

When the system under analysis is what we have called a complex system, it will be necessary to reduce the level of detail that we consider in its representation, otherwise its complexity makes any analysis unfeasible. The answers obtained from these models will typically lead to breadth concepts, that can eventually exceed the knowledge boundaries of the specific discipline from which the question originally arose. This is probably why controversies are found around this concepts, given the intrinsic vagueness implicit, in a sense that we justify as follows.

For any concept we can distinguish between its intension and its extension [2, 3]. The intension refers to the properties or characteristics connoted by the associated concept, whereas the extension is the domain of applicability of the concept in the physical world, i.e. the actual objects it refers to.

In our case we will depart from experimental data in our analysis. In an ideal scene, we may think that we start with a one to one map between data and the intension of the concepts built from this data. Moreover, we may consider that the extension is clearly set, given that the objects we have measured directly provide the extension of the properties measured. However, it is easy to find examples where this starting point is neither achieved. Let us consider for instance a high-throughput sequencing experiment where we aim to recover the genes contained in an environmental sample, and further consider that we intend to associate each gene found to a given bacterial specie.

First, there is not an accepted definition of bacterial specie nowadays [4], and it is used instead an operative definition which depends on arbitrary thresholds for its determination (the Operative Taxonomic Units, (OTUs)). Second, even if a given gene is believed that it is found in a particular OTU, we cannot assert that this gene was actually found in such OTU in the experiment. The reason is that, in this kind of experiments, all the DNA found is fragmented, amplified and then reconstructed on the basis of previous knowledge (unless we deal with *de novo* sequencing). In this way, a gene which is believed to belong to an OTU a, will be assigned to that OTU. However, it may happen that the target gene observed was actually horizontally transferred [5] to a different OTU b, and it is from this OTU from which the gene was actually extracted. Therefore, in this example we observe intensional uncertainty (in the definition of OTU) and extensional uncertainty (in the incorrect assignation of the gene to OTU a when it was actually found in OTU b).

Although in complex biological systems we do not consider concepts such as *ghost*, whose intension can be expressed but whose extension is empty, we actually deal with concepts where we can recognize vagueness in the definition, either intensional, extensional or both [2]. Intensional vagueness arises when it is not possible to completely determine the properties constituing a given concept, as it was the case for bacterial species in the previous example. Extensional vagueness in the other hand arises when we are not able to sharply separate the objects of observation to which a given concept is referring to, this was the case for the gene. Both types of vagueness are interlinked, as it is not possible to rule out the intensional vagueness if there exist extensional vagueness and viceversa. This typically happens with general concepts [6], which are a frequent matter of debate in the scientific literature but which intrinsic vagueness may be even considered an asset [7]. The difference between organism and machine [2], the different definitions of stability [8] or the exact intension of concepts such as function, autopoiesis and complexity [9] are already classical examples of such discrepancies.

Generality is not the single source of vagueness in concept analysis. Any change in the properties of the referred observational objects will be an important source of vagueness, as one must determine the intension of concepts associated to variable properties. This is particularly challenging when it affects to the quantitative variables in such a way that a qualitative change is observed. An accessible example may be the concept of phase transition, where the description of these phenomena has generated great interest due to its difficulties. For instance, from the macroscopic point of view there is a continuous interest in finding order parameters that may allow us to monitor the transition from a given phase to the other. And finding a microscopic correspondence has generated much attention with the development of important theories such as the renormalization group theory [10], that predict universal patterns in these transitions that allow us to predict critical changes [11].

Following the classification of concepts proposed by Georgescu-Roegen [12] we will refer to concepts containing any source of vagueness as dialectic. The dialectic notion we propose is compatible with the Hegelian notion of dialectics, where vagueness is implicit in the gradualness transition from quality to quantity [13]:

> "It is said, natura non facit saltum [there are no leaps in nature]; and ordinary thinking when it has to grasp a coming–to–be or a ceasing–to–be, fancies it has done so by representing it as a gradual emergence or disappearance. But we have seen that the alterations of being in general are not only the transition of one magnitude into another, but a transition from quality into quantity and vice versa, a becoming–other which is an interruption of gradualness and the production of something".

Dialectic concepts are even better understood in contraposition to arithmomorphic concepts, which are those concepts that can be discretely differentiated. Following the words of Georgescu-Roegen: "[arithmomorphic concepts] conserve a differentiate individuality identical in all aspects to that of a natural number within the sequence of natural numbers". Arithmomorphic concepts are suitable for formal reasoning, and therefore compatible with a quantitative treatment.

Scientific modeling may be viewed as an activity aiming to construct arithmomorphic schemes, and systems that must be described through dialectic concepts challenge scientific modeling. Concepts such as the adjective democratic, being defined with a wide variety of implicit qualitative variables changing in space and time, make difficult to propose an arithmomorphic scheme oriented to provide an objective measure (therefore discretely differentiated) of this concept, and it must be understood as dialectic. With objective we mean that it is possible to provide a logical or mathematical definition. Although, if it is not found, it does not exclude that there is an

intersubjective definition compatible with the scientific method, but such an alternative will be probably continuously revisited. An illustrative example of what is understood as arithmomorphic scheme and its relation with dialectic concepts comes from the classification problem, that will be discussed in detail below. When a classification is obtained it means that we find an extensionally sharp conceptual scheme compatible with the mathematical definition of equivalence class. If we do not find such an objective classification we can still obtain an intersubjective classification but its extent and interpretation will be subject to continuous debate.

In this paper, we are going to deal with systems whose properties are well defined in an starting ideal scenario –and then their description could be regarded as arithmomorphic–, but they lead to new properties with either intensional or extensional vagueness. We will define what should be understood as a vague concept from a logical framework that, after building a topology in a classical sense, provides an extensional definition of vagueness. This definition, when applied to the analysis of the system's microstates, clarifies different concepts such as the ambiguity in classification schemes, the identification of dynamical constraints or the vagueness associated to the concept of emergence.

2.2 System definition: preliminaries

We start proposing a glossary of terms concerning the system definition, some of them close to those definitions proposed by Ryan in [14]. We will call an (object of) observation o_i, to a set of basic magnitudes associated to a given entity. Each of this magnitudes is a function f_M of the cartesian product of a collection of M sets –where at least one of them is determined by an experimental measurement–, onto real numbers $\mathbb{R}$, i.e. $f_M : A \times B \times ... \times M \rightarrow \mathbb{R}$. The non measurable sets may refer for instance to a set of measurement units (grams, meters,...) to a set of reference frameworks, or any other set necessary to determine the final magnitude. For simplicity, we will consider that any variation in the magnitudes is a consequence of a variation in the outcome of a measurement and thus, in the following, we will not distinguish between magnitude and measurement when objects of observation are discussed.

In this way, we will consider that our system is characterized by a bunch of M quantitative and/or qualitative (i.e. binary) magnitudes $X = \{x_k, k = 1, .., M\}$. Given that we are interested in complex systems, we will consider that our system consists on a large number of entities, that we denote with N. We will call *scope* to this selection of objects whose size, N, will be very relevant for us and it must be noted that this choice implicitly determines the spatial boundaries of the system. Determining the scope is already a difficult task for large dynamical systems. These difficulties arise, on the one hand, from the identification of these entities because, when the number is large, a complete characterization may be unfeasible. On the other hand, it will be also difficult to define the separation between system and environment, as this separation cannot be achieved many times using strictly objective arguments [15, 16].

The bunch of variables selected X are intended to be sufficient to answer the questions that will be addressed in the research, therefore fully describing the system. For simplicity in the exposition, we start considering an ideal scenario where all these variables can be quantified for any entity within the system, leading to $N \times M$ specific values. This assumption will not affect our conclusions, as we can assign a vanishing value to any variable from which the associated magnitude is not observed for one or several entities.

Each of these variables have a *resolution* $r(x_k)$, which is the finest interval of variation that we set for that variable, and it may be established from different arguments. For instance, the resolution may be limited by the intrinsic error in the measurement (which is an ontological

limitation). Another possibility arises when the expected influence of a given variable on the system's description is small for a given shift in the value, and a coarser discretization is then justified (which would be an epistemological limitation). If we call $I(x_k)$ to the domain of a variable, the number of possible values considered will be $\zeta_k = I(x_k)/r(x_k)$. We will call resolution R of the system to the finest variation that allow us to distinguish two states of the system, $R = max(\{\zeta_k\}_{k=1}^{M})$.

This choice of variables together with the set of viable values will be called the *focus* F of the knowledge subject, and we can provide an upper quantitative bound for the focus: $F \sim M \times R$. We finally call the *scale* to the set of specific values $\{N, M, R\}$. In this way a factor multiplying any of these values will be understood as a change in the scale of the scope (if we modify N), or the focus (if we modify M or R). We must additionally note that, following this definition, the scale is both an ontological attribute as determined by N, but it also depends on the epistemological attributes determined by M and R, and thus the breadth of the focus is essentialy set by epistemological choices. Interestingly, it has been suggested that emergent behaviours are the consequence of a change in the scope [14] and not in the focus [17].

2.3 Microscopic and macroscopic descriptions

Following the definitions introduced and inspired into Statistical Mechanics, we aim now to differentiate two types of variables providing a microscopic versus a macroscopic description of the system. A macroscopic property arises from the observed dynamical evolution of a system during a certain period of time during which, even if the microscopic variables are continuosly changing, the macroscopic variables remain invariant. In this way, the most important difference between both description stands on the temporal scale and on the focus, but not essentially in the scope.

We call microstate μ to a vector where each cell contains the specific value of the microscopic variable x_k measured at a given time for the observed microscopic object o. Similarly, we will call macrostate ξ to another vector where each cell contains the specific value of the macroscopic variable y_k measured for the observed macroscopic object $\hat{o}$. Taking the separation of scales between the microscopic and macroscopic descriptions above mentioned, we observe that a macroscopic description is obtained when a whole set of microstates is considered, i.e. measuring a single macrostate ξ implies that we have considered a set $\{\mu\}$ of microstates, what is known as an *ensemble* of microstates. In some cases, the macroscopic variables y_k can be obtained applying a surjective map f over the microscopic variables $f(x_{ik}) \rightarrow y_k$. For instance, if we deal with an incomplete (statistical) microscopic description of an ensemble $P(\mu)$, we can obtain a coarse determination of a macroscopic variable y_k through a weighted averaging of the correspondent microscopic variable, x_k, over the ensemble $\langle x_k(\mu)P(\mu)\rangle$.

Nevertheless, it is not always evident which is the microscopic property describing macroscopic features of collective emergent behaviours. When the system is constrained to a certain region of the phase space there is a breaking of symmetry, namely there is not equiprobability in the values of the variables, thus losing ergodicity [18], p. 186. This is due to the existence of external or internal constraints limiting the behaviour of the system. And, as we will attempt to clarify, a necessary condition for determining a microscopic property associated to every microstate visited, requires the determination of the existing constraints. We will see that the nature of the different constraints acting on the system determine its epistemological accessibility and thus our ability for providing a satisfactory explanation of an emergent behaviour.

To finish, we would like to underlie that, what is considered a macrostate and a microstate, may change if we move from one scale of description to another scale. Let us consider for instance

a system described within certain temporal scale by a set of microstates $\{\mu_i\}$ which are associated to the observation of a single macrostate ξ. Let us now assume that the system evolves under a longer path thus changing the macrostate, and that we store T snapshots of the dynamics, leading to an ensemble of macrostates $\{\xi_u\}_{u=1}^{T}$. It is possible to consider that each of these macrostates is now a microstate $\hat{\mu}$ for a new system with a larger scope and lower resolution $\xi_u \rightarrow \hat{\mu}_i$. Given that the scope of a macrostate will be always larger than that of a microstate ($N_\xi \geq N_\mu$), whereas it occurs the opposite with the resolution ($R_\xi \leq R_\mu$), in this exercise we have increased the scope and reduced the resolution. This movement along the different scales will be very relevant when evolutionary systems are considered, given that we will need to distinguish at least two spatial and temporal scales. For instance, a change in the scale of observation is needed if we move from the analysis of a single individual to the analysis of a population of individuals.

Note that this change in the scale requires an effort to reduce the system description, but this kind of reduction has been performed from the very first step in our definitions. For the definition of scope, we have neglected entities. For the focus, we have neglected variables and probably restricted their viable values assuming a lower resolution. Furthermore, any map between microstates and macrostates again considers a reduction in the information provided by the microstates. In general, for both very broad or very detailed questions the technical complexity increases and a reduction in the description is unavoidable, and it is important to remark that this exercise does not mean that the approach is reductionist. Reductionism should be considered an epistemological attitude where it is accepted the assumption stating that any macroscopic description is a simple extrapolation of the properties of the microscopic description [19]. Instead, we accept that in complex systems there are discontinuities between the different levels of description and that, for each new level, new properties may arise. We are interested here to understand which are the minimum epistemological conditions in order to say that a microscopic description is a valid representation of an emergent macroscopic observation.

2.4 A topological description of the phase space induced by measurable properties

In the following sections we introduce to the reader a novel application of topological notions derived from logic, whose novelty relies on its ability to formally describe epistemological questions that are hardly addressed by other approximations. A nice introduction for computational scientists to the generalization of the approach we introduce here, called formal topology, can be found in [20], and a relevant application to the epistemological determination of what should be understood as a vague concept is found in [21].

Indeed, we aim to show here that well known difficulties discussed around the concept of emergence, emergent property, emergent behaviour or emergent theory [9, 14, 22, 23, 24, 17], can be essentially attributted to an intrinsic vagueness arising from the simultaneous coexistence of a microscopic and a macroscopic description, and the application of topological notions allow us to clarify the different sources of vagueness. Thus, our effort in the application of a novel formalism is primarly justified because we find a more expressive picture of the epistemological problems we face when dealing with complex systems.

2.4.1 Measurable properties, concepts and their extension.

Let us start introducing some definitions, most of them already provided and justified in [21], that we recover here for completeness –although we will not reproduce proofs and lemmas derived from

these definitions given that can be found in the work of Boniolo and Valentini–.

For the shake of simplicity we will start considering that our objects of observation $o \in O$ are the components of a complex system at a given time, i.e. we focus on a single microstate μ with N components described by M variables with resolution R. Each of these components is what we consider for the moment an object of observation. We will move later towards a description where each object of observation is a microstate, becoming the whole space of objects the observed phase space. In general, all the definitions considered in the following for a single microstate are easily generalized for other objects with a different scale.

Definition: We call a *property* or *characteristic* $c_a = x_k^*$ to the specific value x_k^* of a variable x_k, out of the ζ_k possible values, measured over an object of observation o. In this way, if we consider two different measurements of our variables for the same entity, each observation will constitute a different object (of observation).

Definition: We call *focus* F to the whole set of characteristics considered by the observer: $F = \{x_k^l;\ k = 1, .., M;\ l = 1, .., \zeta_k\} = \{c_a;\ a = 1, .., \tilde{M}\}$, with $\tilde{M} = M \times \sum_k \zeta_k$. We have made explicit here the discrete nature of the conceptual setting and we make more precise the relation between resolution and focus, which achieves a suitable description in terms of characteristics, in turn leading to the definition of concepts.

Definition: We call a *concept* ν to any non-empty finite subset of F: $\nu = \{c_1, ..., c_P\}$, with $P \leq \tilde{M}$. Therefore we have defined here the intension of concepts. Given that a single characteristic is indeed a concept, $\nu = \{c\}$, in the following and for the shake of simplifying the exposition, the terms concept, measurement, property or characteristic will be equivalent unless a distinction is required.

2.4.2 Binary operations

From the previous definition it is immediate to propose binary operations to build new concepts.

Definition: (Conjuction of concepts). Let $\nu_1 = \{c_1, ..., c_P\}$ and $\nu_2 = \{d_1, ..., d_Q\}$ be two concepts. Then the conjuction of $v_1$and v_2 is the concept:

$$\nu_1 \wedge \nu_2 = \{c_1, ..., c_P, d_1, ..., d_Q\} \tag{1}$$

The conjuction of concepts is in turn a concept which consists on the set of all the characteristics contained in both concepts. On the other hand we may aim to extract, given two concepts, the common characteristics they share, which is expressed in the following binary operation:

Definition: (Disjunction of concepts). Let $\nu_1 = \{c_1, ..., c_P, b_1, ..., b_L\}$ and $\nu_2 = \{d_1, ..., d_Q, b_1, ..., b_L\}$ be two concepts. Then the disjunction of $v_1$and v_2 is the concept:

$$\nu_1 \vee \nu_2 = \{b_1, ..., b_L\} \tag{2}$$

The disjunction of concepts is a concept which consists on the set of all common characteristics contained in both concepts.

We note here that the definition of concepts arising from the basic characteristics and extended through binary operations, readily leads to a focus partition. The exploration of this partition

and how it in turn induces another partition on the set of objects, stands on the basis of our epistemological approach. Understanding the relationship between this partition and the objects of observation requires to first determine a constitutive relationship between any single charactersic belonging to the focus F and the set of objects O, that we build invoking an extensionalist setting. The constitutive relationship will express that the objects become cognitively significant by means of the characteristics measured, and in turn by the concepts we build from them:

Definition: (Constitution relation). Let F be the focus over a set O of objects. Given any $o \in O$ and any $\nu \in F$, we introduce a binary relation, $\Vdash$, that we call constitution relation, such that by $o \Vdash \nu$ we mean that ν is one of the concepts constituting o.

With the constitution relation we establish the means by which the objects of observation are expressed via the conceptual apparatus of the knowing subject. In addition, we would like to know which objects are constituted by a given concept, what is provided by the following map:

Definition: (Extension of a concept). Let $\nu \in F$ be a concept. Then the extension Ext of ν is the subset of objects of O constituted by ν, that is:

$$Ext(\nu) = \{o \in O \mid o \Vdash \nu\} \tag{3}$$

We note here that an immediate consequence of the above setting is that any object of observation has associated necessarily a concept, i.e. it is just accessible by means of the conceptual apparatus of the knowing subject. This assertion, if accepted in general, would lead to a Kantian epistemological positioning [21]. In our case it should be seen as a simple consequence of the fact that our objects of observation are built from the measurements of a reproducible experimental setting, and thus this is true by construction. Finally, we aim to know what is the extension of a subset U of concepts $U = \{\{\nu_1\}, ..., \{\nu_L\}\}$.

Lemma: Let U be a subset of the set F of concepts. Then, the extension of U is defined by setting

$$Ext(U) = \bigcup_{\nu \in U} Ext(\nu) \tag{4}$$

We should not confuse a concept built by conjuction of different characteristics with a subset of concepts containing the same characteristics (remember that a characteristic is itself a single concept). In the former case we look for objects containing *all* the characteristics –and thus its extension, as we will see below, is reduced by the fact that we consider objects that should contain more and more characteristics–, whereas a subset of characteristics extends over objects containing *any* of the characteristics, being its extension the union of the extension of characteristics.

2.4.3 Topological notions

From the previous definitions we can introduce a theorem and some more definitions that are on the basis of our topological approach to the analysis of complex systems. Again, we remind that most of these definitions are introduced and discussed in [21].

Theorem 1: If the map Ext satisfies the extension condition, then the family $\{Ext(U)|U \subseteq F\}$ is a topology over the set O, where U is a subset of concepts of the focus F.
This map is central in our arguments. The basic characteristics are defined in terms of measurements over specific objects, and thus the extension provides a map between these characteristics and the sets of objects. If we call power set $\wp$ to the set containing all the possible ways that another set, in our case O, can be divided into, a topology will be a subset of the power set verifying some particular properties. A topology is a collection of subsets called *open sets*, which include the empty set and the whole set, verifying: 1) the arbitrary union of open sets is another open set in the topology. 2) The binary intersection of open sets is also another open set in the topology. In this way we say that a topology is a subset of $\wp$ which is *closed* under arbitrary union and binary intersection.

What we are expressing here is that, once we have built our conceptual setting from measurements, the extension function induces a partition in the set of objects, and this partition fullfills the conditions for being a topology. In this way we can make profit of the topological notions of open and closed sets that we will use along our exposition and whose definition will be made more precise in the following.

Definition: (Open set) Let A be a subset of the set O of objects. Then A is an open set if it coincides with its interior $Int(A)$ where,

$$Int(A) = \{o \in O \mid (\exists \nu \in F)\ o \Vdash \nu\ \&\ ext(\nu) \subseteq A\} \tag{5}$$

Definition: (Closed set) Let A be a subset of the set O of objects. Then A is a closed set if it coincides with its closure $Cl(A)$ where,

$$Cl(A) = \{o \in O \mid (\forall \nu \in F)\ o \Vdash \nu \Rightarrow (\exists o' \in O)\ o' \in ext(\nu)\ \&\ o' \in A\} \tag{6}$$

Definition: (Border) Let A be a subset of the set O of objects. Then the border $Bd(A)$ of A is the set

$$Bd(A) = Cl(A) \cap \bar{A} \tag{7}$$

where $\bar{A}$ stands for the complement of the set A with respect to O.

Definition: (Vagueness) Let ν be any concept and U be any set of concepts. Then ν is a vague concept if $Bd(Ext(\nu))$ is not empty, and U is a vague set of concepts if $Bd(Ext(U))$ is not empty.
In this way and, as we anticipated in the phylosophical context, we are are able to work with an extensionalist notion of vagueness. In the following results, we depart assuming that there is no intensional vagueness arising from the experimental data, what allow us to focus on the extensional vagueness. Then we will consider difficulties in our experimental setting, further discussing the intensional vagueness.

3 Results

3.1 The top-down approaches in complex systems: the reduction of dimensionality

3.1.1 Extension of concepts obtained through binary operations: focusing on disjunction

We aim to explore how can be obtained the extension of those concepts built through binary operations over sets of concepts. When we obtain a new concept τ via conjuction, for instance $\tau = \nu_1 \wedge \nu_2$, the extension of the new concept will be the intersection of the sets of objects associated to each of the starting concepts $Ext(\tau) = Ext(\nu_1) \cap Ext(\nu_2)$. Aiming to fully identify a single object requires to determine a sufficiently large number of concepts in order to sharply separate it from the other objects, being conjuction the basic operation that permits to reach more precise descriptions.

Let us take as an example the description of a set of proteins $\{o_\alpha\}$ provided by the sequence of their amino-acid composition, which is embedded within an evolutionary phase space. Each amino-acid molecule in the protein is a component of the system which, considering the most basic description, is described by its position in the sequence and by a single variable whose specific value consists on one out of the 20 natural amino-acids encoded by DNA. In this way, an example of concept within this description would be something like $\nu_i =$ *"cysteine in position i"* –which in turn is built by conjuction of the more basic characteristics "*cysteine*" and "*i*"–. A protein sequence o_α will be subsequently built by conjuction of a set of such a kind of concepts describing the amino-acid observed at each position, i.e $\alpha = (\nu_1 \wedge \nu_2 \wedge ... \wedge \nu_N)$ (see Fig. 1). The sequence becomes uniquely determined under this description, i.e. the extension of the sequence maps exactly one object of observation, namely, the protein under study: $Ext(\nu_1 \wedge \nu_2 \wedge ... \wedge \nu_N) = Ext(\alpha) = o_\alpha$. In summary, conjuction underlies bottom-up approximations, where we focus in an accurate description through the compilation of concepts.

Let us now have a look to another kind of concepts λ, which are obtained via disjunction, $\lambda = \nu_1 \vee \nu_2$. Following the equation 2, given that $\nu_1 = \{c_1, ..., c_P, b_1, ..., b_L\}$ and $\nu_2 = \{d_1, ..., d_Q, b_1, ..., b_L\}$, the extension of the concept $\lambda = \{b_1, ..., b_L\}$ will be given precisely by $Ext(\lambda) = Ext_{i<j}(b_i \wedge b_j)$ (see Fig. 1). From this definition we note that $Ext(\{\{\nu_1\}, \{\nu_2\}\}) = Ext(\nu_1) \cup Ext(\nu_2) \subseteq Ext(\nu_1 \vee \nu_2)$ and thus this is an operation that allow us to look for commonalties, which may be extended to objects that are not included in the sets of objects over which the concepts ν_1 and ν_2 are extended. Disjunction stands out as a relevant operation to look for breadth concepts, and it is consistent with the intuition stating that these concepts tend to overtake the boundaries of our starting focus.

In addition, it is remarkable to observe that to obtain these concepts the basic operation arises from comparison of objects. Indeed, since long it has ben recognized the importance of comparisons for the proper determination of any object that may be viewed as a negative determination through the exploration of the limits of the object, as was stated by Hegel [13]:

> "the object, like any determinate being in general, has the determinateness of its totality outside it in other objects, and these in turn have theirs outside them, and so on to infinity. The return–into–self of this progression to infinity must indeed likewise be assumed and represented as a totality, a world; but that world is nothing but the universality that is confined within itself by indeterminate individuality, that is, a universe."

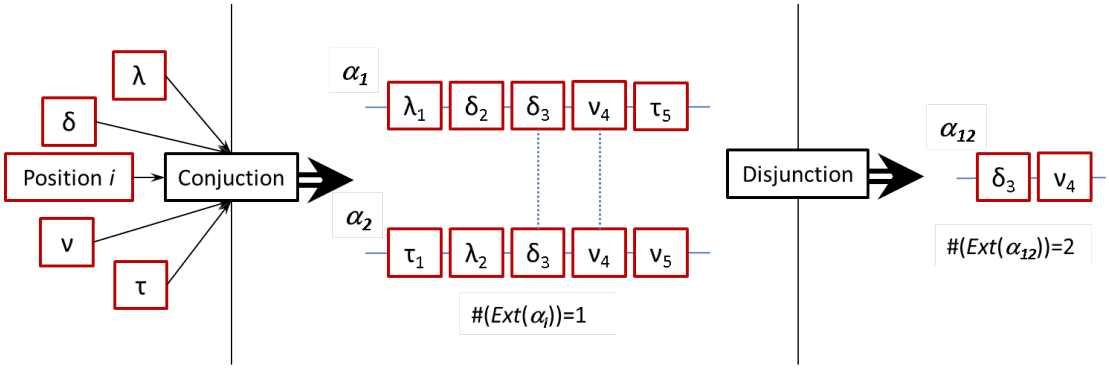


Figure 1: Starting from the knowledge subject's conceptual apparatus (left), two sequences α_1 and α_2 are built through conjuction of the basic concepts, being themselves concepts (center). These sequences uniquely determine a single object, for instance a protein sequence, and thus $\#(Ext(\alpha_i)) = 1$. By comparison of both sequences we observe two common concepts (linked by dotted lines) that we extract through binary disjunction leading to a concept α_{12} (right) containing less basic concepts but whose extension is larger than the original sequences ($\#(Ext(\alpha_{12})) = 2$) being its scope larger. In the case of proteins it may be understood as a signature of their common ancestry, i.e. of their homology.

With disjunction we explore the progression of an object into other objects that may eventually lead to the identification of new objects exceeding the individuality of the entities themselves, and then back: the identification of these objects with a larger scope reinforce the individuality of the starting objects. Recovering the example of the set of proteins sequences $\{o_\alpha\}$, given two protein sequences α and α' with N amino-acids, one of the questions we are typically interested to answer is which is the percentage of sequence identity shared between both sequences, i.e. the fraction of identical amino-acids that are similarly located in both sequences. The natural operation to obtain this subset of common amino-acids is disjunction. Indeed, the number of shared concepts (in this case amino-acids), normalized by the total number of amino-acids N leads to the sequence identity. Extending the example to a larger set of proteins, if we find a common region of amino acids shared by all of them, we are dealing with a new object that exceeds the sequences themselves. This new concept, which represents a vestige of their common ancestor, is on the basis of a breadth concept, namely the "homology" between these sequences, whose determination reinforce the identity of the starting sequences.

Of course this operation may become much more complex if the length of the sequences is not the same. When this happens, we need to solve a harder poblem that consists on identifying which positions should be considered equivalent, which is the intended task for alignment algorithms such as BLAST [25], but the essence of this new problem still relies on the aplication of concept disjunction. In general, the search of similarity measures, dissimilarity measures or distances is an essential task in Biology and Ecology [26] aiming to understand, following a top-down approach, the information shared between the different objects of observation. And, in general, this operation is on the basis of dimensionality reduction techniques such as principal components analysis [27]. Following our framework, these are techniques aiming to obtain a representation with the minimum number of concepts explaining the maximum variability in the space of objects. In this way we are able to *talk* about the set of objects using a subset of concepts, which is equivalent to the more

traditional notion of dimensionality reduction.

3.1.2 The problem of classification and the determination of system boundaries

In order to better understand the application of disjunction and dimensionality reduction we provide a simple classification example, which is a critical activity in the earlier stages of development of any science [28], and it is particularly challenging in Biology [29], where they are continuously revisited. For instance, species definition is a classical example where discrepancies do exist nowadays, see for instance the differences between the Ecdysozoa and Coleomata hypothesis [30, 31], or the difficulties in the definition of bacterial species [4]. The access to new quantitative data such as whole genomes, open new opportunities to improve these schemes.However, the intrinsic variability of the data is so prevalent that it is mandatory to consider the epistemological basis of any proposal [29], what typically leads to vigorous debates, –see for instance the debate around the classification of protein structures that we will treated in more detail below [32]–.

In general, when the entities considered in any attempt of classification contain an increasing amount of dialectic concepts there are also increasing chances of dealing with a frustrated problem. Frustration is understood here as an irreconciliable balance between two or more tendencies –in this case between two or more classification schemes–, consistently with the notion considered in physics [33].

In this section we analyse an example where the vagueness is attribuible to difficulties in the extension of the different concepts, i.e. we address extensional vagueness. Let us consider that we deal again with a set of entities $\{o_\alpha\}$ described by a set of concept sequences $\{\alpha\}$ arising from four variables $\{\nu, \tau, \lambda, \delta\}$ whose specific values lead to the concepts needed to fully specify any entity. Two different values of the same variable will be differentiated by subindices, e.g. ν_1 and ν_2. Thus, once we experimentally obtain the values of the variables we can build the concepts and uniquely determine the sequences by concept conjuction. Let us start with the following example, focusing in only three sequences (see Fig. 2):

$$\begin{aligned} \alpha_1 = &\quad \nu_1 \wedge \lambda_1 \wedge \tau_1 \wedge \delta_1 \\ \alpha_2 = &\quad \nu_1 \wedge \lambda_1 \wedge \tau_1 \wedge \delta_2 \\ \alpha_3 = &\quad \nu_1 \wedge \lambda_1 \wedge \tau_2 \wedge \delta_3 \end{aligned}$$

To make more explicit the relation between the concepts and the sequences, we show in the following the extension of these concepts. Given that each object of observation o_α is uniquely determined by a sequence α, for the shake of simplicity in the notation we will denote the object with its sequence, i.e. with its formal counterpart[1]. We also consider the fact that there exist other sequences, apart from those we are explicitly analysing, what we are going to specify in terms of arbitrary subsets $\{\hat{\alpha}\} \subset \{\alpha\}$. Finally, we will denote with the subindex x to any other arbitrary values for these variables. Keeping in mind these considerations, the extension of the different concepts would read:

[1] A clarification following the notion of formal points should be included here

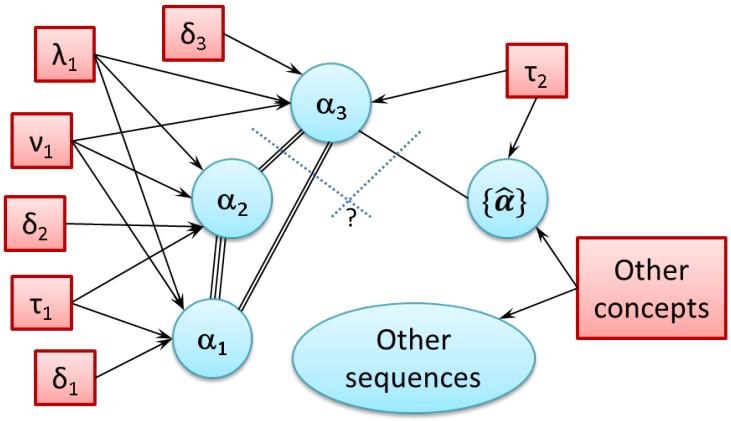


Figure 2: Example of a classification scheme. Concepts are depicted as squares and objects as circles, and we focus in the example on three specific entities α_i $(i = 1, 2, 3)$. Arrows indicate the extension of the concepts and continuous lines the similarity among objects built by concept disjunction, the number of lines indicate the number of concepts shared. Dotted lines crossing over a question mark indicate the two possible thresholds that may be considered to determine the classification (see main text for details).

$$\begin{aligned} Ext(\nu_1) &= \{\alpha_1, \alpha_2, \alpha_3\} \\ Ext(\nu_x) &= \{\hat{\alpha}\} \\ Ext(\lambda_1) &= \{\alpha_1, \alpha_2, \alpha_3\} \\ Ext(\lambda_x) &= \{\hat{\alpha}\} \\ Ext(\tau_1) &= \{\alpha_1, \alpha_2\} \\ Ext(\tau_2) &= \{\alpha_3, \{\hat{\alpha}\}\} \\ Ext(\delta_i) &= \{\alpha_i\} \end{aligned}$$

We remind that by $\{\hat{\alpha}\}$ we stand an arbitrary subset and thus we do not know for the moment the actual extension of those concepts extending on such sets. Therefore we cannot build propositions such as $Ext(\nu_X) = Ext(\lambda_X)$ because we do not know the actual limits of their extensions, we just know that these objects do not extend on the three sequences we are focusing in, although they may do in others. For instance, the concept τ_2 extends on the object described by α_3 and on an arbitrary set. We finally observe that there are concepts δ_i which guarantee the determination of any sequence, given that these are extended only in unique entities.

The next step in order to formally get more insight into the conceptual structure of this system will be to look for an appropriate similarity measure S for sequences' comparison. We propose a simple measure which is built on concepts' disjunction:

$$S(\alpha_i, \alpha_j) = \#(\alpha_i \vee \alpha_j)$$

Where the function $\#(\cdot)$ expresses the cardinality of the set being the result of concept disjunction, i.e. the number of concepts contained. Applying this measure to our example we get

$$\begin{array}{cc} S(\alpha_i,\alpha_i)=4 & \\ S(\alpha_1,\alpha_2)=3 & \\ S(\alpha_1,\alpha_3)=S(\alpha_2,\alpha_3)=2 & \\ S(\alpha_1,\alpha_i)=S(\alpha_2,\alpha_i)=0 & (\alpha_i\neq\alpha_3) \\ S(\alpha_3,\alpha_i)=1 & (\alpha_i\in\{\hat{\alpha}\}) \\ S(\alpha_3,\alpha_i)=0 & (\alpha_i\notin\{\hat{\alpha}\}\bigcup\alpha_2\bigcup\alpha_3) \end{array}$$

The result of this analysis can be represented with a graph (see Fig. 2). A graph or network represents a set of entities together with their relationships, where the entities are represented as nodes in the network and their relationships as links between nodes. Given that the nature of the entities and their relationships is not specified a powerful flexibility is conferred to this kind of network representations, being the basis of complex networks theory [34]. In our case, the sequences represent the nodes and the links represent their concepts' similarities, and we are going to see that this representation provides an immediate intuition of the topological notions previously introduced.

We are already equipped with a formal setting to deal with the following question: Is it possible to objectively find any concept describing the three sequences selected? In other words, we would like to know whether, under this representation, these sequences are different enough from any other sequence, and whether we can *talk* about their distinctive features using exclusively the conceptual setting we constructed. Thus, starting from the rather rough information we handle –given that we do not have information for all the relations in the space– we would like to propose a suitable classification where our three sequences would belong to the same cluster. It is convenient to remind here that a classification consists on a set of equivalence classes. Each class consists on a set of elements linked by an equivalence relationship $\sim$ together with the following properties that must hold for any three elements a, b and c belonging to the class:

$$\begin{array}{cc} a\sim a & (Reflexivity) \\ if\ a\sim b\ then\ b\sim a & (Symmetry) \\ if\ a\sim b\ and\ a\sim c\ then\ b\sim c & (Transitivity) \end{array} \tag{8}$$

It is easy to see that the similarity measure S we have proposed readily verify the requirements of reflexivity and symmetry. In order to determine an equivalence class we need to further verify that there exist a threshold S_0 such that transitivity holds for the elements belonging to the class.

We can start in our example considering the trivial classifications, that would arise if we consider either $S_0=0$ and $S_0=4$. If we consider $S_0=0$ we deal with the whole set which, by definition of topology, is both open and closed and it is a rather trivial classification, with all the elements joined into a single cluster. If we consider instead $S_0=4$, given that the concepts δ_i safely separate any sequence from the others, we will obtain as many clusters as sequences, an exercise that we have already performed to define the system and that do not provide further information about the specific sequences of interest.

Thus, in order to extract common information it seems convenient to neglect the concepts δ_i and then exploring the classifications arising from threshold values $S_0=2,3$ (see Fig. 2). Starting with $S_0=3$ we would join the sequences $\{\alpha_1\}$ and $\{\alpha_2\}$, sharing the concept $\nu_1\wedge\lambda_1\wedge\tau_1$. The set $\{\alpha_1,\alpha_2\}$ of sequences is an open set (by definition because there exist a concept whose extension coincides with the set), but it is also closed because the other candidate that may belong to their closure, $\{\alpha_3\}$, is safely separated by means of τ_2, and thus $Cl(\{\alpha_1,\alpha_2\})=\{\alpha_1,\alpha_2\}$ being a closed set. Therefore the concept $\nu_1\wedge\lambda_1\wedge\tau_1$ describes without vagueness this set and this threshold seems to be optimal to obtain a proper classification.

If we choose now the threshold $S_0 = 2$ we consider the set of sequences $\{\alpha_1, \alpha_2, \alpha_3\}$ described in this case by the concept $\nu_1 \wedge \lambda_1$. This set is also an open set but its closure is the set $\{\alpha_1, \alpha_2, \alpha_3, \{\hat{\alpha}\}\}$, and $\nu_1 \wedge \lambda_1$ is a vague concept given that the border of the set it extends over contains $\{\hat{\alpha}\}$. Therefore choosing this threshold would lead to an ambiguous classification.

Deciphering whether there exists an acceptable classification is not a trivial task and different approximations have been proposed, but our formalism provides a clear picture of the problem and its relation with complex networks. For instance, the fact that the set $\{\alpha_1, \alpha_2, \alpha_3\}$ is not closed can be understood in terms of the high betweeness [34] of the sequence $\{\alpha_3\}$, i.e. there is a large number of paths joining $\{\alpha_1, \alpha_2\}$ and $\{\hat{\alpha}\}$ passing through $\{\alpha_3\}$. Indeed, "cutting" the network in $\{\alpha_3\}$ may be the solution if we use an algorithm for the search of modules such as the one proposed by Newman and Girvan [35], which is based on the interplay between betweeness [34] and the existence of cores.

Another test may be performed invoking the definition of equivalence class. For instance, even if transitivity is enforced once a given threshold is set, we can still test whether we deal with a legitime classification looking at the similarity between the elements in the clusters with respect to those elements that have been left aside. If we consider that $\{\alpha_1, \alpha_2\}$ is a cluster transitivity holds internally, but it seems also consistent with respect to the outer elements because $S(\alpha_1, \alpha_i) = S(\alpha_2, \alpha_i)$, $(\forall i \neq 1, 2)$. On the other hand, if we additionally include in the same cluster to $\{\alpha_3\}$ we observe that $S(\alpha_3, \alpha_i)$ is not always equal to $S(\alpha_1, \alpha_i)$ (nor to $S(\alpha_2, \alpha_i)$ $(\forall i \neq 1, 2, 3)$), what may be considered an indirect evidence of a transitivity violation within the class (indirect because we perform the test with the elements that are outside the class) [36].

Therefore the existence of disjoint clusters in a network together with the number of internal links – and ideally the definition of cliques [34]– is concomitant with the notion of closed sets and in turn with the absence of vagueness in the classification. Moreover, considering the relationships between the elements in the set and those outside the set may allow us to evaluate the quality of our selection. But it is important to remind that we have started with a reasonable definition in terms of intension of our entities (at least for the three entities of interest). In this way, the problems we find to establish an aritmomorphic scheme arise in this case from an extensionalist vagueness.

We observe that the problem of classification is similar to the determination of system boundaries in a complex system, where there are particular difficulties in biological systems. The reasons for these difficulties are, first, that living beings are open systems that exchange energy and matter with the environment. Thus it is not easy to individuate the boundaries of the system when there is not a clear physical separation with the environment –such as a membrane in the cells– [15].

In addition, what we consider the environment of the system –containing both biotic and abiotic entities–, is sensitive to the system's behaviour. Thus, even if the physical boundaries of the system are well established, when the environment shows a relevant response according to the system's dynamics, an analysis of the system excluding the environment may be unfeasible, and we may consider that our description of the system should be expanded to explicitly include some parts of the environment.

The approximation to this problem is equivalent to the one followed for the classification problem: we look for an objective threshold where the similarity between objects sharply determine equivalence classes. But, in this case, instead of looking for similarities between the particular values of the variables describing the objects, we should look for strengths in the interactions between components. Looking for interactions we focus on how the values of the variables change rather than in the particular values, and we may say that a system is sharply defined if the strength of the interactions between its components is significantly higher than the interactions with respect

to those objects that are left aside in the definition of system. If this description is not achieved, we will observe a picture similar to the one found for an ecosystem, where the boundaries are difficult to individuate and it should be considered a *continuum* of organisms and processes.

3.1.3 An example: The protein fold.

We would like to show now a real example mapping the toy example introduced above, where the existence of ambiguities in the definition of a classification has motivated interesting research. Let us consider that the concepts defined in the previous example correspond to sets of coordinates of the $\alpha-$carbons of the proteins considered, i.e. a coarse grained description of their structure. Under this representation, sharing a concept means that we observe, for at least a couple of proteins, a region of consecutive $\alpha-$carbons (determined by their cartesian coordinates) arranged similarly in the space, where the equivalence has been properly determined with an structural alignment algorithm [37]. The question we aim to address is whether it is possible to find well defined clusters of proteins sharing structural similarities, clusters that are known as protein folds[2].

A canonical definition of fold was provided by Alexey Murzin [39] saying that two protein domains belong to the same fold if they share "the same major number and direction of secondary structures with a same connectivity". This definition is rather loose, because it still needs to specify which secondary structures must be considered *major* and which are simply embellishments, otherwise it would open the door to rather subjective relationships. In our previous example, these embellishments would correspond for instance to the concepts δ_i that were helpful to determine each single object but could be neglected when looking for concepts mapping a larger scope. In figure 3 we show two examples of real proteins illustrating the definition of protein fold, a concept that has motivated the development of projects aiming to classify protein structures [40, 39, 41].

Is it justified to expect the existence of protein folds? The answer is yes, at least in principle. For instance, the molecular clock hypothesis recognized that, after gene duplication, protein sequences accumulate amino acid substitutions almost linearly in time. This means that, if we consider a gene duplication event leading to two different branches a and b and, after this event, we observe a new duplication event in the branch a leading to two sub-branches a_1 and a_2, we expect that the similarity between these genes fullfills the relation $S(a_1, b) \approx S(a_2, b)$. What we obtain is a transitive relation between a_1, a_2 and b, and we can say that these genes belong to the same equivalence class. Indeed, we remind that the representation of homologous genes is typically a phylogenetic tree which, in an ideal scenario of constant substitution rates it becomes ultrametric [44], and thus its analysis would lead to an unambiguous classification.

In addition, it was shown in a seminal paper by Chothia and Lesk [45], that the protein structures diverge linearly in time, what was further confirmed and quantified with larger sets of proteins for different families [46]. Thus it may be expected that the space of protein structure similarites ressembles the properties found in the space of protein sequence similarities. The analysis of the protein structure space [47, 36], together with computational modelization of in silico evolutionary processes [47, 48], seem to provide support to the existence of a finite number of protein folds where any new structure should be included [49].

However, there are more evolutionary events to take into account apart from gene duplications and substitutions, such as gene insertions or deletions. These events, although less frequent [50], may have dramatic effects on the protein structures [51], increasing the number of local similarities shared between proteins belonging to different putative folds and in turn becoming any attempt of

[2]We actually talk about protein domains, which are the minimum units which autonomously function and evolve. Identifying the structural domains for a given protein is the aim of the domain decomposition problem [38].

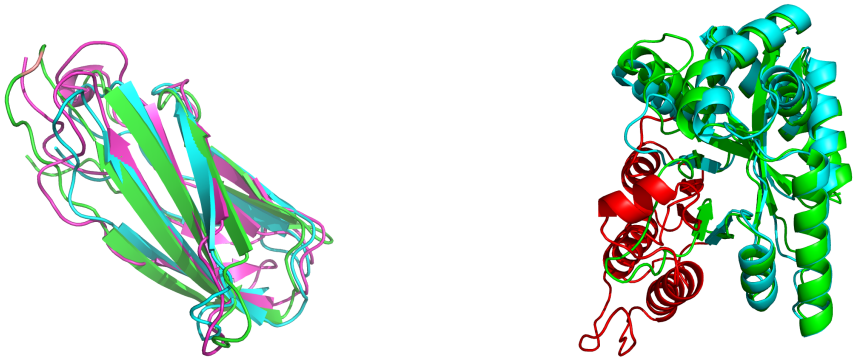


Figure 3: (Left). Example of three proteins with pdb id. 1ttf (green), 1ten (cyan) and 1cfb (pink) belonging to the inmunoglobulin CATH superfamily. This superfamily has a low structural diversity and, in this case, it seems justified to consider that all three proteins belong to the same fold. (Right). The Bacterial Luciferase (1luc, in green and red) is represented with the Nonfluorescent Flavoprotein (1nfp, cyan). Although both proteins are homologs (with a sequence identity around the 30%), and they share an important fraction of their structure, the Nonfluorescent Flavoprotein present a large deletion of the region depicted in red in the Bacterial Luciferase. Determining whether both share the same fold or not with objective arguments would be a complex task where the whole protein structure space should be considered. All the proteins have been aligned with Mammoth Multiple [42] and represented with Pymol [43].

classification frustrated. This fact, together with a systematic finding of short regions of proteins repeated among different folds [52] –in many cases functionally meaningful [53] and apparently leading to a dictionary of short motifs that may even cover any protein structure [52]–, also justifies the view claiming that the protein structure space should be considered continuous rather than discrete [54].

Although one may think that the origin of this discussion stands on an intensional vagueness associated to the determination of protein structures –given the multiplicity of evolutionary events, the dynamical nature of the protein structures or the, sometimes low, experimental resolution–, these difficulties still remain when a non redundant set of representative structures is considered [36]. Thus, this discussion clearly stands on the extensional vagueness, as we have formally defined in the conceptual setting we introduced. Any attempt for solving this discussion should consider objective approximations as those suggested in our toy example, and we find some attempts in the literature based on concepts such as the identification of a percolation transition [47] or the control of transitivity violations [36], providing objective arguments to evaluate the extent by which we can accept a discrete or a continuous view of the protein structure space [32].

The central biological result underlying this example, namely the existence of favored global and local structural regions in the evolutionary process that may be related with the dominance of either gene duplication or more dramatic events, can be re-read pointing towards the existence of constraints, i.e. regions of the fold and sequence space that are not visited by the evolutionary processes. We motivate this negative relecture appealing to a falsifiable positioning: we will work starting from a model where the system is free of constraints and we will progressively incorporate constraints –that will lead to the different null hypothesis–, that become in this way more and more stringent and that we will attempt to reject.

This epistemological positioning highlights the importance of the identification of constraints as an starting point in the modeling process. And, in particular when we deal with complex biological systems, these constrains arise from the interplay between physics –analysed in the previous example from the ensemble of protein structures–, and evolution –that was given by the protein sequences–, to maintain the biological function. Moreover, understanding this interplay has motivated the development of new models of protein evolution that have "bring back molecules into evolution" [55]. Therefore, in the following sections we will focus in the epistemological problem arising around constraints identification and its linkage with interesting questions such as the detection of emergent properties.

3.2 Identification of constraints

Computational modeling aims first to reproduce observed patterns through *in silico* experiments and to further predict new ones that may be testable in *wet labs*, being this last step critical within a scientific framework [29]. Building a formal model requires as a preliminary step the identification of both the viable values and the constraints acting on the observed system. This will be our focus in this section and we will not deal with other problems arising after this step, such as observability [56].

The identification of viable values requires some prior knowledge on the performance of the system in the absence of constraints. For instance, in the case of protein sequences, we know which are the natural aminoacids or the typical lengths. In the case of trophic chains we know which are the kind of interactions expected between any pair of animals considered. In this way, a constraint should be understood as a set of values, belonging to any of the variables we have identified as bein relevant in the description of the system, that are limited in the experimental data.

The procedure to identify these constraints relies on the application of disjunction over those concepts describing the microstates. Looking for common properties in the ensemble, we will decipher the biases that the performance of the system has with respect to its expected behaviour when the constraints are absent. And the other way around, if some properties are found more frequently than it was expected, a negative reading of this result leads to identify which values are not observed. In this way, using the formalism followed, we will express the different constraints talking about their effects, namely identifying any concept c such that it describes a viable value of the system which extension is the emptyset, $Ext(c) = \emptyset$, i.e. it is not observed.

3.2.1 The three bits system

The toy model we are going to consider consists on a system of three entities which physical state is described by a single binary variable, i.e. a system modelled with only three bits. From an experimental point of view, there may be three distinguishable entities (from a molecule to a population) described with binary variables. We can think in sets of genes that are considered expressed (not expressed) when the amount of the correspondent protein is above (below) certain threshold, species observed (absent) in certain environmental sample or, from a strictly computational experiment, the attractor of a boolean network. Each measurement performed over these entities will be considered an observation, and each of them may take a value of one or zero. For a system composed by three entities we can observe, given that there are no constraints, $2^3 = 8$ microstates $\mu_k = (x_1, x_2, x_3)$ associated to a three bits system (with $k = 1, .., 8$; and $x_i = \{0, 1\}$; see table 1). If the three entities can be distinguished we are interested in the following focus:

$c_1 =' ON\ at\ object\ 1'$; $d_1 =' OFF\ at\ object\ 1'$;

Microstate	
$\mu_1 = (1,1,1)$	$\mu_5 = (0,1,1)$
$\mu_2 = (1,0,0)$	$\mu_6 = (1,0,1)$
$\mu_3 = (0,1,0)$	$\mu_7 = (1,1,0)$
$\mu_4 = (0,0,1)$	$\mu_8 = (0,0,0)$

Table 1: Three bits microstates for a free system.

Microstate
$\mu_1 = (1,1,1)$
$\mu_2 = (1,0,0)$
$\mu_6 = (1,0,1)$
$\mu_7 = (1,1,0)$

Table 2: Three bits microstates for a system with a single constraint of scope one

$c_2 =' ON\ at\ object\ 2'$; $d_2 =' OFF\ at\ object\ 2'$;

$c_3 =' ON\ at\ object\ 3'$; $d_3 =' OFF\ at\ object\ 3'$;

Each microstate is defined in terms of this focus through concepts e_k $(k = 1, ..., 8)$ built by conjuction of characteristics. For instance, the microstate $\mu_7 = (1,1,0)$ is defined in terms of the basic characteristics as $e_7 = c_1 \wedge c_2 \wedge d_3$, being in turn a concept.

We can define for this system $\binom{N}{n_{ext}}$ possible combinations of constraints involving n_{ext} variables, and thus the number of final microstates will depend on the number of constraints and their scope, i.e. the number of components influenced by the constraint. In the following, we are going to consider examples with a different number and type of constraints, all resulting in the same number of microstastes (four out of the eight viable states). We aim to explore the epistemological procedure we would apply to get access to the constraints in the system following our formalism. Given that we build our conceptual setting starting from the basic characteristics measured on the system and then performing binary logical operations, we expect that the results obtained for the different systems are fairly comparable –in terms, for instance of the informational content of the propositions found–.

System with a single constraint of scope one. The first system we consider is a system where the first bit is constrained to a fixed value (c_1), leading to the following observations: $\{\mu_1, \mu_2, \mu_6, \mu_7\}$ that we explicitly show in table 2.

In order to find the system constraints we can start comparing the concepts e_i, which determine the different microstates, by disjunction. We provide a compact representation with a network in Fig. 4 where each concept e_i is linked with a concept e_j if they share a basic concept, c or d. Although the constraints determine the microstates, these act on the variables so that we need one step further to identify them. Then, we move from a network of microstates to a network of basic concepts, and we link two concepts c_i or d_i if they extend on the same microstates (see Fig. 4). More formally, we link two concepts c_i and c_j with a directed edge if $Ext(c_i) \subseteq Ext(c_j)$, and with an undirected edge if $Ext(c_i) \cap Ext(c_j) \neq \emptyset$. In this way we compactly represent all the dependencies present in the system, resembling for instance the information that we would recover if we build a network of variables from a covariance matrix: positive (negative) correlation arises

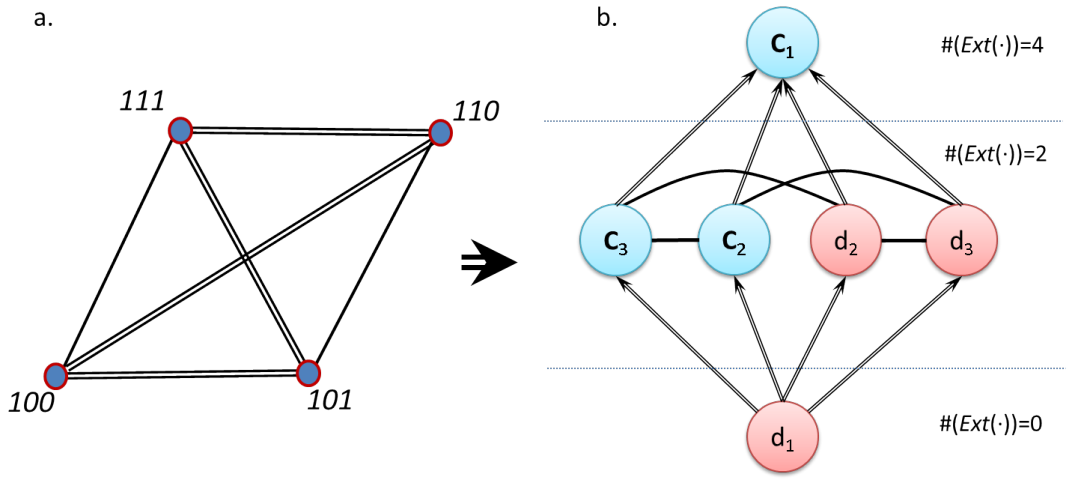


Figure 4: Graph of microstates for a three bits system with a single constraint of scope one (left). Each node represent a microstate and it is linked with another microstate if they share the same observation for any component,where the number of links represent the number of concepts shared. (Right) Graph of the concepts extracted from the analysis of the microstates. Two links c_i and c_j are linked with a directed edge if $Ext(c_i) \subseteq Ext(c_j)$ and with an undirected link if $Ext(c_i) \cap Ext(c_j) \neq \emptyset$. The concepts are hierarchically ordered by the number of microstates they map, naturally arising the single constraint on d_1.

when similar (dissimilar) values are found between two objects. This fact is represented in our representation by subordination (directed edges) coocurrence (undirected) or exclusion (absent link).

From this second network it is easy to observe that one of the values of the first variable, d_1, is never observed, a fact that we can express with the proposition:

$$Ext(d_1) = \emptyset$$

We confirm in this way that the system is topologically representable and, what this proposition simply states is that, in order to identify that a given microstate belongs to this system, it is necessary to evaluate that the value measured on the first component of the system is different from zero.

A convenient description of the ensemble is given by its probability distribution:

$$P(\mu) = \delta(x_1, 1)/2^{n+1}$$

where δ(a,b) is the Kronecker's delta and n is the number of bits.

System with two constraints of scope two. We select now four microstates that are obtained imposing one constraint among each pair of variables. Taking the microstates $\{\mu_1, \mu_4, \mu_5, \mu_8\}$ that are explicitly shown in table 3, and repeating the procedure done in the previous example (see Fig. 5), we observe that the disconnected components in the graph lead to the following constraints, which can be expressed with the propositions:

Microstate
$\mu_1 = (1,1,1)$
$\mu_4 = (0,0,1)$
$\mu_5 = (0,1,1)$
$\mu_8 = (0,0,0)$

Table 3: Three bits microstates for a system with two constraints of scope two.

Microstate
$\mu_1 = (1,1,1)$
$\mu_2 = (1,0,0)$
$\mu_3 = (0,1,0)$
$\mu_4 = (0,0,1)$

Table 4: Three bits microstates for a system with a single constraint of scope three.

$$\begin{aligned} Ext(c_1 \wedge d_2) &= \emptyset \\ Ext(c_2 \wedge d_3) &= \emptyset \\ Ext(c_1 \wedge d_3) &= \emptyset \end{aligned}$$

It is easy to observe that one of these constraints is redundant. Given that c_2 and d_2 cannot be observed simultaneously, if c_1 is observed it means that $c_2$is also observed and thus d_3 cannot be observed. And the other way around, if d_3 is observed c_2 will not be observed and thus c_1 cannot be observed. Therefore, the third constraint $Ext(c_1 \wedge d_3) = \emptyset$, is a consequence of the other two. The identification of these constraints allow us to write down the probability distribution of the ensemble as:

$$P(\mu) = \delta(\delta(H_{12} + 1, 1), H_{23} + 1)/2^{n+1}$$

where $H_{ij} = H(x_i - x_j - 1)$ is the Heaviside function, which reflects the interplay between the variables, and introducing the delta function we sequentially impose the constraints. Note that the redundancy of the third proposition can be also seen if we replace any of the Heaviside functions by H_{13}, as we do not recover the observed ensemble but a larger one.

System with a single constraint of scope three (the parity bit system). Our last example consists on the set of microstates having an even number of ON bits, i.e. a single constraint involving all three components. This system has been previously introduced by Bar-Yam as a toy example of a particular type of emergent behaviour called *strong emergence,* that will be discussed in detail later [57]. For this system, given that we find two random values in two randomly selected bits, the third bit is constrained in such a way that the number of bits in the microstate is always even. This rule is used in the control of message transmission, where the last bit (called the parity bit) is used to monitor the presence of errors in the chain transmitted. The microstates we will consider are $\{\mu_1, \mu_2, \mu_3, \mu_4\}$, explicitly shown in table 4

In this case the graph of concepts intuitively resembles an sphere in the sense that there are not "borders" –i.e. disconnected concepts from which propositions about the constraints are simply derived (see figure 6)–. Thus, the identification of constraints is possible only because we already know the viable values. Indeed, a parallel analysis of the free system highlights a lower coocurrence

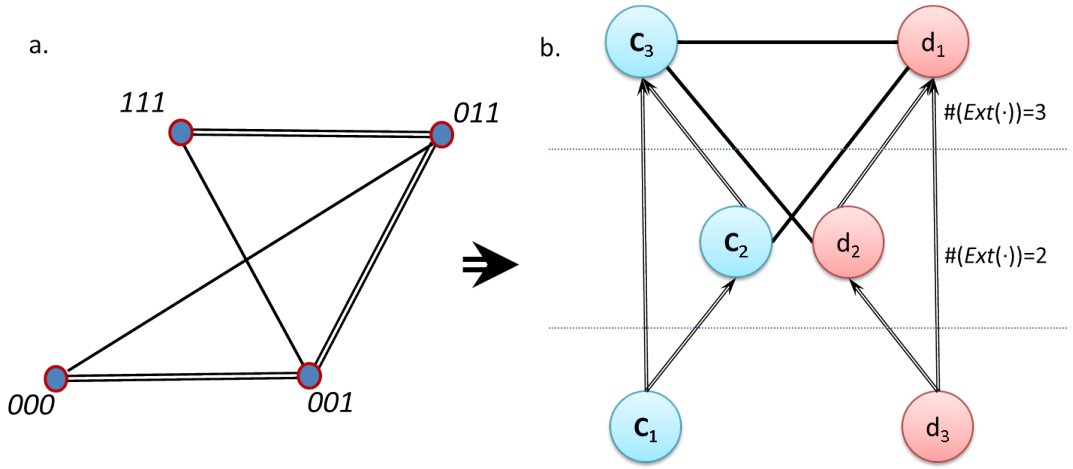


Figure 5: Graph of microstates for a three bits system with a three constraints of extension two (left). Each node represent a microstate and it is linked with another microstate if they share the same observation for each component. (Right) Graph of the concepts extracted from the analysis of the microstates. Two links c_i and c_j are linked with a directed edge if $Ext(c_i) \subseteq Ext(c_j)$ and with an undirected link if $Ext(c_i) \cap Ext(c_j) \neq \emptyset$. The concepts are hierarchically ordered by the number of microstates they map. We identify the constraints observing those links that being viable are absent, for instance there is no link between d_3 and c_2.

of the different variable values but there will be no differences in the final graph we obtain. The constraints cannot be inferred from the system itself but rather by comparison with respect to the free system. The comparison with the free system bring to the surface the following propositions:

$$\begin{aligned} Ext(d_1 \wedge c_2 \wedge c_3) &= \emptyset \\ Ext(c_1 \wedge d_2 \wedge c_3) &= \emptyset \\ Ext(c_1 \wedge c_2 \wedge d_3) &= \emptyset \\ Ext(d_1 \wedge d_2 \wedge d_3) &= \emptyset \end{aligned}$$

And what we observe is that the most compact way to talk about this system within this formalism is to write down all the microstates that are *not* observed. From these constraints we have no clue to write down the probability distribution of the ensemble that we obtain with a rather *ad hoc* definition:

$$P(\mu) = \delta(mod_2(\sum_i x_i), 1)/2^{n+1}$$

where $mod_2(\cdot)$ is the module two function that allow us to test whether the sum of the bits is odd or even.

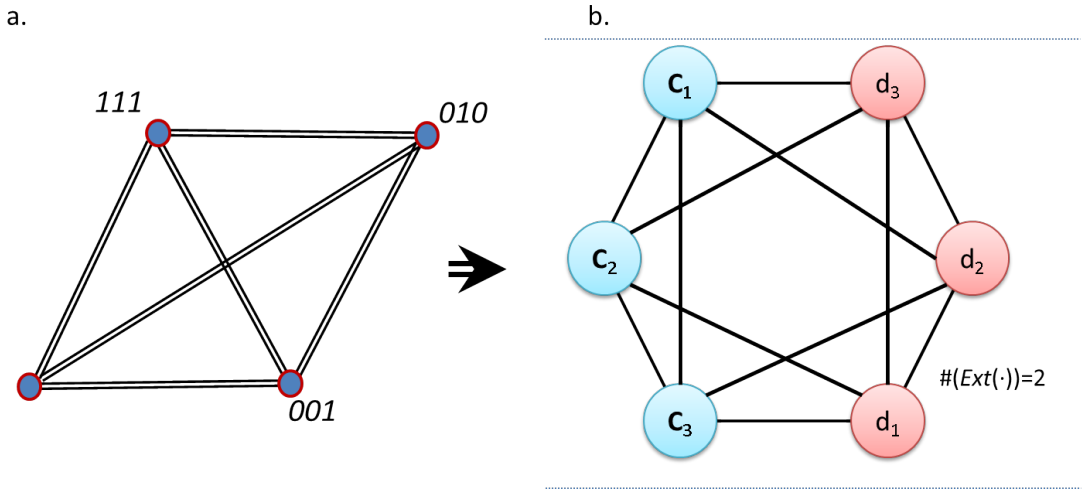


Figure 6: Graph of microstates for a three bits system with a single global constraint (left). Each node represent a microstate and its linked with another microstate if they share the same observation for each component. (Right) Graph of the concepts extracted from the analysis of the microstates. Two links c_i and c_j are linked with a directed edge if $Ext(c_i) \subseteq Ext(c_j)$ and with an undirected link if $Ext(c_i) \cap Ext(c_j) \neq \emptyset$. In this example we can just identify the existence of constraints because we know the space of viable values. The graph of concepts is equivalent to the graph we would obtain for a free system, being just observed a reduction in the number of objects mapped by each concept (from $\#Ext(\cdot)$=4 towards $\#Ext(\cdot)$=2).

3.3 Emergence

3.3.1 Weak and strong emergence

One of the most important questions being a matter of controversy in the analysis of emergent properties, is whether there exist emergent behaviours that are not epistemologically accessible, being probably the consciousness its most paradigmatic example. To answer this question we observe that there should be still a previous question to be answered, namely how the epistemological accessibility can be quantified. Bedau proposed that this question could be solved if we consider that an epistemologically accessible pattern should be reproduced with a computational model, and he coined for the process leading to such patterns the term weak emergence [24].

Another interesting notion of emergence that has been considered fundamental as opposed to epistemological, has been called *strong emergence* [57]. In Physics it is accepted that knowing the positions and velocities of particles is sufficient to determine the pairwise interactions. This assumption is frequently found in Physics-inspired models of collective behaviour, where individual motion results from averaging responses to each neighbour considered separately. Nevertheless, Bar-Yam argues that this assertion would not hold if the system is embedded in responsive media, such as the motions of impurities embedded in a solid, or further in any process where global optimization (instead of local) is involved.

In this way, if there is in the system a constraint acting on all the components simultaneously and it is strong enough –i.e. it is a global constraint–, it is not possible to determine the state of the system considering only pairwise interactions. In some sense the parts are determined *downward* from the state of the whole, and the toy example he provides is precisely the parity bit system we analysed in our previous examples.

The advantage of the formalism we introduced is that it allow us to explore from (logical) *first principles* the constraints present in the system. Furthermore, given that we are limited to the consideration of few binary operations, it is possible to fairly compare the complexity of the constraints present in the different systems from the number of propositions found and the number of concepts involved, what can be easily quantified in terms of the information content. In particular, we observe that larger is the scope and/or the number of constraints, more difficult will be to describe our system.

From the examples depicted, we have seen that in the first two examples, it is possible to identify the constraints and from them we have proposed a model of the system through a probability distribution, where each function used (either the Kronecker δ or the Heaviside function) have a clear parallelism with the propositions found, in the sense that a sequencial application of these functions allow us to build the description of the model following a bottom-up process. In some sense, the top-down analysis of the system allowed us to propose a bottom-up process to reproduce the observed behaviour. Therefore, from this model it would be easy to simulate the observed behaviour and we would say –following the ideas of Bedau– that the system is epistemologically accessible.

For the parity bit system, on the other hand, it was not possible to identify any proposition that would allow us to express the constraints involved. Indeed, the propositions found have the same information content than the definition of the system itself, and are rather of doubtful utility. In addition, it is difficult to observe a clear parallelism between the probabilistic model proposed and the propositions, we have provided instead an *ad hoc* definition that requires the global evaluation of each microstate, which is difficult to see as a bottom-up process to describe this behaviour.

In summary, we observe that when a global constraint is present in the system, the top-down approach we follow to discover the constraints fails to recover propositions from which we could

build bottom-up models to reproduce the pattern. In this sense, it seems that the notion of strong emergence is justified, maybe not as a kind of system which is not epistemologically accessible but rather as a limiting case of weak emergence, where a bottom-up approximation to simulate the system is not valid. We will discuss below which are the consequences of this fact in the simulation of complex systems.

3.3.2 Quantification of emergent behaviours: scaling, intervention and Granger causality

The next question we aim to address is whether there exist any procedure to quantify the strength of any observed emergent behaviour. From the formalism we followed a natural choice arise if we consider the information needed to describe the system $I(syst)$ and the different constraints $I(cons)$ in terms of number of propositions and the number of concepts contained, i.e. number of constraints and their scope. In this way, we propose as a definition of emergence strength ES:

$$ES = \frac{I(cons)}{I(syst) - I(cons)} \tag{9}$$

Following this definition, if the constraints in the system can be characterized with a low amount of information $I(cons) \ll I(syst)$ and $ES \to 0$. On the other hand, if we need to describe constraints as much is information as we need to describe the system $I(cons) \approx I(syst)$ then $ES \to \infty$, reflecting the notion of epistemological inaccessibility claimed for strong emergence.

Note that, with this definition, any collective behaviour can be understood as an emergent behaviour and, indeed, it is natural to say that it is. What differentiate one process from others is how difficult is to characterize the process and the processes we typically understand as emergent processes should have a large emergence strength.

Nevertheless, with this definition there are still several questions opened. First, which is the appropriated description of the system over which this information should be quantified. Second, how we know that we have achieved a complete characterization of the constraints of the system and, as a consequence, third, which is the experimental procedure needed to achieve such characterization.

The first question following our formalism seems easy to be answered, because we use similar propositions for defining the system and expressing the constraints, and these are therefore easily comparable. However, it is difficult to know whether this kind of approximation would be helpful for real systems. A more realistic possibility could arise from the evaluation of the algorithmic information content [58] of a program containing the expression of the constraints. For the parity bit system, such a program would require the inspection of all the variables further applying two functions, namely a sum and a division –within the module 2 function–. This program would be algorithmically more complex than the evaluation of the previous examples. For instance, for the system with a single constraint of scope one, we would need to evaluate a single variable with a boolean function, which clearly requires a shorter algorithm.

But a more interesting alternative arises from the observation of the scaling behaviour of the constraints, i.e. how the number of propositions needed to describe the constraints changes when the system size increases. We observe that, when the number of bits N increases, for the first example the number of propositions needed to describe the same constraints remains equal to one. In the second example, this number increases proportionally to N^2, and for the parity bit system it increases as 2^N. This fact is translated not in the complexity of the algorithm we develop to evaluate the microstates, that would be exactly the same, but in the computational time needed to

halt once it has identified the pattern. Whereas this time would be exactly the same for the system with a single constraint of scope one, for the parity bit system we will need as much time as we would need to identify a random chain. In this way the uncertainty we face in this experiment for the parity bit system does not diminish by a previous knowledge of the constraints in the system, whereas for the other systems may be much lower.

This perspective seems to provide an answer to the first question exposed above, because we could consider the emergence strength taking as variables, rather than the information content, the number of components n_{cons} needed to evaluate that a microstate belongs to the ensemble with respect to the total number of components N:

$$ES = \frac{n_{cons}}{N - n_{cons}} \tag{10}$$

equation having the same properties than Eq. 9. In addition, this procedure would allow us to recognise whether we have a complete description of the system in terms of its constraints –as we should recover the ensemble of observed microstates–. The remaining question is, if we still do not deal with a complete description, which experimental procedure we may follow to achieve such a description.

The observation that the scaling behaviour of the constraints seems to be related with the emergence strength may provide an answer. Because there is an artificial manner for changing the size of the system, which is neglecting components, i.e. reducing N. Therefore, we can *intervene* in the system neglecting components and monitor which is the relative change in our knowledge of the system arising from this intervention, that can be measured with the number of constraints lost. Indeed, intervention has been highlighted as a basic ingredient to link computational modeling with the scientific method [59]. Furthermore, the fact that the effects of our intervention can be quantified through the observation of the constraints lost, means that we can relate the causal effects that a component has on other components, which is entirely compatible with the notion of Granger causality.

3.3.3 Traceability

We return to the formalism provided by concrete topology, to formalize the definition of emergence strength integrating this notion within a methodology procedure compatible with the scientific method. We first define what is considered a novel macroscopic property, which identification is typically the starting point of any research.

Definition: (Novel macroscopic property). We will say that an observed macroscopic property is a novel property if it is observed only in the presence of certain constraints limiting the viable values of the system.

Therefore, given that the phase space of the system Ω is restricted to a smaller observed region $\Omega^O \subset \Omega$, what we say is that there exists a macroscopic concept $\hat{c}$ such that $Ext(\hat{c}) = \Omega^O$, and we would like to explore this region both in terms of macroscopic $\{\hat{c}\}$ and microscopic $\{c\}$ concepts. We now introduce the conditions that allow us to consider that a macroscopic description is in correspondence with a microscopic description, leading to the concept of traceability.

Definition: (Traceability). Given a novel macroscopic property $\hat{c}$ describing the observed phase space Ω^O, i.e. $Ext(\hat{c}) = \Omega^O$, we will say that the macroscopic description obtained is traceable if we find an appropriate function or algorithm applied on

microscopic properties $f : \{c\} \rightarrow q$ such that the new concept q derived describes the ensemble of microstates, i.e. $Ext(q) = Ext(\hat{c}) = \Omega^O$.
This definition provides the means by which we can quantify the correspondence between both descriptions within the framework proposed. This definition does not require that we are able to relate macroscopic and microscopic properties, but only to establish a correspondence between microscopic and macroscopic variables describing the same region of the observed phase space Ω^O. This approximation is what allow us to talk about emergent properties circumventing any epistemological discontinuity between both descriptions.

3.3.4 Intervention and loss of traceability

Our next objective is to incorporate the concept of intervention, namely how the phase space of the observed system changes if we actively ignore any of the components of the system. Neglecting variables in the observation process may imply that we are no longer able to identify any of the constraints, and thus we would expect that the system is able to visit a larger region of the phase space, gaining symmetry in the variable values. If we start with a situation where perfect traceability exists this procedure would lead to a region larger than the one observed, being the observed region *covered in excess*. This excess will be a measure of how far we are from perfect traceability of the system, and we formalize it as follows.

Let us start considering a system with N components where has been observed a macroscopic property $\hat{c}$ whose traceability has not been properly determined through a microscopic concept q, thus $\Omega^O = Ext(\hat{c}) \subset Ext(q)$. As we said, this means that we have identified some constraints but we are still not able to establish a perfect map between the microscopic and macroscopic description.

Definition: (Coverage Excess). Given a complex system of N components where a novel macroscopic property is described by $\hat{c}$ and whose traceability is being approximated with the microscopic concept q, we call coverage excess to the fraction of the phase space covered by q which is not representative of the novel macroscopic property:

$$CE = \frac{\#(Ext(q)) - \#(\Omega^O)}{\#(\Omega) - \#(\Omega^O)} \tag{11}$$

Where in this case the function $\#(x)$ returns the number of objects (not the number of concepts) contained in the set x. Following this definition we get a number between zero and one where, if the microscopic property maps the whole ensemble, there will be no coverage excess and thus $CE = 0$. But, if we are not able to find even a single property closely mapping the ensemble, the coverage excess will be maximal and $CE = 1$. From this definition it is immediate to define the degree of traceability of the system.

Definition: (Traceability Degree). Given a complex system of N components where a novel macroscopic property is described by the macroscopic property $\hat{c}$ and whose traceability is being approximated with the microscopic concept q, we determine its degree of traceability with the magnitude:

$$Trace = 1 - CE \tag{12}$$

In this way we achieve perfect traceability when there is no coverage excess ($Trace = 1$) and no traceability at all when the coverage excess is maximal and $Trace = 0$.

Next, we aim to quantify the loss of traceability we experience when the accessibility to the system is lost in some degree. To address this task, we are going to consider that we are able to obtain information from S components, a magnitude that we call the *sampling effort* of the experiment, verifying that $S \leq N$. This change may be enforced by any change in the experimental conditions, or may be voluntarily generated through the intervention of the observer over the information provided by the system. Assuming that we have achieved some degree of traceability, we are interested in understanding whether the traceability is lost when we move to a situation where either the number of components increases $N \rightarrow N'$ with $N' > N$, or the sampling effort decreases $S \rightarrow S'$, with $S' < S$. The idea is that, under this new scenario and even if we are able to derive a new concept q' from the analysis of the microstates, it will typically happen that $Ext(q) \subset Ext(q')$, because we will lose some constraints in our analysis.

Definition: (Relative Loss of Traceability). Consider a complex system of N components and sampling effort S, with a given degree of traceability that has been determined through the microscopic concept q that we assume it covers in excess the observed phase space by a quantity CE. Let us further consider that there is a change in either the number of components $N \rightarrow N'$, the sampling effort $S \rightarrow S'$, or both, i.e. $(N' - N) + (S - S') = \Delta N + \Delta S \geq 1$ and, under this new scenario, we identify a new microscopic property q' that covers in excess the phase space by a quantity CE'. We measure the relative loss of traceability of the system due to this change with the magnitude:

$$rLT = \frac{CE' - CE}{\Delta N + \Delta S} = \frac{1}{\Delta N + \Delta S} \frac{\#(Ext(q')) - \#(Ext(q)))}{\#(\Omega) - \#(\Omega^O)} \tag{13}$$

With the above expression we measure the relative coverage excess increase for a given change in the variables controlling our access to the system, N and S. But this measure still depends on our current state of knowledge of the system. Thus we aim to establish an absolute value that can be associated to the novel macroscopic property.

Definition: (Absolute Loss of Traceability). We call absolute loss of traceability aLT to the relative loss of traceability obtained when there is perfect traceability, $Trace = 1$, and we observe an unitary change in the variables N and S, i.e. $\Delta N + \Delta S = 1$. Given that this measure may be still sensitive to the specific components over which we intervene, we should consider a systematic procedure where we compute the measure for different realizations of our intervention. For instance, in the above examples around the three bit systems, let us consider that we decide to intervene over a randomly selected bit. In the first system, the one with a single constraint of scope one, we will completely lose traceability each time we neglect the variable over which the constraint is applied, but we will have no loss of traceability when we neglect any of the other variables, and thus on average $aLT = 1/3$. For the second system, where we find pairwise constraints, irrespective of the variable neglected we will expand the microstates from four to six and thus $aLT = 1/2$. Last, for the parity bit system we find that the loss of traceability is maximal for any variable selected $aLT = 1$.

3.3.5 Traceability and emergence strength

From the definition of aLT it is immediate to propose a positive definition.

Definition: (Traceability strength). We define the traceability strength associated to a novel macroscopic property as

$$TS = 1 - aLT \tag{14}$$

In this way we simply state that the traceability strength is obtained estimating the maximum loss of traceability observed in a perfectly traceable system, which is a condition to assert that our measure is associated to the macroscopic property, after an unitary change. We are already equipped to introduce a class of novel macroscopic properties particularly interesting in complex systems, namely the emergent properties.

Definition: (Emergent property). We will say that a novel macroscopic property is an emergent property if its emergence strength ES is significantly different from zero where:

$$ES = \frac{1 - TS}{TS} \tag{15}$$

This equation has the same behaviour than the equations 9 and 10: emergent behaviours with a loss of traceability close to one will have a very high emergence strength (infinity if traceability has been completely lost) and will be close to zero if there is almost no loss of traceability, what justifies its exclusion from being classified as an emergent property. In our examples we get $ES = 0.5$ for the system with a single constrain of scope one, $ES = 1$ for the system with three constraints of scope two and $ES = \infty$ for the parity bit system.

Following this definition we deal with the problem of emergence maintaining ourselves within an epistemological framework, given that it is determined through changes in our experimental setting, i.e. the sampling effort or the system size. In addition, we make expressive the limiting case where we have no access to the underlying constraints in the system, which is understood in this framework invoking to a lack of information. The current discussion around the definition of weak and strong emergent properties [57] can be rationalized using the emergent strength: if the emergence strength is infinite we deal with a strongly emergent pattern whereas if the value is finite we deal with a weak emergent pattern with the associated strength as an indicator of its traceability.

Note that we have also included the term *significant*, emphasizing that this measure requires an statistical experimental setup, although this that not mean that they will not be observed in a single microstate after its identification. For instance, it has been claimed that flocking is not observed in a single microstate [14]. However, once we have characterized flocking with microscopic variables, for example once we derive the pairwise distribution of interactions among neighbours [60] we can also test whether any behaviour, even locally, is consistent with the model derived. This fact already provides the means to build an hypothesis testing experiment where the null hypothesis is that the group of birds do not present flocking behaviour.

3.3.6 Emergence and vagueness

The coexistence of the microscopic and macroscopic descriptions can be problematic when perfect traceability is not achieved. Moreover, in this section we aim to show that a novel macroscopic property is vague unless traceability is achieved. Let us start considering the following topologies

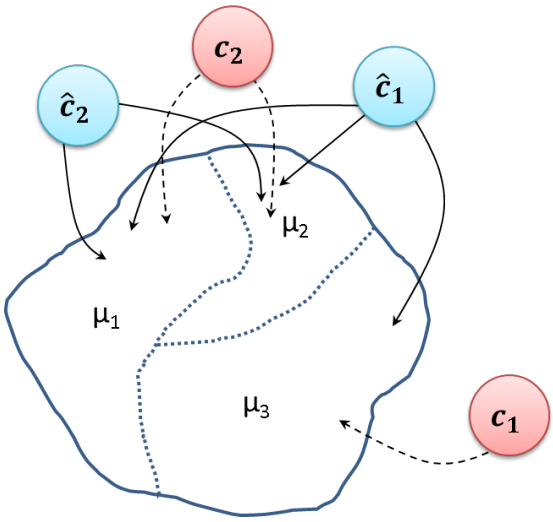


Figure 7: Simultaneous representation of the macroscopic and microscopic descriptions (See text for details).

described by macroscopic concepts $\hat{c}_0$, $\hat{c}_1$ and $\hat{c}_2$ for a system composed by three sets of microstates μ_1, μ_2 and μ_3 (see Fig. 7).

In the first topology, the system is free of constraints and the extension of the concepts are:

$$\begin{aligned} Ext(\hat{c}_0) = & \quad \emptyset \\ Ext(\hat{c}_1) = & \quad \{\mu_1, \mu_2, \mu_3\} \end{aligned}$$

whereas in the second case, there is some constraint limiting the observation of μ_3, and a new macroscopic property $\hat{c}_2$ arises:

$$\begin{aligned} Ext(\hat{c}_0) = & \quad \emptyset \\ Ext(\hat{c}_2) = & \quad \{\mu_1, \mu_2\} \end{aligned}$$

It is important to observe here that it is not possible to simultaneously observe both behaviours, either we observe $\hat{c}_1$ or we observe $\hat{c}_2$. However, we can say that there is a novel behaviour just if we know the viable values of the system in the absence of constraints, and thus it seems to be justified the consideration of both topologies simultaneously to reach a complete understanding of the system:

$$\begin{aligned} Ext(\hat{c}_0) = & \quad \emptyset \\ Ext(\hat{c}_1) = & \quad \{\mu_1, \mu_2, \mu_3\} \\ Ext(\hat{c}_2) = & \quad \{\mu_1, \mu_2\} \end{aligned}$$

Analysing this final topology we observe that, apart from the empty set and the whole set, there is only one more open, which is $\{\mu_1, \mu_2\}$. And we observe that the closure of $\{\mu_1, \mu_2\}$ is $\{\mu_1, \mu_2, \mu_3\}$, and thus its border is $\{\mu_3\}$, from which we can say that the concept $\hat{c}_2$ describing the novel macroscopic property is vague.

Next, let us consider a microscopic description. As we have seen, the analysis of the microscopic description requires looking for a characteristic property belonging to the ensemble of microstates of

the phase space associated to the novel macroscopic property. Furthermore, there is an interesting observation which is that, once we have identified a relevant property, this property is associated to the microstates even if we observe the system free of constraints, whereas in the macroscopic description the novel property is observed only when the constraints are present. Therefore, what differentiates one macroscopic behaviour from the other is not a change in the microscopic properties but rather the set of microstates observed, and thus the microscopic description remains invariant. Following the example, if there is traceability we will deal with a microscopic description such as:

$$\begin{aligned} Ext(c_0) = &\quad \emptyset \\ Ext(c_1) = &\quad \{\mu_3\} \\ Ext(c_2) = &\quad \{\mu_1, \mu_2\} \end{aligned}$$

where, what is meant, is that the microscopic property selected returns a value c_2 for those microstates represented in $\{\mu_1, \mu_2\}$ and that, for those microstates not belonging to the phase space observed in the presence of constraints, a value of c_1 is obtained. In this case $\{\mu_1, \mu_2\}$ and $\{\mu_3\}$ are sharply separated and the interesting observation arising here is that both $\hat{c}_2$ and c_2 describe the same region of the phase space but the former is a vague concept. Nevertheless, as soon as we simultaneously consider the macroscopic and microscopic description and traceability is proved, the vagueness in the concept $\hat{c}_2$ vanishes by means of its correspondence with c_2, making possible the determination of the emergence strength for the novel property.

4 Conclusions

The analysis of complex systems are nowadays increasingly considered with central questions such as complexity [61], emergence [24], autopoiesis [15] or self-organization [62], becoming more and more present in the literature. However, little attention has been made to the epistemological challenges that these questions involve in scientific research. Here, we have introduced a novel link between concrete topology and complex systems –with particular emphasis in biological systems– in order to provide a more expressive analysis of these epistemological problems, where vagueness has been identified at the heart of these problems.

First, we have related the concept of dimensionality reduction with concept disjunction. We showed that, concept disjunction, is the logical operation underlying the development of similarities and distances, and thus it is on the basis of the development of classification schemes. Indeed, we have analysed the problem of classification, which is an important problem in biology [29], and we found that extensional (and not intensional) vagueness is the source of frustration in classification schemes. We have also seen that this problem is related with the determination of system boundaries, although we must focus on the strength of the interactions instead on the similarity between variables.

Next, we wondered whether disjunction allows for identifying constraints when an ensemble of microstates is considered. Using a simple model of three bits, we observed that the number of propositions needed to describe the constraints was related with the number and scope of the constraints. This analysis already allowed us to propose a definition of emergence strength, that integrates the notion of weak [24] and strong emergence [57].

Nevertheless, it seems difficult to apply this measure in real systems, and we explored a practical definition working around a central concept: the traceability of a system, which provides the means to link the macroscopic and microscopic descriptions, and the problems arising from their

coexistence. We have defined the traceability of the macroscopic description as a measure dependent on our ability to determine the intension of a concept describing the microscopic ensemble. In order to provide a definition of emergence, we have assumed perfect traceability and then we have proposed a procedure, based in the intervention of the observer [59], to determine the emergence strength of the system. The relative loss of traceability measured during this process allows for driving any experimental procedure towards capturing the existing constraints, a proposal close to the notion of Granger causality [22]. We close our exposition providing an argument to show that the vagueness associated to the concept of emergence has its origin in the lack of traceability between the macroscopic and microscopic descriptions.

Our analysis readily suggests the proposal of an epistemological program that will be presented elsewhere. Nevertheless, we would like to point out that the concept of strong emergence implies an important challenge in the modelization of complex biological systems. Understanding emergent properties in complex biological systems requires accounting for at least two different spatio-temporal scales. First, an intrinsic scale where those processes considered essential to keep the system far from equilibrium take place, what we call the physical scale. And second, an scale where evolutionary events, that may modify the physical performance, become fixed in the population (the evolutionary scale). This interplay between different spatio-temporal scales represent a characteristic difficulty in the analysis of emergent properties in biological systems, where disentangling the contribution of each kind of process in the observed pattern is often a matter of intense conceptual research (see, for instance, [63, 64])

David Bohm pointed out that, in the earlier stages of any science, the interest is focused on "the basic qualities and properties that define the mode of being of the things treated in that science" [28], being tasks such as comparative analysis and classification the cornerstones in its earlier development. It is just after a sufficient characterization of the entities under study where we will find a growing interest on "processes in which things have become what they are, starting out from what they once were, and in which they continue to change and to become something else in the future" [28], i.e. in the evolutionary scale.

However, it is difficult in Biology to "define the mode of being" of biological entities without taking into account the "processes in which things have become what they are" precisely due to this interplay between the physical and evolutionary scales. The "basic qualities and properties" even when they may be well defined for a given process within a physical scale, should be simultaneously understood within an evolutionary context, what recalls why "nothing in Biology makes sense except in the light of evolution"[65].

In this way, understanding emergent properties in complex biological systems requires to handle both scales, and it may be interpreted that the determination of a given biological state depends on the bottom-up processes taking place in the system, and in the top-down processes selecting the system. Therefore, we could consider that understanding these behaviours requires to simultaneously face the existence of collective behaviours that are both weak and strong emergent. This fact can be viewed in the problem of the determination of the object of selection.

On the one hand, each system may be determined taking into account the strength of the interactions of the different components, as we explained. In this case, we are focusing in the physical scale. On the other hand, the object of selection may be determined taking into account the relative effects that the interactions have on the fitness of individuals. And, it may happen, that there is not a perfect map between the object of selection and the physical systems, because the effect of the interactions in the fitness of the individuals is so strong that their genomes coevolve with individuals of other species [66]. If this were the case, we may observe that the physical systems determine their states in the evolutionary scale *downward*, given that they have strong

coevolutionary forces with other systems. Therefore, any observed collective pattern would reflect the interplat between bottom-up physical behaviours and top-down evolutionary selection.

In summary, our exposition intends to reflect the constant will of human knowledge to handle dialectic concepts within an arithmomorphic scheme. This contraposition between dialectic and arithmomorphic concepts is well reflected in philosophical terms with the metaphor expressed by Nietzsche [1] for the tragedy between Dionysus (dialectics) and Apollo (arithmomorphic). The process by which we are able to reach an arithmomorphic scheme is well described by Hegelian logic, where the thesis, being arithmomorphic, is challenged by the dialectical nature of the processes, generating an antithesis until a new conceptual synthesis is reached.

We hope that our approach contributes to build the epistemological framework required for a health growth of new disciplines such as Systems Biology, where there are several voices claiming for a solid epistemological framework [67, 68]. We would like to remark again that, in these disciplines, the fact that the description of the system is necessarily reduced does not imply that it is a reductionist approach. It is rather a necessary epistemological exercise to deal with complex systems what allows to the scientist to propose general questions, which otherwise would not be possible to handle. We claim that these approaches circumvent those difficulties arising from the study of systems with intrinsically dialectic concepts, opening a door to understand emergent properties and the establishment of general laws. But it would be nothing but time and experiments what will allow us to test both the predictive power of these new approaches and the skepticism.

References

[1] F. Nietzsche, *Thus Spoke Zarathustra*. GoodBook LLC, 2014.

[2] M. Bunge, *Emergence and Convergence: Qualitative Novelty and the Unity of Knowledge*. University of Toronto Press, 2003.

[3] G. Frege, P. T. Geach, and M. Black, "On concept and object," *Mind*, vol. 60, pp. 168–180, Apr. 1951.

[4] F. Cohan, "What are bacterial species?," *Annu Rev Microbiol*, vol. 56, pp. 457–487, 2002.

[5] E. V. Koonin, K. S. Makarova, and L. Aravind, "Horizontal gene transfer in prokaryotes: quantification and classification 1," *Annual Reviews in Microbiology*, vol. 55, no. 1, pp. 709–742, 2001.

[6] W. G. Stock, "Concepts and semantic relations in information science," *Journal of the American Society for Information Science and Technology*, vol. 61, no. 10, pp. 1951–1969, 2010.

[7] S. Strunz, "Is conceptual vagueness an asset? arguments from philosophy of science applied to the concept of resilience," *Ecological Economics*, vol. 76, pp. 112–118, 2012.

[8] A. R. Ives and S. R. Carpenter, "Stability and diversity of ecosystems," *science*, vol. 317, no. 5834, pp. 58–62, 2007.

[9] C. Emmeche, S. KÃžppe, and F. Stjernfelt, "EXPLAINING EMERGENCE: TOWARDS AN ONTOLOGY OF LEVELS," *Journal for General Philosophy of Science*, vol. 28, pp. 83–117, Jan. 1997.

[10] K. G. Wilson and J. Kogut, "The renormalization group and the ϵ expansion," *Physics Reports*, vol. 12, no. 2, pp. 75–199, 1974.

[11] M. Scheffer, J. Bascompte, W. A. Brock, V. Brovkin, S. R. Carpenter, V. Dakos, H. Held, E. H. van Nes, M. Rietkerk, and G. Sugihara, "Early-warning signals for critical transitions," *Nature*, vol. 461, no. 7260, pp. 53–59, 2009.

[12] N. Georgescu-Roegen, *The Entropy Law and the Economic Process*. Harvard University Press, first ed., Jan. 1971.

[13] G. W. F. Hegel and G. D. Giovanni, *Georg Wilhelm Friedrich Hegel: The Science of Logic*. Cambridge University Press, Aug. 2010.

[14] A. J. Ryan, "Emergence is coupled to scope, not level," *Complexity*, vol. 13, no. 2, pp. 67–77, 2007.

[15] H. Maturana, E. Lenneberg, and E. Lenneberg, "{Biology of Language, the Epistomology of Reality}," in {*Foundations of Language Development, a Multidisciplanary approach*}, vol. 2, The UNESCO Press, 1975.

[16] T. R. Alley, "Organism-environment mutuality epistemics, and the concept of an ecological niche," *Synthese*, vol. 65, no. 3, pp. 411–444, 1985.

[17] B. B. Machta, R. Chachra, M. K. Transtrum, and J. P. Sethna, "Parameter space compression underlies emergent theories and predictive models," *Science*, vol. 342, no. 6158, pp. 604–607, 2013.

[18] S. Y. Auyang, *Foundations of complex-system theories: in economics, evolutionary biology, and statistical physics*. Cambridge University Press, 1999.

[19] P. W. Anderson *et al.*, "More is different," *Science*, vol. 177, no. 4047, pp. 393–396, 1972.

[20] G. Sambin, "Some points in formal topology," *Theoretical computer science*, vol. 305, no. 1, pp. 347–408, 2003.

[21] G. Boniolo and S. Valentini, "Vagueness, kant and topology: a study of formal epistemology," *Journal of Philosophical Logic*, vol. 37, no. 2, pp. 141–168, 2008.

[22] A. K. Seth, "Measuring autonomy and emergence via granger causality," *Artificial Life*, vol. 16, pp. 179–196, Jan. 2010.

[23] J. de Haan, "How emergence arises," *Ecological Complexity*, vol. 3, pp. 293–301, Dec. 2006.

[24] M. A. Bedau, "Weak emergence," *Noûs*, vol. 31, no. s11, pp. 375–399, 1997.

[25] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of molecular biology*, vol. 215, no. 3, pp. 403–410, 1990.

[26] P. Legendre and L. Legendre, *Numerical Ecology*. Elsevier, July 2012.

[27] H. Hotelling, "Analysis of a complex of statistical variables into principal components.," *Journal of educational psychology*, vol. 24, no. 6, p. 417, 1933.

[28] D. Bohm, *Causality and Chance in Modern Physics*. University of Pennsylvania Press, Jan. 1971.

[29] E. R. Dougherty and U. Braga-Neto, "Epistemology of computational biology: mathematical models and experimental prediction as the basis of their validity," *Journal of Biological Systems*, vol. 14, no. 1, pp. 65–90, 2006.

[30] G. K. Philip, C. J. Creevey, and J. O. McInerney, "The opisthokonta and the ecdysozoa may not be clades: Stronger support for the grouping of plant and animal than for animal and fungi and stronger support for the coelomata than ecdysozoa," *Molecular Biology and Evolution*, vol. 22, pp. 1175–1184, May 2005.

[31] H. Philippe, N. Lartillot, and H. Brinkmann, "Multigene analyses of bilaterian animals corroborate the monophyly of ecdysozoa, lophotrochozoa, and protostomia," *Molecular biology and evolution*, vol. 22, pp. 1246–1253, May 2005.

[32] R. I. Sadreyev, B.-H. Kim, and N. V. Grishin, "Discrete-continuous duality of protein structure space," *Current Opinion in Structural Biology*, vol. 19, pp. 321–328, June 2009.

[33] K. Binder and A. P. Young, "Spin glasses: Experimental facts, theoretical concepts, and open questions," *Reviews of Modern Physics*, vol. 58, pp. 801–976, Oct. 1986.

[34] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, "Complex networks: Structure and dynamics," *Physics Reports*, vol. 424, pp. 175–308, Feb. 2006.

[35] M. Girvan and M. E. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences*, vol. 99, no. 12, pp. 7821–7826, 2002.

[36] A. Pascual-García, D. Abia, Á. R. Ortiz, and U. Bastolla, "Cross-over between discrete and continuous protein structure space: Insights into automatic classification and networks of protein structures," *PLOS Computational Biology*, vol. 5, p. e1000331, Mar. 2009.

[37] H. Hasegawa and L. Holm, "Advances and pitfalls of protein structural alignment," *Current Opinion in Structural Biology*, vol. 19, pp. 341–348, June 2009.

[38] T. A. Holland, S. Veretnik, I. N. Shindyalov, and P. E. Bourne, "Partitioning protein structures into domains: Why is it so difficult?," *Journal of Molecular Biology*, vol. 361, pp. 562–590, Aug. 2006.

[39] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, "SCOP: a structural classification of proteins database for the investigation of sequences and structures," *Journal of molecular biology*, vol. 247, pp. 536–540, Apr. 1995.

[40] C. Orengo, A. Michie, S. Jones, D. Jones, M. Swindells, and J. Thornton, "CATH - a hierarchic classification of protein domain structures," *Structure*, vol. 5, pp. 1093–1109, Aug. 1997.

[41] L. Holm and C. Sander, "Dali/FSSP classification of three-dimensional protein folds," *Nucleic Acids Research*, vol. 25, pp. 231–234, Jan. 1997.

[42] D. Lupyan, A. Leo-Macias, and A. R. Ortiz, "A new progressive-iterative algorithm for multiple structure alignment," *Bioinformatics*, vol. 21, no. 15, pp. 3255–3263, 2005.

[43] W. L. DeLano, "The pymol molecular graphics system," 2002.

[44] R. Rammal, G. Toulouse, and M. A. Virasoro, "Ultrametricity for physicists," *Reviews of Modern Physics*, vol. 58, pp. 765–788, July 1986.

[45] C. Chothia and A. M. Lesk, "The relation between the divergence of sequence and structure in proteins.," *The EMBO Journal*, vol. 5, pp. 823–826, Apr. 1986.

[46] A. Pascual-García, D. Abia, R. Méndez, G. S. Nido, and U. Bastolla, "Quantifying the evolutionary divergence of protein structures: The role of function change and function conservation," *Proteins: Structure, Function, and Bioinformatics*, vol. 78, no. 1, pp. 181–196, 2010.

[47] N. V. Dokholyan, B. Shakhnovich, and E. I. Shakhnovich, "Expanding protein universe and its origin from the biological big bang," *Proceedings of the National Academy of Sciences*, vol. 99, pp. 14132–14136, Oct. 2002.

[48] K. B. Zeldovich, P. Chen, B. E. Shakhnovich, and E. I. Shakhnovich, "A first-principles model of early evolution: Emergence of gene families, species, and preferred protein folds," *PLoS Comput Biol*, vol. 3, p. e139, July 2007.

[49] N. Fernandez-Fuentes, J. M. Dybas, and A. Fiser, "Structural characteristics of novel protein folds," *PLoS Comput Biol*, vol. 6, p. e1000750, Apr. 2010.

[50] A. N. Lupas, C. P. Ponting, and R. B. Russell, "On the evolution of protein folds: Are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world?," *Journal of Structural Biology*, vol. 134, pp. 191–203, May 2001.

[51] N. V. Grishin, "Fold change in evolution of protein structures," *Journal of Structural Biology*, vol. 134, pp. 167–185, May 2001.

[52] J. D. Szustakowski, S. Kasif, and Z. Weng, "Less is more: towards an optimal universal description of protein folds," *Bioinformatics*, vol. 21, pp. ii66–ii71, Oct. 2005.

[53] A. Goncearenco and I. N. Berezovsky, "Prototypes of elementary functional loops unravel evolutionary connections between protein functions," *Bioinformatics*, vol. 26, pp. i497–i503, Sept. 2010.

[54] J. Skolnick, A. K. Arakaki, S. Y. Lee, and M. Brylinski, "The continuity of protein structure space is an intrinsic property of proteins," *Proceedings of the National Academy of Sciences*, vol. 106, pp. 15690–15695, Sept. 2009.

[55] C. O. Wilke, "Bringing molecules back into molecular evolution," *PLoS computational biology*, vol. 8, no. 6, p. e1002572, 2012.

[56] Y.-Y. Liu, J.-J. Slotine, and A.-L. Barabási, "Observability of complex systems," *Proceedings of the National Academy of Sciences*, vol. 110, no. 7, pp. 2460–2465, 2013.

[57] Y. Bar-Yam, "A mathematical theory of strong emergence using multiscale variety," *Complexity*, vol. 9, no. 6, pp. 15–24, 2004.

[58] M. Gell-Mann and S. Lloyd, "Information measures, effective complexity, and total information," *Complexity*, vol. 2, no. 1, pp. 44–52, 1996.

[59] F. Boschetti, "Causality, emergence, computation and unreasonable expectations," *Synthese*, vol. 181, no. 3, pp. 405–412, 2011.

[60] W. Bialek, A. Cavagna, I. Giardina, T. Mora, E. Silvestri, M. Viale, and A. M. Walczak, "Statistical mechanics for natural flocks of birds," *Proceedings of the National Academy of Sciences*, vol. 109, no. 13, pp. 4786–4791, 2012.

[61] R. Lopez-Ruiz, H. L. Mancini, and X. Calbet, "A statistical measure of complexity," *Physics Letters A*, vol. 209, no. 5, pp. 321–326, 1995.

[62] P. Bak, C. Tang, and K. Wiesenfeld, "Self-organized criticality: An explanation of the 1/f noise," *Physical review letters*, vol. 59, no. 4, p. 381, 1987.

[63] M. Vellend, "Conceptual synthesis in community ecology," *The Quarterly review of biology*, vol. 85, no. 2, pp. 183–206, 2010.

[64] O. X. Cordero and M. F. Polz, "Explaining microbial genomic diversity in light of evolutionary ecology," *Nature Reviews Microbiology*, vol. 12, no. 4, pp. 263–273, 2014.

[65] T. Dobzhansky, "Nothing in biology makes sense except in the light of evolution," *The american biology teacher*, vol. 35, no. 3, pp. 125–129, 1973.

[66] E. Mayr, "The objects of selection," *Proceedings of the National Academy of Sciences*, vol. 94, no. 6, pp. 2091–2094, 1997.

[67] F. Mazzocchi, "Complexity in biology. exceeding the limits of reductionism and determinism using complexity theory," *EMBO Reports*, vol. 9, pp. 10–14, Jan. 2008.

[68] M. H. V. Regenmortel, "Reductionism and complexity in molecular biology," *EMBO Reports*, vol. 5, pp. 1016–1020, Nov. 2004.

Chapter 2

Protein systems

We can never hope for a lasting peace and better times till Botanists come to an agreement among themselves about the fixed laws in accordance with which judgement can be pronounced on names.

Carl Linneaus

Summary

In this chapter we present results related with the analysis of protein systems, where we focus in the evolution of protein structures and its relation with protein sequence and protein function.

In the first article [Pascual-García et al. (2009)], we build relationships between protein structures computing their structural similarities, through the structural alignment algorithm developed in our laboratory MAMMOTH [Lupyan et al. (2005)], reviewed in [Pascual-García (2014)]. From these relationships, we have designed a procedure to determine whether there exists a similarity threshold that generates equivalence classes. Given three objects a, b, and c, an equivalence class is defined when we find an equivalence relationship R endowed with the following properties:

1. *Reflexivity*: aRa
2. *Symmetry*: $aRb \Rightarrow bRa$
3. *Transitivity*: if aRb and $bRc \Rightarrow aRc$

We note that these three properties hold for the phylogenetic distance (defined as the time from last common ancestor) if the molecular clock hypothesis holds. In this way, we look for an objective threshold monitoring a measure that we call *transitivity violations*, which quantifies whether the

transitivity property in the definition of equivalence class is accomplished, and that we justify as follows. If we consider a gene duplication event leading to two different branches a and b and, after this event, we observe a new duplication event in the branch a leading to two sub-branches a_1 and a_2, we expect that the similarity between these genes fullfills the relation $S(a_1, b) \approx S(a_2, b)$ (if the molecular clock holds). The result we obtain is a transitive relation between a_1, a_2 and b, and we can say that these genes belong to the same equivalence class, i.e. to the same cluster, and thus a classification of genes would be possible. Following this reasoning, we classify protein structures with different agglomerative clustering algorithms and we measure whether transitivity is fullfilled quantifying the error we incur joining clusters through the transitivity violations. We observe that the transitivity violations increase linearly until some clustering step after which it starts growing exponentially. We identify this transition automatically and this is the threshold selected to define an optimal classification.

We compare our results with the expert classifications of protein structures SCOP [Murzin et al. (1995)] and CATH [Orengo et al. (1997)], and we rationalize some discrepancies between both, and also between these classifications and our results. The similarity between the potential classifications obtained following our procedure and expert classifications is computed along the clustering. It is shown that the maximum similarity with SCOP and CATH is obtained at the fold level, but it is reached after the cross-over. This indicates that joining clusters until the fold level generates large transitivity violations. On the other hand, our results show that the classification at cross-over is more consistent with the superfamily level. Therefore, within the high similarity region, proteins within the same cluster are homologs which have diverged after gene duplication, which justifies to classify related domains on a tree. On the other hand, in the low similarity region, the relationships reflect either the occurrence of more dramatic evolutionary events –domains joined in the same cluster after cross-over differ substantially in size–, or examples of functional convergence. For this region, it is not further justified a tree representation and protein domains should be represented rather as a network.

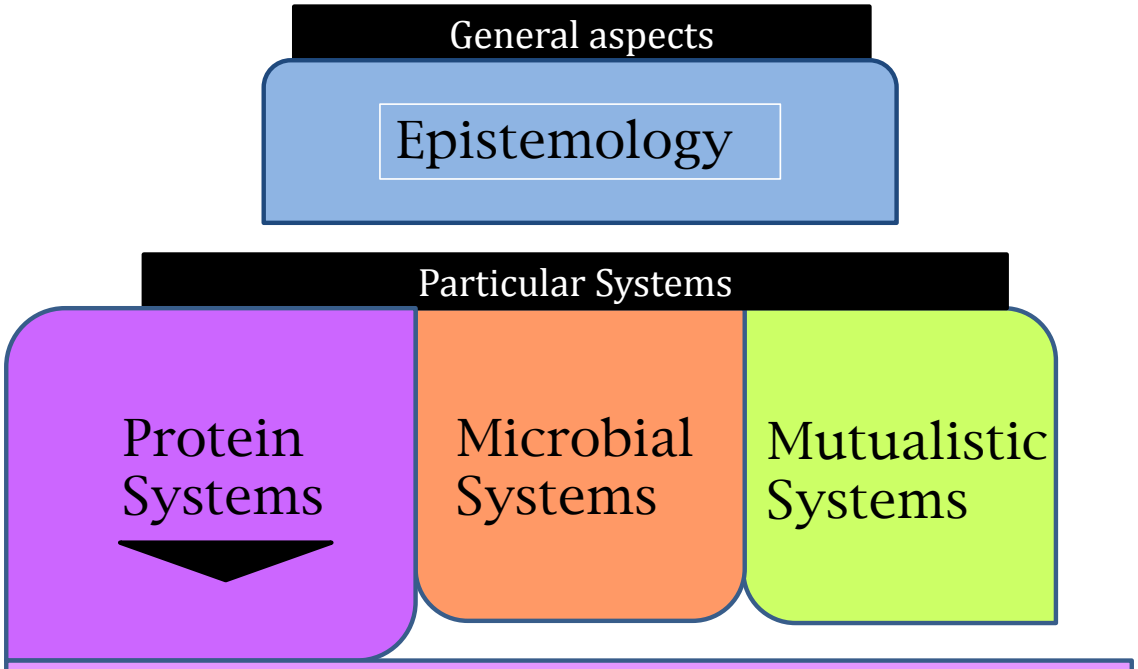
As it was previously mentioned, the existence of an intrinsic structural threshold for the determination of equivalence classes has been interpreted as a signature of the dominance of gene duplication events and subsequent divergence in the high similarity region. This observation drove us to look for a structural measure that would reflect a proportional relation with the sequence divergence –and thus it should be expected that this measure also correlates with time of divergence–, in a second paper [Pascual-García et al. (2010)]. If such measure exists, we would further explore the role of protein function in the conservation or divergence of protein structure.

Focusing on the interactions between the amino-acids, we worked around

the number of native contacts (amino-acids physically interacting), and proposed a measure quantifying the contact divergence between two proteins which reflects the desired properties. Indeed, using this measure we observed that the contact divergence increases linearly with the sequence identity until a point after which the structural divergence abruptly explodes. After building relationships based on sequence and functional similarities, the contact divergence allowed us to quantify the rate of divergence of the structure with respect to the sequence for four large superfamilies, finding that the structure is two to four times more conserved than the sequence. In addition, we further explored the role of function change and function conservation in these trends. Interestingly, function conservation goes together with a global conservation of the structure, finding most of the pairs of proteins with the same function constrained to the region of high structural similarity (before the contact divergence explosion). Nevertheless, it is possible to observe function change with global conservation of the structure also within this region. Both findings may reflect the existence of selection pressures acting either over the whole structure –for instance, favoring fast folding– or over a particular region –that may be the case if a particular region, such as an active site, performs a function in the protein–.

In summary, in these works we shed some light on the complex relationship between protein sequence, structure, function, and evolution.

2.1. Article [PROT-1]



1) Pascual-García A, Abia D, Ortiz AR, Bastolla U. (2009) Cross-over between discrete and continuous protein structure space: insights into automatic classification and networks of protein structures. *PLoS Comput Biol.* 5(3):e1000331.

2) Pascual-García A., Abia D., Méndez R., Nido G.S., Bastolla U. (2010) Quantifying the evolutionary divergence of protein structures: The role of function change and function conservation. *Proteins*. 78(1):181-196

3) Nido G.S., Méndez R., Pascual-García A ., Abia D. and Bastolla U. (2012) Protein disorder in the centrosome correlates with complexity in cell types number. *Mol. BioSyst.* 8, 353-367

4) Sánchez-Nido, G., Romano, L., Bastolla, U. and Pascual-García , *A .* (2015) Structural Bioinformatics within an Snake Puzzle. *In preparation*

5) Pascual-García, A. (2014) Evolución de estructura de proteínas.

6) Pascual-García, A. (2014) Modelos simplificados de plegamiento de proteínas.

7) Pascual-García, A. (2014) Alineamiento de estructura de proteínas.
Published in: "Bioinformática con Ñ".
Editors: Sebastián A., and Pascual-García, A.
ISBN: 978-84-617-1976-X.

OPEN ACCESS Freely available online PLoS COMPUTATIONAL BIOLOGY

Cross-Over between Discrete and Continuous Protein Structure Space: Insights into Automatic Classification and Networks of Protein Structures

Alberto Pascual-García, David Abia, Ángel R. Ortiz†, Ugo Bastolla*

Centro de Biología Molecular 'Severo Ochoa' (CSIC-UAM), Cantoblanco, Madrid, Spain

Abstract

Structural classifications of proteins assume the existence of the fold, which is an intrinsic equivalence class of protein domains. Here, we test in which conditions such an equivalence class is compatible with objective similarity measures. We base our analysis on the transitive property of the equivalence relationship, requiring that similarity of A with B and B with C implies that A and C are also similar. Divergent gene evolution leads us to expect that the transitive property should approximately hold. However, if protein domains are a combination of recurrent short polypeptide fragments, as proposed by several authors, then similarity of partial fragments may violate the transitive property, favouring the continuous view of the protein structure space. We propose a measure to quantify the violations of the transitive property when a clustering algorithm joins elements into clusters, and we find out that such violations present a well defined and detectable cross-over point, from an approximately transitive regime at high structure similarity to a regime with large transitivity violations and large differences in length at low similarity. We argue that protein structure space is discrete and hierarchic classification is justified up to this cross-over point, whereas at lower similarities the structure space is continuous and it should be represented as a network. We have tested the qualitative behaviour of this measure, varying all the choices involved in the automatic classification procedure, i.e., domain decomposition, alignment algorithm, similarity score, and clustering algorithm, and we have found out that this behaviour is quite robust. The final classification depends on the chosen algorithms. We used the values of the clustering coefficient and the transitivity violations to select the optimal choices among those that we tested. Interestingly, this criterion also favours the agreement between automatic and expert classifications. As a domain set, we have selected a consensus set of 2,890 domains decomposed very similarly in SCOP and CATH. As an alignment algorithm, we used a global version of MAMMOTH developed in our group, which is both rapid and accurate. As a similarity measure, we used the size-normalized contact overlap, and as a clustering algorithm, we used average linkage. The resulting automatic classification at the cross-over point was more consistent than expert ones with respect to the structure similarity measure, with 86% of the clusters corresponding to subsets of either SCOP or CATH superfamilies and fewer than 5% containing domains in distinct folds according to both SCOP and CATH. Almost 15% of SCOP superfamilies and 10% of CATH superfamilies were split, consistent with the notion of fold change in protein evolution. These results were qualitatively robust for all choices that we tested, although we did not try to use alignment algorithms developed by other groups. Folds defined in SCOP and CATH would be completely joined in the regime of large transitivity violations where clustering is more arbitrary. Consistently, the agreement between SCOP and CATH at fold level was lower than their agreement with the automatic classification obtained using as a clustering algorithm, respectively, average linkage (for SCOP) or single linkage (for CATH). The networks representing significant evolutionary and structural relationships between clusters beyond the cross-over point may allow us to perform evolutionary, structural, or functional analyses beyond the limits of classification schemes. These networks and the underlying clusters are available at http://ub.cbm.uam.es/research/ProtNet.php

Citation: Pascual-García A, Abia D, Ortiz ÁR, Bastolla U (2009) Cross-Over between Discrete and Continuous Protein Structure Space: Insights into Automatic Classification and Networks of Protein Structures. PLoS Comput Biol 5(3): e1000331. doi:10.1371/journal.pcbi.1000331

Editor: Philip E. Bourne, University of California San Diego, United States of America

Received August 20, 2008; **Accepted** February 11, 2009; **Published** March 27, 2009

Funding: This work was supported by the Ramon y Cajal program of the Spanish Science Ministry of Education and Science, Project 'Centrosoma 3D-Bioinformatics' of the program Consolider-Ingenio 2010 of the Spanish Ministry of Education and Science, Project BIO2005-0576 from the Spanish Ministry of Education and Science, Project 200520M157 from the Comunidad de Madrid, and Research Foundation "Ramon Areces".

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: ubastolla@cbm.uam.es

† Deceased.

Introduction

Structural genomics projects [1] aim at an exhaustive exploration of the space of protein structures realized in evolution [2,3], speeding up considerably the rate at which new protein structures are resolved. In this context, structural classification of proteins [4–9] has become essential for uncovering remote evolutionary relationship that can not be inferred from sequence information alone, and it will have important consequences on our understanding of protein evolution, the sequence to structure to function relationships, the recognition of remote homologs and the modelling of their structures.

This dramatic growth of the number of known protein structures calls upon automatic classification methods that are

Author Summary

Making order of the fast-growing information on proteins is essential for gaining evolutionary and functional knowledge. The most successful approaches to this task are based on classifications of protein structures, such as SCOP and CATH, which assume a discrete view of the protein structure space as a collection of separated equivalence classes (folds). However, several authors proposed that protein domains should be regarded as assemblies of polypeptide fragments, which implies that the protein–structure space is continuous. Here, we assess these views of domain space through the concept of transitivity; i.e., we test whether structure similarity of A with B and B with C implies that A and C are similar, as required for consistent classification. We find that the domain space is approximately transitive and discrete at high similarity and continuous at low similarity, where transitivity is severely violated. Comparing our classification at the cross-over similarity with CATH and SCOP, we find that they join proteins at low similarity where classification is inconsistent. Part of this discrepancy is due to structural divergence of homologous domains, which are forced to be in a single cluster in CATH and SCOP. Structural and evolutionary relationships between consistent clusters are represented as a network in our approach, going beyond current protein classification schemes. We conjecture that our results are related to a change of evolutionary regime, from uniparental divergent evolution for highly related domains to assembly of large fragments for which the classical tree representation is unsuitable.

objective and based only on structural information. The most used structural classifications of proteins, such as SCOP [4] and CATH [5], are manually curated, and therefore they are slow to update. For instance, the last update of SCOP at the moment of writing this paper took from october 2006 to november 2007 (13 months), and the last update of CATH took from may 2006 to january 2007 (9 months). This makes automatic classifications with similar quality to that of CATH and SCOP highly desirable.

However, this goal raises the question of whether, and up to which point, the classification of protein structures is justified. This question is addressed in this paper, where we ask whether an automatic classification based on an objective similarity measure can be uniquely defined.

Several authors studied the agreement between SCOP and CATH classifications [10–13], concluding that an overall agreement exists, but it is not satisfactory from a quantitative point of view. This problem is partially due to the fact that SCOP and CATH differ in the way in which they split the proteins into domains [12], which are the units of protein classifications. Nevertheless, they often classify differently even domains that are defined in the same way. Sam and coworkers [13] found out that more than 25% of the domain pairs classified in the same SCOP fold are not significantly similar under two measures of structure similarity.

The other side of the coin is that several structures classified in different folds present a significant structural similarity due to the presence of common substructures, a fact noted for instance by the group of Orengo and later by other groups [14,15], which in principle makes multiple classifications possible.

The first and most successful automatic classification of protein domains is the database FSSP [8], which is based on the DALI algorithm [9] and on its structure similarity measure. Though this similarity measure is overall consistent with the CATH and SCOP classifications important differences exist [11,12]. Other approaches aiming at the automatic classification of protein structures have been recently proposed by Rogen and Fain [16], Sam et al. [17], Zemla et al. [18] and by the group of Sippl [19]. However, the FSSP database and its more recent followers do not address the question to which extent structure classification is possible and unique. This question is the subject of the present paper.

Is Protein Structure Space Discrete or Continuous?

Some of the above difficulties are related with the very essence of protein classification schemes, which assume that it exists an intrinsic level of structure similarity for defining equivalence classes of protein structures. In SCOP, such an equivalence class is called *fold* [20]. Two proteins are defined to belong to the same fold if they share "the same major number and direction of secondary structures with a same connectivity" [4]. In CATH, the corresponding classification level is called *topology*, defined as "the overall shape and connectivity of secondary structures" [5]. These apparently clear definitions are in practice subject to substantial arbitrariety, first because it is not always clear which secondary structure elements belong to the structural core defining the fold and which ones are regarded as optional "embellishments", and second because one has to allow a certain extent of structural divergence in the protein core.

The difficulties presented above have led several authors to propose that the space of protein structures is continous [13,21,22]. This view is supported by the studies that underline the importance of substructures below the level of the globular domain, such as the autonomously folding units of Tsai et al [23], the loops of standard size (approximately 30 residues) of Berezowski and Trifunov [24], or the recurrent fragments of Tendulkar et al. [25] and Szustakowski et al. [26]. Expanding an old idea by Ohno [27], Lupas et al. [28] proposed that the most ancient folds have arisen through an evolutionary process consisting in assembling polypeptide fragments together. These and similar ideas have suggest that the basic unit of protein classification should be substructures below the domain level, defined by Shindyalov and Bourne [22] as "highly repetitive near-contiguous pieces of polypeptide chain that occur frequently" in a set of non-redundant protein structures. If protein domains can be regarded as a combination of such substructures, the resulting structure space should be seen as continuous rather than discrete.

A similar spirit is present in the approaches of Efimov [29] and in particular Taylor, who proposed to enumerate in a kind of periodic table all possible arrangements of secondary structure elements compatible with simple stability rules [30], consistent with the view that evolution of protein structures proceeds by combining simpler modules, resulting in a continuous structure space.

Homology and Structure Similarity Are Not Always Consistent

Another basic assumption of CATH and SCOP is that evolutionary relationships at the superfamily level imply structure similarity at the fold level. Although this assumption is most of the times correct, it was observed already in Ref. [31] that sequence divergence beyond ≈40% identity sometimes implies large structural variations. Grishin [32,33] has monitored several examples in which proteins belonging to the same superfamily diverged to the point where they do not share a common fold under the loose definition given above. Interestingly, many of these fold changes take place together with insertions or deletions of

large polypeptide fragments, although an interesting example of secondary structure switching has been reported between two homologues regions of distant related proteins [34,35]. Viksna and Gilbert [36] recently quantified these fold changes in protein evolution, finding that some of them are relatively common. The occurrence of fold change implies that the classification level based on evolution, as the superfamily, and the classification based on structure, as the fold, should not be necessarily consistent, as already recognized by the group of Orengo [14].

Results

Objective Fold Definition and Transitive Property

Given the above, one can ask whether protein classifications entirely based on a quantitative measure of structure similarity are possible at all, and if so to which extent.

In formal terms, a protein fold is an equivalence class of protein structures. Mathematically, an equivalence relationship must possess the three property of symmetry, reflexivity and transitivity. Whereas symmetry and reflexivity are automatically fulfilled by any relationship based on a similarity measure, transitivity is not. For transitivity to hold, every time that a is similar to b and b is similar to c, then a must also be similar to c. In other words, you can not make a big step from a to c by making an intermediate small step through b. Note that transitivity is not the same as the familiar triangular inequality, $d_{ac} \leq d_{ab} + d_{bc}$, which characterizes similarity measures obtained from a properly defined distance. Rather, transitivity is guaranteed by the much stronger property of ultrametricity [37], $d_{ac} \leq \max(d_{ab} + d_{bc})$, i.e., the distance travelled in two steps can not be larger than the longer of the two steps. An ultrametric set can be uniquely classified in the form of a tree.

Uniparental evolution satisfies transitivity. The importance of gene duplication for protein evolution [27] is a reason to expect that protein structural similarity fulfils the transitive property. The distance across the gene tree, i.e., the time spent since the divergence of two genes, is ultrametric (the time spent from the divergence of a and c can not be larger than the time either from the divergence of a and b or from the divergence of b and c), and therefore it is naturally endowed with the transitive property. Therefore, a phylogenetic tree naturally induces a hierarchical classification for every similarity threshold. If pairs of proteins are related through gene duplication, and if their structural dissimilarity correlates with the time of divergence, as it happens for suitable sequence dissimilarities when evolution is neutral, the transitivity property will approximately hold. However, directed evolution where new conformations are positively selected, for instance to fulfill a new function, may violate the last hypothesis.

Fragment assembly violates transitivity. Gene duplication is not the only possible mechanism for the evolution of protein domains. Complex proteins are formed from a combination of individual domains with independent evolutionary history. For this reason, the domain and not the complete protein is the basic unit for protein classification. However, there is increasing evidence that globular domains may be formed by combining fragments below the domain level [23–26,28], and it has been observed that many structurally unrelated proteins share common substructures [14,26,29]. If two domains a and b are similar because of a partial substructure A, while b and c are similar because of a different partial substructure C, then a and c are not similar and transitivity is violated. Several authors refer to this kind of situation by saying that protein space is continuous, since one can connect two different structures a and c with two small steps passing through b.

Transitivity Violation and Automatic Stop of the Clustering

If b is similar to both a and c but a and c are not similar, there is no classification simultaneously compatible with all the pairwise similarity relationships. Borrowing a term from statistical physics, we can say that the classification problem is *frustrated* [38] when transitivity is violated. We expect that, if this situation is common for many triplets, there is an exponentially large number of substantially different classifications that are almost optimal, in the sense that they violate a small and similar number of pairwise relationships. Conversely, if the transitive property approximately holds, we expect that a well-defined unique globally optimal classification exists, and all sub-optimal classifications are very similar to it.

We expect that the validity of the transitive property strongly depends on structure similarity. Domain pairs with high similarity share most of their structure, and we expect that transitivity approximately holds for them, so that at high similarity the structure space is made of discrete clusters. However, less stringent similarities may be due to partial substructures, and we expect that the transitive property will be violated, and the clustering will strongly depend on the algorithm used.

We propose here a measure to quantify the violation of the transitive property at each step of a hierarchical clustering algorithm. In this way, we aim at detecting the minimum similarity at which transitivity still holds and clustering is justified. At lower similarity, the space should be regarded as continuous, and the significant similarities between clusters should be represented as a network rather than a tree.

Let us consider three elements or clusters ABC, with the convention that $S(A,B) \geq S(B,C) \geq S(A,C)$. Violation of the transitive property occurs if $S(B,C)$ is large while $S(A,C)$ is small, so that B is an intermediate point between A and C. Therefore it is natural to define the transitivity violation of the triangle ABC as $S(B,C) - S(A,C)$. Such a quantity depends on the absolute scale and the offset of the similarity measure, i.e., it is not invariant if we multiply all similarities times a scale factor or we add to them a constant. To remove this dependency, we divide $S(B,C) - S(A,C)$ times the difference between the largest and smallest similarities, $S(A,B) - S(A,C)$, defining the transitivity violation associated to the triangle ABC as

$$\mathrm{TV}(ABC) = \frac{S(B,C) - S(A,C)}{S(A,B) - S(A,C)}. \qquad (1)$$

Notice that, by definition, Eq. (1) is comprised between zero and one because $S(B,C) \leq S(A,B)$ The maximum violation $\mathrm{TV} = 1$ happens when $S(B,C) = S(A,B)$ while $S(B,C) > S(A,C)$.

Another way to interpret this formula is the following. Because of transitivity, only five clustering configurations of the elements A, B and C are possible: all elements joined, all separated, two joined and the third one separated. For a threshold S_0, we say that the link (x,y) is violated if either x and y are joined despite $S(x,y) < S_0$ (overunification) or x and y are separated despite $S(x,y) > S_0$ (oversplitting). For thresholds S_0 such that $S_0 > S(B,C)$ or $S_0 < S(A,C)$ there is one and only one configuration that satisfies all links. However, if $S(B,C) > S_0 > S(A,C)$, no one of the five possible configurations satisfies all links, since either A and C are incorrectly joined, or B and C are incorrectly separated. The volume in the space of the threshold parameter S_0 such that some links are violated quantifies the violation of transitivity as $S(B,C) - S(A,C)$. On the other

hand, if $S_0 > S(A,B)$ all elements are separated, and if $S_0 < S(A,C)$ all elements are joined, so that only values of S_0 such that $S(A,B) > S_0 > S(A,C)$ correspond to non-trivial clustering. Therefore, Eq. (1) represents the ratio between the volume of parameter space for which transitivity is violated and the volume for which non-trivial clustering exist.

Yet a third way to look at the above equation is the following. Most hierarchical clustering algorithms join at each step t the two most similar clusters A and B and then recompute the similarity of the new cluster AB with any other one C. For the average linkage algorithm, we use the formula $S(AB,C) = w_A S(A,C) + w_B S(B,C)$, where w_A and w_B are proportional to the number of elements in sets A and B. The error made by substituting the original similarities $S(A,C)$ and $S(B,C)$ with the combined one is $\delta = w_A|S(AB,C) - S(A,C)| + w_B|S(AB,C) - S(B,C)| = w_A w_B (S(B,C) - S(A,C))$, and it is proportional to Eq. (1).

Finally, $S(B,C) - S(A,C)$ also quantifies the violation of ultrametricity, since in an ultrametric set the two longest sides of any triangle must be equal [37], which implies that $S(B,C) = S(A,C)$. Eq. (1) is normalized in such a way that the value 1 corresponds to the maximum possible violation of ultrametricity, $S(B,C) = S(A,B)$.

Now let us consider the step t of the clustering algorithm in which clusters A and B are joined. We define the transitivity violation at this step as the weighted sum of the transitivity violations for all triangles involving A and B:

$$\mathrm{TV}(A+B \rightarrow AB) = \sum_{C \neq A,B} w_C \mathrm{TV}(ABC), \qquad (2)$$

where w_C is proportional to the number of elements in cluster C, and for each triangle we label as B the element such that $S(\mathrm{A,B}) \geq S(\mathrm{B,C}) \geq S(A,C)$.

Cross-Over in Transitivity Violations

The main result obtained in this study is the existence of a cross-over in the behavior of transitivity violations. This cross-over point determines an intrinsic condition for stopping the hierarchical clustering algorithm. We call the classification obtained at this point "automatic classification".

The results that we present here are based on a set of 2890 domains that are decomposed very similarly in the SCOP and CATH databases (see Methods), so that the domain decompositions are more likely to be accurate and differences between CATH and SCOP on this set can not be attributed to their different ways of decomposing proteins into domains. We compute structure similarities using the Mammoth-mult algorithm [39], which is one of the fastest algorithms for such a purpose and is comparable in accuracy to other state of the art algorithms [40]. The similarity measure that we use is based on the contact overlap, normalized in such a way as to eliminate the dependence on the domain size for pairs of unrelated domains, and for clustering we use the average linkage algorithm (see Methods). These choices yielded the best results, as described below, and the results presented will refer to them unless otherwise stated.

We plot in Figure 1 the transitivity violations as a function of the step t of the clustering algorithm. For large t the clusters joined are less similar and the transitivity violations increase. The plot can be divided into two regimes: an initial part with slow increase of transitivity violations at large similarity and a final part with faster increase and small similarity. The cross-over between these two regimes can be detected through a two-pieces fit (see Methods). The normalized error of the fit, plotted in Figure 1 versus the trial cross-over point, allows us to detect at its minimum the optimal

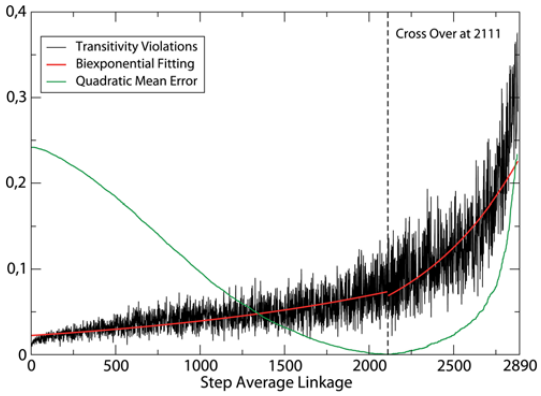


Figure 1. Violations of transitivity, Eq. (2), as a function of the step of the average linkage algorithm. We also plot the mean quadratic error of the two-piece linear fit, whose minimum identifies the cross-over point, plotted as a vertical line;
doi:10.1371/journal.pcbi.1000331.g001

cross-over point, depicted as a vertical line. The classification obtained at this cross-over point is called here "automatic classification", since the threshold similarity at which the clustering algorithm is stopped is automatically determined. We find $t_0 = 2111$, corresponding to joining two clusters with similarity $S_0 = 6.78$. At the stopping point, the automatic classification has 779 clusters.

Robustness of the Method

In order to test the robustness of our method, we repeated the numerical experiments changing all the relevant choices: The alignment algorithm, the similarity measure and its normalization, the clustering algorithm and the set of domains. In all cases, we observed a clear cross-over in the behavior of the transitivity violations, and the cross-over point could be automatically located through our algorithm. Moreover, the cross-over point did not vary very much for different choices (see Table 1).

In order to choose the best options, we measured the transitivity violations, the clustering coefficient, which is the network analogous of the transitive property (see Methods), and the agreement of the automatic classification with SCOP and CATH as assessed through the weighted kappa measure, which is a normalized measure of consistency between two classifications (see Methods). These measures tend to be consistent, i.e., choices yielding larger clustering coefficient tend to yield smaller transitivity violations and larger weighted kappa as well. This justifies the use of the weighted kappa to assess the method, despite the problems that we will discuss in the following and that limit the best possible agreement between the automatic classification and SCOP or CATH. In particular, we considered the following options:

1. As **structure alignment** method, we used either the multiple [39] or the pairwise [41] version of the MAMMOTH algorithm. As it has been recently assessed through an extensive test [40], MAMMOTH multiple is of comparable accuracy to other state of the art structure alignment tools and faster than most of them, while its pairwise version is even faster, but at the expense of accuracy. Moreover, the two algorithms are based on different principles, since Mammoth pairwise optimizes the local superimpositions of heptamers whereas Mammoth-mult optimizes the global superimposition of the two structures. Nevertheless, we

Table 1. Robustness of the automatic classification.

Set	Ali	Score	Norm	Cl. Al.	N.Clu.	Clus.co.	T.V.	WKSS	WKSF	WKCS	WKCF
SCOP 2890	MM	Cont.	Gauss	AL	**779**	**0.90**	**0.072**	**0.54**	**0.69**	**0.58**	**0.32**
SCOP 2890	MM	**TM**	**No**	AL	740	0.87	0.101	0.59	0.60	0.55	0.22
SCOP 2890	MM	**PSI4-p**	EV	AL	768	0.88	0.088	0.51	0.57	0.51	0.24
SCOP 2890	MM	**PSI6-p**	EV	AL	855	0.87	0.113	0.54	0.58	0.52	0.27
SCOP 2890	MM	**PSI4-t**	EV	AL	788	0.88	0.084	0.49	0.60	0.48	0.26
SCOP 2890	MM	Cont.	**No**	AL	883	0.88	0.069	0.57	0.50	0.53	0.27
SCOP 2890	**MP**	Cont.	**No**	AL	950	0.86	0.070	0.51	0.54	0.53	0.23
SCOP 2890	**MP**	**PSI4-p**	EV	AL	797	0.77	0.089	0.47	0.44	0.49	0.19
SCOP 2890	**MP**	**PSI4-t**	EV	AL	758	0.88	0.085	0.51	0.54	0.51	0.25
SCOP 2890	MM	Cont.	Gauss	**SL**	876	0.90	0.167	0.24	0.48	0.54	0.69
SCOP 2890	MM	Cont.	Gauss	**CL**	730	0.90	0.080	0.26	0.47	0.43	0.10
CATH 2890	MM	Cont.	Gauss	**A**L	776	0.90	0.079	0.50	0.71	0.54	0.36
SCOP 5041	MM	Cont.	Gauss	AL	1353	0.92	0.063	0.61	0.52	-	-
CATH 7073	MM	Cont.	Gauss	AL	2287	0.91	0.068	-	-	0.51	0.14

The qualitative features of the classification at the cross-over point are robust with respect to different methodological choices. First column, set of domains at less than 40 percent sequence identity: either 2890 domains from SCOP, or the corresponding 2890 domains from CATH, or 5041 domains from SCOP, or 7073 domains from CATH. The number of superfamilies and folds is, respectively: SCOP 2890: 779, 466; CATH 2890: 873, 473; SCOP 5041: 1094, 660; CATH 7073: 995, 1852. 2nd column, alignment algorithm: either the multiple structure alignment algorithm MAMMOTH multiple (MM) or its pairwise version (MP), faster but much less accurate. 3rd column, similarity measures: either Contact Overlap (Cont.) or TM score (TM) or percentage of structure identity (PSI). This can have a tolerance of either 4Å or 6Å , and it can be normalized either with respect the length of the shortest domain, PSI partial (PSI-p), or with respect to the geometric average, PSI total (PSI-t). 4th column, normalization with respect to length: either none, or Gaussian statistics (Gauss) or extreme value statistics (EV) 5th column, clustering algorithms: either average linkage (AL), or single linkage (SL) or complete linkage (CL). The results presented are the following. Number of clusters at the cross-over point (6th column), clustering coefficient averaged until the cross-over similarity (7th column), mean transitivity violations(8th column) and weighted kappa with respect to SCOP superfamilies (9th column), SCOP folds (10th column), CATH superfamilies (11th column) and CATH topologies (12th column), The first line in bold face refers to the selected choices, used in the presented results. In the following lines we evidence in bold face the variables that have changed with respect to the reference.
doi:10.1371/journal.pcbi.1000331.t001

obtained very similar results with the two algorithms, which shows that the whole methodology is not very sensitive to the accuracy of the alignment. We used the more accurate MAMMOTH-mult algorithm as the standard option.

2. We used several different measures of **structure similarity**. First, we used measures that require optimal rigid-body superimposition of the aligned residues. Such is the the percentage of structure identity (PSI), which counts the percentage of aligned residues that superimpose within a given threshold after optimal rigid body superimposition. In order to examine the influence of this threshold, we used the standard value 4Å as used in the standard MAMMOTH score and the larger tolerance 6Å. We normalized the PSI either through the length of the shorter protein, Eq. (5), which does not penalize matches that are only partial (we refer to it as the Partial PSI) or through the geometric mean length, Eq. (6) (Total PSI). As an alternative to an arbitrary tolerance parameter we tested the TM score [42], which uses a length dependent threshold that makes this score almost independent of the size of the aligned proteins. Second, we used the contact overlap, Eq. (7), which does not depend neither on the optimal rigid body superimposition nor on a tolerance parameter, although it depends on the parameter used to define contacts, i.e., interatomic interactions in the native structure. Most of the results presented here are obtained with the overlap as similarity score.

In order to remove the dependence on protein length for unrelated proteins, we normalized the PSI and the overlap as in Eq. (8). The parameters used in this expression were determined by fitting mean and standard deviation of the similarity of unrelated structures with respect to the length used to normalize the PSI, using either Gaussian statistics Eq. (9), or extreme value statistics, Eq. (10), as in the original Mammoth paper.

The best similarity score was selected based on the value of transitivity violations and the clustering coefficient evaluated up to the automatic cross-over point (see Methods). Using these criteria, the best score was the contact overlap (see Figure S1).

The normalization with respect to domain size did not modify the clustering coefficient considerably. However, measures that omit the normalization yield much lower agreement with expert classifications, and their cross-over points are rather distinct, whereas all the normalized scores have almost the same cross-over points. Therefore, normalized scores were used as the standard.

3. As **clustering method**, we considered average linkage (AL), single linkage (SL) and complete linkage (CL). We also used the neighbour joining algorithm (NJ), finding results very similar to those with average linkage (data not shown). For this comparison, we did not use the clustering coefficient, since it does not depend on the clustering algorithm.

The plot of transitivity violations for the three algorithms is shown as Figure S2, plot A. Not surprisingly, we found the best results with the average linkage algorithm, which can be interpreted as an algorithm trying to minimize the combination of oversplitting and overunification transitivity violations. The complete linkage only minimizes overunification errors, since it separates all structures that are below the similarity threshold. Its transitivity violations are only slightly larger than for the average linkage, but its weighted kappa is much smaller. The single linkage only minimizes oversplitting errors, since it joins all pairs above the similarity threshold. Correspondingly, it generates larger clusters. Its transitivity error is much larger than for complete and average linkage.

Remarkably, single linkage clustering agrees much better than average linkage with the CATH classification at topology (fold)

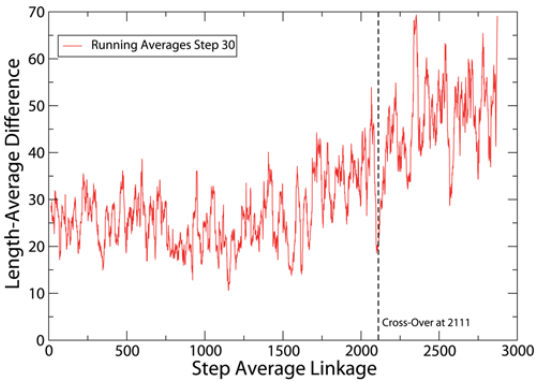


Figure 2. Difference between the mean lengths of the two joined clusters, Eq. (3), versus the average linkage step. The cross-over of transitivity violations is depicted as a vertical line. One can see that length differences are significantly larger after the cross-over. To improve the representation, we performed running averages with window size of 30 steps.
doi:10.1371/journal.pcbi.1000331.g002

level. This is not surprising, since CATH uses single linkage clustering, but it is an interesting observation, since it illustrate that one basic difference between CATH and SCOP arises from their reliance on different clustering procedures. However, superfamilies agree much better with the average linkage classification for both CATH and SCOP. More important, the transitivity violation is an intrinsic criterion, not based on any reference classification, which clearly favors the average linkage algorithm (see also the Discussion).

4. As **domain set**, we used the consensus domains (2890 domains), the ASTRAL40 set of domains corresponding to SCOP release 1.63 (5041 domains), and the set of non-redundant domains at the 35 percent sequence identity threshold corresponding to CATH release 3.1.1 (7073 domains).

The number of domains per fold as defined by SCOP (1.67, 2.05) and CATH (1.64, 2.30) increases with the size of the set, as we would expect from the fact that the cluster size is power law distributed, so that smaller samples are more likely to have smaller averages. The same happens at the level of superfamily. In contrast, the number of domains per cluster does not increase for larger samples, being 3.71 and 3.73 for SCOP domains and 3.71 and 3.09 for CATH domains. This indicates that our method tends to stop the clustering process relatively earlier for larger samples. In fact, larger samples are more likely to contain proteins that evidence transitivity violations. The plots of transitivity violations are qualitatively very similar, and are represented in Figure S2, plot B.

Length Differences

At each clustering step, we measure the difference between the average domain length of the two joined clusters A and B,

$$\text{Length difference} = \left| \frac{\sum_{a \in A} L(a)}{n_A} - \frac{\sum_{b \in B} L(b)}{n_B} \right| \qquad (3)$$

One can see from Figure 2 that the length difference is significantly larger after the cross-over point when transitivity violations increase faster. This observation is consistent with the intepretation that the regime of large transitivity violations takes place when the joined clusters are more likely to share only partial substructures. This behavior of the length difference is very robust with respect to changes in the clustering algorithm, similarity score, or set of domains.

Statistics of the Cluster Size

At the cross-over point, we find a broad distribution of the number of domains per cluster, with power-law probability density, $p(n) \approx n^{-2.4 \pm 0.1}$. This result agrees with the distribution of the number of proteins predicted to belong to specific folds in various genomes, which follow power-laws [43] with exponents between 2.5 and 4.0, approaching 2.5 for large genomes [44]. It also agrees very well with the automatic clustering by Dokholyan et al. [45], who found an exponent of 2.5 using as similarity measure the Dali score [9], with single linkage clustering and threshold derived from the statistical analysis of the domain similarity network. We also measured the cluster size distribution in the SCOP classification with 40 percent sequence similarity threshold to reduce redundancy, finding $p(n) \approx n^{-2.1 \pm 0.3}$ for folds and $p(n) \approx n^{-2.0 \pm 0.1}$ for superfamilies.

Therefore, the exponent of the distribution of the number of domains per cluster agrees reasonably between the SCOP and the automatic classification. Nevertheless, this agreement is not an evidence of the consistency between classifications, since the same size distribution can be found also for clusters obtained from random networks with the same statistical properties [45].

Comparison of Automatic and Expert Classifications

Weighted kappa. We compared the automatic classification with SCOP and CATH measuring their weighted kappa, which is plotted in Figure 3 versus the step of the average linkage. At first kappa increases steadily, since most joined domains belong to the same superfamily or fold, then it reaches a plateau and it decreases steeply when most of the joined domains belong to different folds or superamilies. The maximum of kappa is reached earlier, i.e., at larger number of clusters, for superfamilies than for folds, as expected since there are more superfamilies than folds. The maximum kappa for folds is larger than for superfamilies, which seem at first sight surprising, since structural similarity is on the average larger within a superfamily than within a fold. However, kappa can be decomposed into the contributions of related and unrelated pairs, with weights proportional to the number of related and unrelated pairs, respectively, see Eq. (20). For folds, the ratio of related to unrelated pairs, and consequently the weight of related pairs, is larger than for superfamilies. Therefore, kappa will be larger when all domains in the same fold are joined than when all domains in the same superfamily are joined.

The cross over point is located before the maximum weighted kappa for folds, indicating that many clustering steps that join clusters containing domains in the same fold imply large transitivity violations. This suggests that these fold relationships are more compatible with a network than with a classification. The difference between the automatic classification and the classification at the step where the kappa for folds is maximum becomes larger when more domains are added to the set, which makes it more likely to find transitivity violations that prevents clusters from being joined.

These results are robust with respect to the different choices mentioned above. In the following, we analyze in more detail the instances of disagreement between the automatic and the expert classifications.

Splitting of SCOP and CATH superfamilies. At the cross-over point, the great majority of the clusters only contain domains

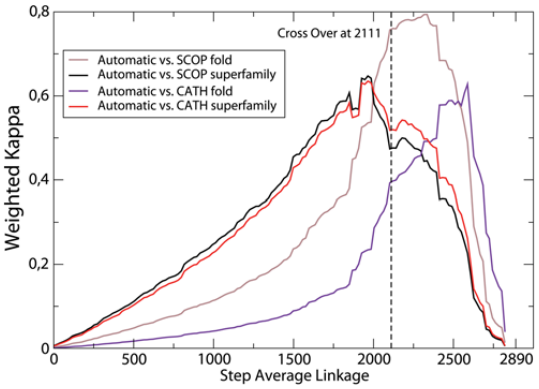


Figure 3. Weighted kappa measuring the agreement the average linkage classifications with step represented in the horizontal axis and SCOP and CATH superfamilies and folds. Notice that the cross-over point, depicted as a vertical line, lies between the maximum agreement with superfamilies and the maximum agreement with folds.
doi:10.1371/journal.pcbi.1000331.g003

in the same SCOP or CATH superfamily. Their number is 632 for CATH superfamilies, 664 for SCOP superfamilies, and 673 over 779 (more than 86 percent) for either SCOP or CATH superfamilies (see Table 2).

Several superfamilies are splitted in various clusters of the automatic classification. This is one of the most common disagreement between the automatic and the expert classifications. This is however not surprising, since it is well known that evolutionarily related proteins may diverge structurally. The number of splitted superfamilies is 115 over 779 (almost 15%) for SCOP and 87 over 885 (less than 10%) for CATH, which splits several superfamilies that are unique in SCOP.

To analyse these splittings, we measured the distribution of structure similarity between each pair of domains in the same

Table 2. Detailed comparison between automatic and expert classifications.

Reference classification	Num. clust.	Homogeneity	Joining probability
SCOP SF	779	85.2	68.0
CATH SF	885	81.1	66.4
SCOP or CATH SF	-	86.3	69.1
SCOP folds	466	92.0	44.5
CATH folds	473	91.4	10.7
SCOP or CATH folds	-	95.4	45.0

First column: reference classification. Second column: Number of clusters in the reference classification. Third column: Percentage of the 779 clusters in the automatic classification that are pure with respect to the reference classification (in case of CATH or SCOP, it is the fraction of clusters that are pure with respect to either CATH or SCOP). Fourth column: Percentage of the pairs joined in the reference classification that are joined in the automatic classification. In the case of folds, only pairs in different superfamilies are counted.
doi:10.1371/journal.pcbi.1000331.t002

SCOP superfamily, distinguishing split superfamilies from superfamilies contained in just one cluster of the automatic classification. The two distributions are shown in Figure 4A. Similarities in split superfamilies show a bimodal distribution, with one peak at low similarity corresponding to pairs of domains belonging to different clusters and one peak at high similarity corresponding to pairs in the same cluster. This indicates that the splitting is not an artifact of the method, but it reflects a significant difference between split and unsplit superfamilies.

For some cases, the difference between domains in the same superfamily appears to be due to large insertions or deletions of secondary structures, which may produce fold changes in protein evolution [32,33,36]. In fact, we measured the difference in length between proteins in the same superfamily, distinguishing split and unsplit superfamilies. The median size difference is 41 residues for splitted superfamilies, as compared with 22 residues for unsplitted ones. One such example of split superfamilies is shown in

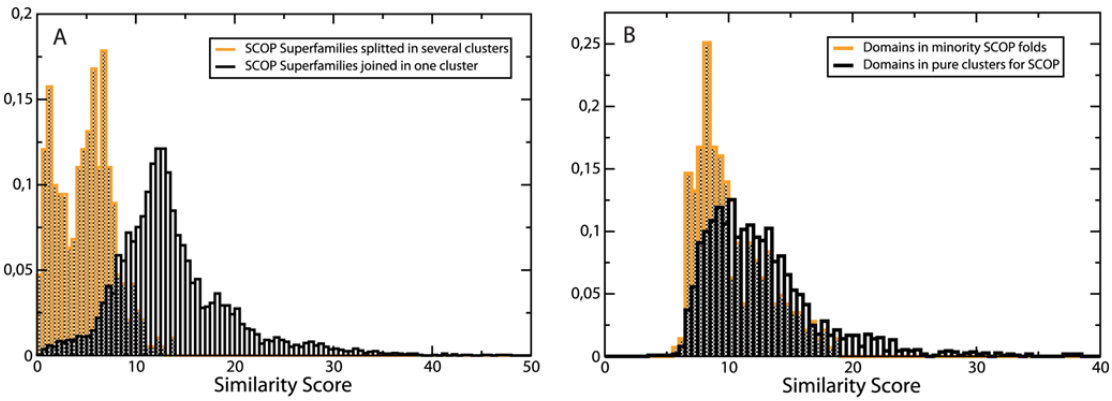


Figure 4. Distributions of intra-superfamily and intra-cluster similarity scores. (A) Distribution of the normalized total similarity score, Eq. (6) and (8), for domain pairs in the same superfamily. The grey bars are obtained for superfamilies that are not split, whereas the white bars are obtained for splitted superfamilies. One can see that splitted superfamilies present a bimodal distribution, with a peak with very small structure similarity. (B) Distribution of the mean intracluster similarity in the automatic classification, Eq. (11). The white bars are obtained for domains in clusters that contain only proteins of the same SCOP fold. The orange bars are obtained for minority domains in clusters containing domains that are mostly of a different SCOP fold.
doi:10.1371/journal.pcbi.1000331.g004

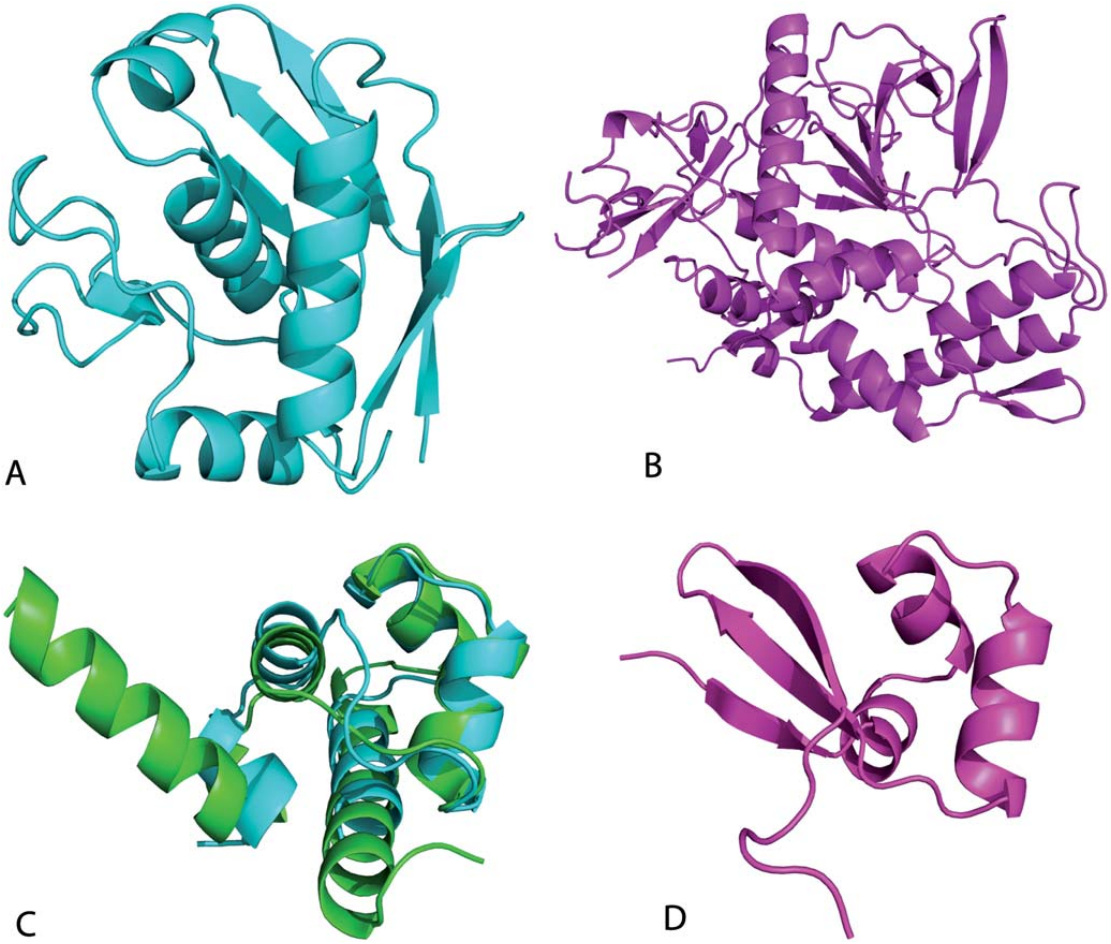


Figure 5. Examples of splitted SCOP superfamilies with large structural changes. Above: Two domains classified in SCOP in the metalloproteases superfamily, but splitted in CATH. Their codes are 1c7ka_ (A) and 1e1h.1 (B), with lengths of 132 and 399 residues respectively. Most of the secondary structure elements in the long protein are not matched in the short one. Below: Lambda repressor-like DNA-binding domains 1lmb3_ and 1r69__ (C) and 1d1la_ (D), which represent a well studied example of possible secondary structure switch in evolution.
doi:10.1371/journal.pcbi.1000331.g005

Figure 5A, showing domains 1c7ka_ and 1e1h.1, both from the SCOP superfamily of metalloproteases (55486). The first domain has 132 residues, and it is automatically classified in a cluster of 5 domains from the same superfamily with average length 163. The second domain has 399 residues and it is not joined with any other domain. Only three of the five beta strands in the main sheet of the large domain superimpose with the corresponding strands in the small domain. The large domain has several additional beta strands and alpha helices. CATH also separates the two domains. It includes the cluster containing 1c7ka_ in the superfamily collagenase, and the domain that we separate in the superfamily metalloproteases.

Another example is the superfamily lambda repressor-like DNA-binding domains (47413). We separate this superfamily in two clusters, one containing the domains with ASTRAL id. 1lmb3_ and 1r69__ and another one containing domain 1d1la_. This is consistent with the CATH classification, which separates them in two different topologies, and even two different secondary structure classes (all alpha and alpha+beta). Domains 1lmb3_ and 1d1la_ constitute possibly a very interesting example of evolutionary secondary structure switch between proteins that could be demonstrated to be homologues [34,35]. Placing both structures in the same fold puts in shadow this very interesting example of divergent structure evolution.

A number of splittings is due to the limited ability of the similarity score to assign significant similarity to short proteins In fact, the average overlap or PSI of unrelated structures is larger for short proteins, and therefore a larger overlap or PSI is required to judge it as significant (see Eq. (8)). As a consequence, there is a bias to split superfamilies with small domains: The mean length of splitted superfamilies is 165 residues versus 180 residues for superfamilies that are not splitted. We show one such example in Figure 6, which represents three short domains of the homeodomain-like superfamily that would be joined at a similarity value

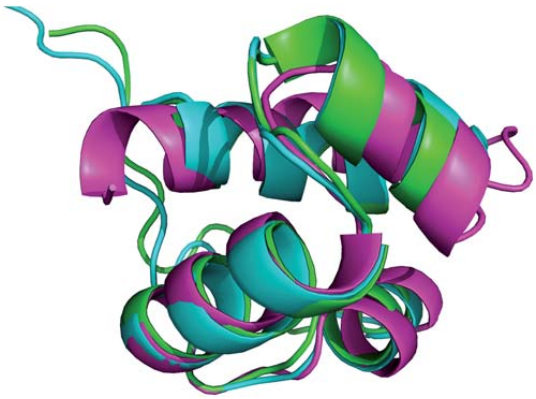


Figure 6. Three small domains of the Homeodomain-like superfamily, with PDB codes 1bl0a1, 1bl0a2 and 1d5ya2 are splitted in two clusters despite very high similarity. These clusters would be joined with $S=6.1$, short after the cross-over. This is an example of the limitation of the similarity measure in recognizing significant similarity when dealing with small structures.
doi:10.1371/journal.pcbi.1000331.g006

slightly below the cross-over (at $\tilde{z}$-score 6.1). A possible solution would be to modify the score so that the similarity does not depend on chain length neither for closely related nor for unrelated proteins. We will study such a modification in following work.

Fold unification. The automatic classification disagrees with CATH or SCOP when two domains in the same cluster belong to different folds. This kind of disagreement is rather rare. Only 142 domains over 2890, i.e., less than 5 percent, are contained in clusters where the majority of domains is from another SCOP fold, and they are distributed in only 63 clusters, so that 92 percent of the clusters contains only domains from the same fold. Similarly, 124 CATH domains over 2890 are minority domains, distributed in 67 clusters. However, these do not coincide with the 62 homogeneous clusters according to SCOP. Only 36 clusters (less than 5 percent) are not homogeneous according to both SCOP and CATH, indicating a very high agreement in cluster composition with the expert classifications (see Table 2).

For analyzing these disagreements, we computed the mean similarity score of each domain with the other domains in the same cluster, distinguishing domains in homogeneous clusters from minority domains in clusters with a majority of domains of a different fold. As one can see in Figure 4B, the two distributions overlap quite considerably, but their median values are significantly different, which means that it may be possible to distinguish some minority domains and "clean" some clusters from them. This possible refinement of the clustering will be studied elsewhere.

Some examples of fold unification are represented in Figure 7. One such case involves SCOP folds Tim Beta/Alpha Barrel (51350) and 7-stranded beta/alpha barrel (51988). They correspond to two distinct CATH topologies with the same names as in SCOP. However, the distribution of domains in the two folds is not the same in SCOP and CATH. We split these two folds into seven clusters. Four clusters are pure for both SCOP and CATH, which agree in classifying them as TIM barrels, two clusters only contain 7-stranded barrels according to SCOP but all domains but one are classified as TIM barrels in CATH, and the last cluster contains, together with 12 TIM barrel domains, one domain, 1m65a_ that is considered 7-stranded in SCOP and TIM barrel in CATH. Visual inspection supports the 7-stranded classification, in agreement with SCOP, but the structure similarity inside the cluster is very high.

In another example, the automatic classification joins domains from the SCOP folds Spectrin repeat-like (46965, corresponding to CATH topology 12058) and STAT-like (47654, corresponding to CATH topology 1201050) in three different clusters. However CATH classifies domain 1lvfa_, which is STAT-like according to SCOP, in the Spectrin repeat-like fold, while a paper of the SCOP team reports that the SCOP release 1.53 changed the classification of domain 1br0 from spectrin repeat to STAT-like, showing that even experts can confound these two folds [46]. Visual inspection shows that the domains that we unify are indeed very similar.

The third example corresponds to two domains from SCOP folds PIN domain-like (PDB code 1o4wa_) and Adenine Nucleotide alpha Hydrolase-like (PDB 1jmva_), which are automatically classified in the same cluster. Besides a very high structure similarity, these folds have an almost identical description in the SCOP database (beta-sheet of 5 strands, order 32145).

Splitting of folds. Another possible disagreement happens when superfamilies that are joined together in the same SCOP fold or CATH topology are splitted in different clusters. This is very frequent: 55.5 percent of the domain pairs in the same SCOP fold but distinct superfamilies are separated. For CATH, this percentage raises to 89.2%. This is not likely to be an artifact of the automatic classification, since the automatic classification agrees with SCOP or CATH at the fold level better than they agree with each other, as discussed in next section. The transitivity analysis suggests that this happens because SCOP and CATH join superfamilies into folds at a similarity level for which transitivity violations are rather large, so that clustering is not justified and unique. At this similarity level different clustering algorithms yield radically different classifications. In contrast, the pairs of domains of the same superfamily that are separated in the automatic classification is significantly smaller, 32% for SCOP and 34% for CATH.

Analysis of Expert Classifications

Comparison between SCOP and CATH. The expert classification schemes CATH and SCOP split proteins into domains differently. Domains in the CATH classification are typically smaller than those in the SCOP classification, with an average of 155 residues compared to 179 residues for SCOP domains (the standard deviations are 88 and 120 respectively). Comparison with a set of expert curated domain decompositions [47] shows that SCOP undercuts domains, whereas CATH decompositions are usually in good agreement with experts [48]. We used here 2890 domains similarly defined in both SCOP and CATH. For this consensus set, we measured the agreement between the SCOP and the CATH classification through the weighted kappa (see Methods). The values found are reported in Table 3, where the automatic classification is also shown for comparison.

There is rather good agreement, $\kappa=0.84$, between CATH and SCOP at superfamily level. The 779 SCOP superfamilies become 885 with CATH (almost 14 percent more), but CATH superfamilies are larger, so that 26320 pairs of domains are in the same CATH superfamily versus 22937 for SCOP, of which 90 percent (i.e., 20695) are common.

The agreement with the average linkage clustering is significantly weaker. Around 68 percent and 66 percent of pairs in the same SCOP and CATH superfamily are in the same automatic cluster, since many superfamilies are split in the automatic classification.

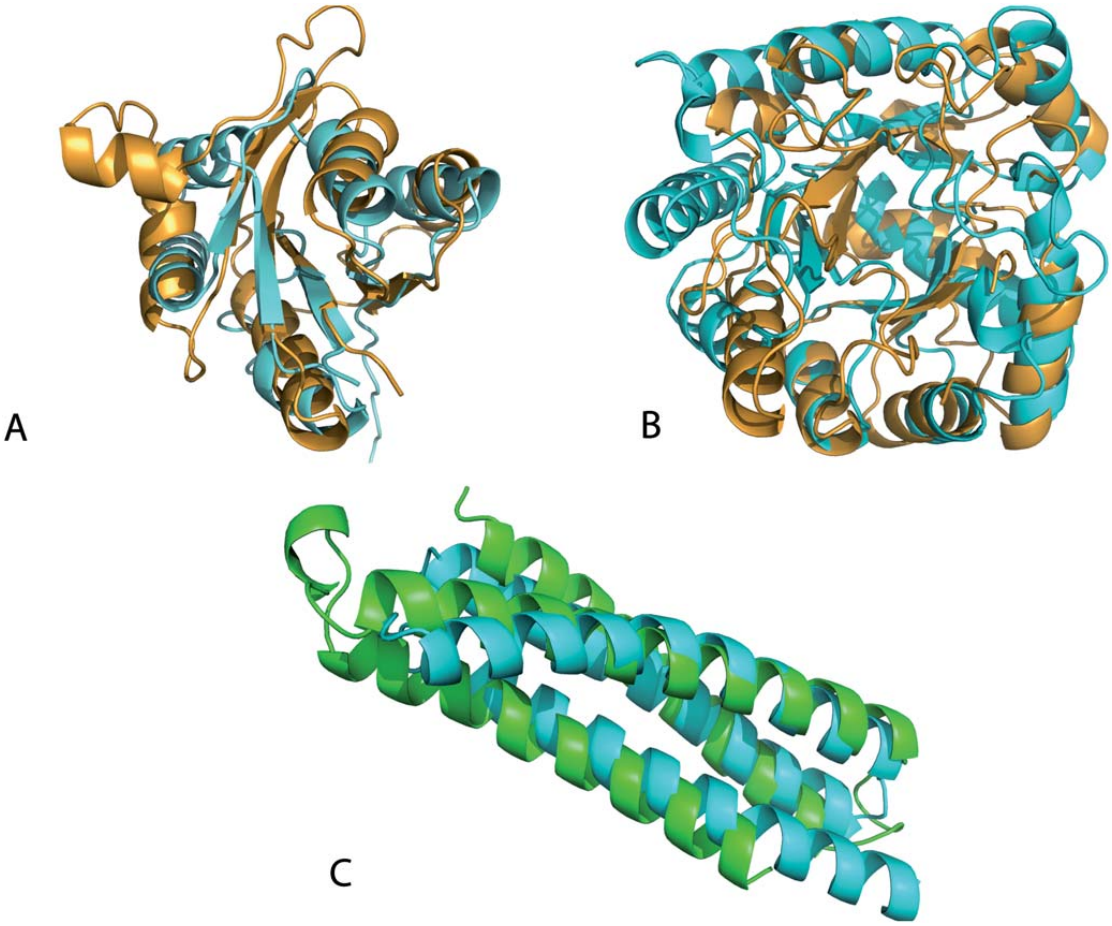


Figure 7. Examples of fold unifications. (A) Domain 1o4wa_ from SCOP fold PIN domain-like and domain 1jmva_ from fold Adenine Nucleotide alpha Hydrolase-like. They have a nearly identical description in the SCOP database in terms of secondary structure elements. (B) The 7-stranded barrel with code 1m65a_ is unified to a cluster with 12 TIM barrel, one representative of which, with code 1j6oa_, is shown for comparison. (C) Unification of two domains from the SCOP folds STAT-like (PDB 1lvfa) and spectrin repeat-like (PDB 2e2aa).
doi:10.1371/journal.pcbi.1000331.g007

In contrast, the agreement between CATH and SCOP at fold level is much poorer, with $\kappa = 0.48$. This suggests that the fold is more subjectively defined than the superfamily. The disagreement comes mainly from the fact that CATH joins many more pairs than SCOP at fold level: there are 3.9 times as many pairs classified as same fold and different superfamily by CATH than by SCOP (137608 versus 35428). More than 94 percent of the domain pairs defined by SCOP in the same fold are joined by CATH, but these commonly joined pairs represent only one third of the pairs in the same CATH topology.

Interestingly, at the fold level the similarity based clustering agrees with the two manual classifications better than they agree with each other, with maximum agreement $\kappa = 0.79$ and $\kappa = 0.63$ for SCOP and CATH, respectively. At the cross-over point, the agreement between the automatic classification and SCOP is $\kappa = 0.69$, much larger than with CATH ($\kappa = 0.32$).

If we perform the clustering using single linkage instead of average linkage, the agreement between the automatic clustering and CATH becomes much better ($\kappa = 0.80$ at the maximum and $\kappa = 0.74$ at the stop point), whereas the agreement with SCOP becomes much poorer. Indeed, CATH uses single linkage clustering, i.e., a new domain is joined to the cluster containing the most similar domain if similarity is above a threshold. This explains why CATH joins more pairs of domains than SCOP at the topology level.

If we compare the average linkage with the single linkage clustering as a function of the clustering step, we find that the single linkage joins many more pairs than the average linkage for the same number of clusters, as expected from the fact that it does not penalize the overunification. The weighted kappa between the two algorithms decreases as the clustering proceeds, as shown in Supporting Figure S3. The disagreement between the two classifications is already important before the cross-over point.

These findings shed light on the comparison between CATH and SCOP. Despite their good agreement at the level of superfamily, CATH and SCOP use different criteria for clustering

Table 3. Comparison of the agreement between different classifications.

	Superfam.	Folds
SCOP vs. CATH	0.84	0.48
Automatic (AL) vs. SCOP	0.54	0.69
Automatic (AL) vs. CATH	0.58	0.32
AL (max) vs. SCOP	0.65	0.79
AL (max) vs. CATH	0.64	0.63
Automatic (SL) vs. SCOP	0.24	0.48
Automatic (SL) vs. CATH	0.28	0.70
SL (max) vs. SCOP	0.51	0.67
SL (max) vs. CATH	0.51	0.80

The agreement is evaluated through the weighted kappa parameter, Eq. (19). The first line compares superfamilies and folds from SCOP and CATH. In the two following lines, the automatic classification at the stop point obtained with average linkage (AL) is compared with SCOP and CATH, respectively, at the levels of superfamilies and folds. The two following lines compare the expert classifications with the AL classification at the points where their weighted kappa is maximum. The four last line are the same, but using as clustering algorithm single linkage (SL), which gives a much stronger agreement with CATH than with SCOP at the fold level, consistent with the fact that CATH uses single linkage.

doi:10.1371/journal.pcbi.1000331.t003

superfamilies. They would nevertheless agree better if the clustering would be stopped at large similarity, where transitivity is approximately fulfilled. Therefore, the discrepancy between CATH and SCOP at fold level has two roots (besides the different in domain decompositions): (1) They use different clustering methods, a procedure effectively similar to average linkage for SCOP and single linkage for CATH. which yields a much larger number of pairs classified as the same fold, despite the number of folds is practically the same. (2) They push the clustering up to a low similarity level at which the two clustering methods diverge considerably.

Classification criteria may vary with time. Another possible source of subjectivity in the definition of the fold is the amount of biological knowledge that the expert curators use. To test the influence of this factor, we analyzed how SCOP folds and superfamilies changed through time. We labelled the age of a SCOP fold or superfamily through its SCOP index. Since the SCOP index depends on the secondary structure class, we normalized separately the index for different secondary structure classes, so that a value of 1 means that the index lies within the first 10% of its class and so on. We measured the mean similarity score for pairs of proteins in the same fold or superfamily. The MAMMOTH similarity score of related domains depends on their length. For superfamilies, we find that the average score depends on the average length of the superfamily, L, as $S \approx L^{0.586}$. Since the folds and superfamilies with index in the 7th and 8th interval are characterized by much longer domains (the average length is 270, compared with average lengths between 131 and 188 for all other intervals), we normalized the MAMMOTH similarity score dividing it by $L^{0.586}$, where L is the average length in the cluster.

One can see from Figure 8 that folds classified since longer time (smaller index) tend to be structurally more diverse. They also contain more domains and more superfamilies (data not shown). There are two possible interpretations of these findings. It is possible that some folds are intrinsically more diverse, and that they are more likely to be discovered and studied first, since they contain a larger number of proteins. But it is also possible that the

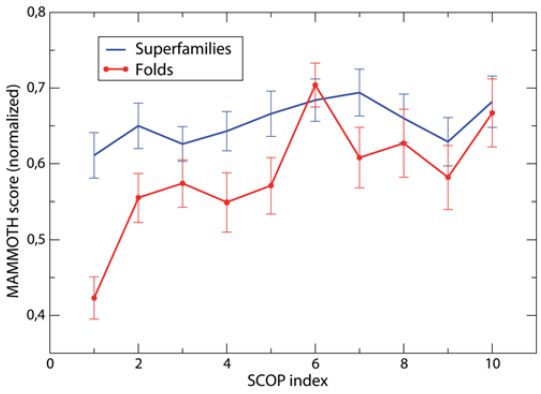


Figure 8. Normalized structural similarity score of the program MAMMOTH (A) and standard deviation of domain length (B) versus the date of the oldest PDB file included in the SCOP fold. Older folds appear to be significantly more structurally diverse, as assessed both through the MAMMOTH score and their length difference.

doi:10.1371/journal.pcbi.1000331.g008

greater biological knowledge available for older folds makes it easier to classify domains in these folds even in the absence of a large structure similarity.

To distinguish between these two interpretations, we measured structure similarity within superfamilies, see Figure 8. Similar as for folds, older superfamilies contain more domains than the more recent ones (11.6 ± 2.2 for the most ancient and 4.1 ± 0.9 for the most recent index interval), but they are not more structurally diverse. This suggests that: (1) Ancient folds are structurally more diverse because they join superfamilies that are more diverse between each other but not within each other. Consistently, ancient folds contain more superfamilies: 3.7 ± 0.8 for folds with the most ancient labels, less than 1.9 ± 0.3 for SCOP labels above the third interval; (2) When there is sequence information to guide the classification, as in the case of superfamilies, the structural diversity remains stable with time, and it does not depend on the size of the superfamily, whereas it changes with time in the case of folds, for which no sequence information is used. This may suggest the existence of a bias to join new superfamilies to a fold known since long time even if the structure similarity is small.

Summarizing, the structure similarity within SCOP superfamilies remained stable through time, whereas the similarity of superfamilies classified into the same fold tends to be lower for ancient folds.

Beyond the Classification: Protein Similarity Network

The cross-over point of transitivity violations determines an intrinsic threshold beyond which protein similarity is better represented as a network rather than as a tree. Protein similarities have been previously represented as a network by other authors. Dokholyan et al. [45] generated the protein domain universe graph using as similarity measure the Z score of the structure alignment program Dali [9]. They found out that, for proper thresholds, the network is scale-free, i.e., the number of links per node is power-law distributed. Performing single linkage clustering over this network, they obtained clusters whose size distribution is also a power-law, reminiscent of the distribution of protein domains per SCOP fold in a genome [43,44]. Krishnadev et al. [49] performed a similar study for the similarity graph of protein

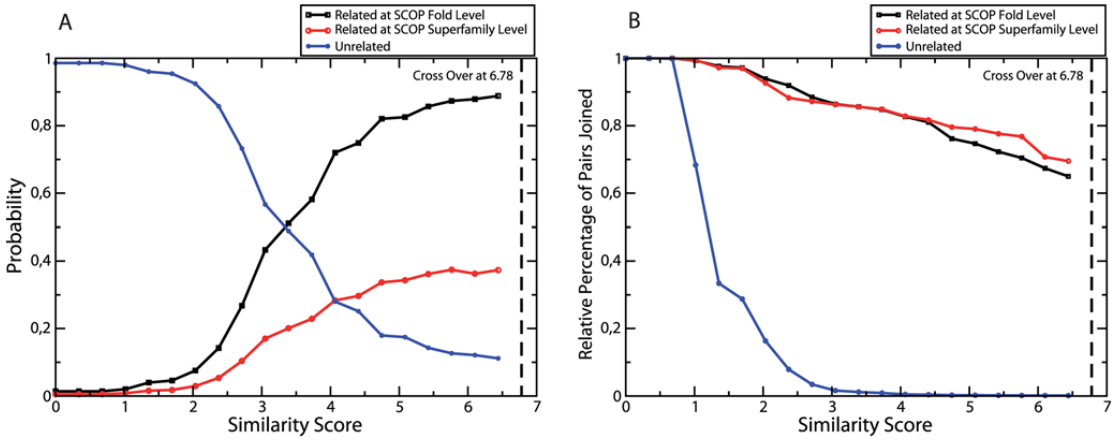


Figure 9. For networks of clusters in the automatic classification joined with the similarity threshold represented in the horizontal axis, we plot in (A) the fraction of links joining clusters that contain two proteins from the same SCOP superfamily (a), the same SCOP fold (b), or different folds (c), respectively; in (B) we plot the probability that a link exists for a pair of clusters of type (a), (b), and (c). In (A), we see that, for $S_0 < 3.5$, the majority of links are from clusters unrelated in SCOP.
doi:10.1371/journal.pcbi.1000331.g009

chains instead of protein domains. They also found scale-free behavior at large enough similarity threshold. They used spectral analysis of the adjacency matrix to partition the graph into clusters.

In contrast to these previous approaches, the graph presented here is not a preliminary step for clustering, but it represents the significant similarity relationships for which clustering is not justified. These relationships not only allow to recover relationships present in expert classifications, such as splitted superfamilies and folds, but also allow to treat on the same ground the cross-fold relationships discussed by several authors, which go beyond expert classifications.

We construct the similarity network by connecting the clusters of the automatic classification that have significant structural similarity. As the similarity threshold is decreased, more and more clusters are connected. Pairs of clusters containing structures from a superfamily splitted in the automatic classification get unified in the network. We measured the probability that a pair of domains is joined in the network as a function of the similarity threshold, distinguishing pairs of domains from the same superfamily, from the same fold, or from different folds. (see Figure 9). Only for similarities as low as $S_0 \approx 2.5$, more than 90% of the domains in the same superfamily are joined. However, already for similarities $S_0 < 3.5$ the majority of the joined domains are from different folds. A reasonable threshold for significant structure similarity, mostly corresponding to pairs of different folds, seems to be S_0 between 3 and 4. Results presented here are obtained using $S_0 = 4$ as threshold for significant structure similarity.

A visual representation of such a network is shown in Figure 10B. One can see that almost all of the structure space is connected, but there is still some structure appearing. If we use a higher similarity threshold but still below the cross-over, such as $S_0 = 6$, the resulting network contains several linear motifs clearly expressing transitivity violations, with *a* connected to *b*, *b* to *c*, *c* to *d*, and so on, but without direct connection between *a* and *c* or *a* and *d*. For comparison, we also show in Figure 10A the network constructed joining clusters at high similarity before the cross-over point ($S_0 \approx 10$) using as threshold the cross-over similarity, $S_0 = 6.78$. This network presents many regions with high density of links, representing clusters that have still to be joined,

In the context of network analysis, the transitive property studied in this paper is analogous to the clustering coefficient (see Methods). Clustering coefficient equal one means that the network is transitive, i.e., if *a* is connected with *b* and *b* is connected with *c*, also *a* is connected with *c*. The high siilarity network obtained before the cross-over point has a high mean clustering coefficient equal to 0.69, which decreases to 0.36 for the network after the cross-over. In general, as one could expect, the clustering coefficient increases with the similarity threshold S_0 (see Figure S1). However this increase is smooth, so that we can not use the clustering coefficient to detect the cross-over point.

Interestingly, the network allows not only to recover similarity relationships at the superfamily and fold level that are below the threshold for clustering, but it may also help to discover new evolutionary or functional relationships that are not contained in SCOP or CATH. For instance, in a recent paper Xie and Bourne proposed a new method to detect remote evolutionary relationships based on the structure similarity of the active site [50]. Using this method, they confirm a previously proposed evolutionary relationship between SCOP superamily Phosphoenolpyruvate carboxykinase (PCK) and the P loop containing nucleotide triphosphate hydrolase (NTH) superfamily. The PCK domain 1ayl_1 used as a seed by Xie and Bourne is joined in the automatic classification with domains 1knxa2 and 1ko7a2, which are classified in SCOP in the PCK superfamily but are classified in CATH in the NTH superfamily. The automatic classification supports the CATH classification. This cluster has a single significant structural link, with average similarity $S = 5.0$, with a cluster containing only domains classified in the NTH superfamily in both CATH and SCOP, and through this link another step connects it to many other clusters in the NTH superfamily or in the NTH fold. The relevant part of the network is represented in Figure S4, from which it is clear that the structurally consistent clusters joined in a network give a richer evolutionary information than a unique fold.

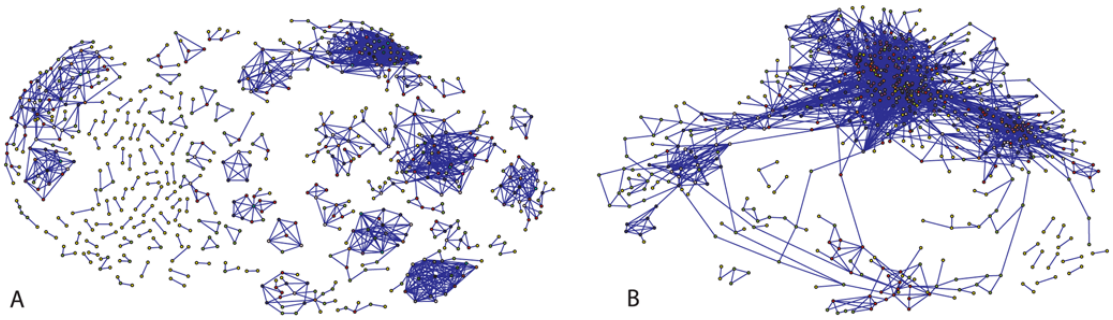


Figure 10. Networks of protein clusters similarities. (A) High similarity clusters ($S=10$) linked using as a threshold the cross-over similarity, $S_0=6.78$. (B) Cross-over clusters ($S=6.78$) linked below the high transitivity regime, up to $S_0=4$.
doi:10.1371/journal.pcbi.1000331.g010

In order to complement structure information with sequence information, we constructed the network connecting clusters that have members belonging to the same superfamily. The networks based on sequence and structure similarity can be accessed at the url http://ub.cbm.uam.es/research/ProtNet.php

Transitivity violations and protein modularity. To investigate protein modularity, we studied the triangles that violate transitivity for a specific threshold S_0, in the sense that $S(a,b)>S_0$, $S(b,c)>S_0$, but $S(a,c)<S_0$. For such triangles, we tested whether the regions of the intermediate structure b having a good match with structures a and c are the same or they are different, by measuring the overlap between these two regions as

$$Q_{bca}=\frac{\min(\mathrm{end}_{ba},\mathrm{end}_{bc})-\max(\mathrm{ini}_{ba},\mathrm{ini}_{bc})}{\min(\mathrm{end}_{ba}-\mathrm{ini}_{ba},\mathrm{end}_{bc}-\mathrm{ini}_{bc})} \qquad (4)$$

where the initial and final residues of the matching regions are denoted as ini_{ba}, end_{ba}, ini_{bc} and end_{bc}, respectively. The value $Q_{bca}=1$ means that all three structures all share the same core over which they are similar. In contrast, the value $Q_{bca}=0$ means that the intermediate structure b shares completely different fragments with structures a and c. This is the most dangerous case for clustering algorithms, which can run the risk to join two structures that do not share any common region. One such example, with ASTRAL codes d1mt5a_, d1bif_1 and d1b3qa1, is shown in Figure 11.

The distribution of the fragment overlap Q_{bca} is bimodal, with peaks at $Q_{bca}=1$ and $Q_{bca}=0$ (see Figure 12). However, triangles with $Q_{bca}=0$ are very rare for large similarity $S_0=10$, where they may correspond to errors in domain decompositions, whereas they become more frequent for similarities below the cross-over point.

Thus, beyond the cross-over point it is likely to find severe violations of transitivity in which two significant matches *ab* and *bc* fall in two completely different regions of protein *b*, consistent with the idea that transitivity violations and the consequent continuity of protein structure space stem from the modularity of proteins. These significant and disjoint partial matches offer a way to operatively define substructures below the domain level. A more detailed study of substructures based on their recurrence will be presented elsewhere.

Discussion

Transitivity Violations

As for all problems for which hierarchical clustering algorithms are applied, for clustering protein structures it is of key importance to determine up to which point the clustering is justified. We propose to test the internal consistency of a clustering method based on a similarity measure by testing the transitive property, which requires that whenever a is similar to b and b is similar to c, then a must be similar to c. Only if the transitive property holds a hierarchical classification can be unambiguously built. If the transitive property is violated for an extensive number of triangles, hierarchical clustering is frustrated [38], and we expect that there is a very large number of unrelated and almost optimal classifications, in each of which a similar number of similarity relationships are violated. We proposed here Eq. (1) to quantify the violations of transitivity of a group of three elements, and Eq. (2) to quantify the violation of transitivity when two clusters are joined.

Transitivity violations as defined here occur either when a pair of domains is joined below the similarity threshold, or when a pair is separated above the same threshold. Another definition, common in the context of sequence comparisons, considers that transitivity is violated only when pairs are separated above threshold. This definition is motivated by the fact that significant sequence similarity demonstrates almost certainly an evolutionary relationship, whereas the lack of similarity does not exclude it. With this definition, the single linkage algorithm does not produce any transitivity violation, since it joins all pairs above threshold. In fact, the term transitivity is often used as a synonymous of single linkage clustering.

Nevertheless, several reasons make the definition of transitivity adopted here more suitable in the context of structure classification. The first reason also applies to sequence comparisons, and it is based on protein modularity. If a domain b is made of two fragments A and C, with A similar to domain a and C similar to domain c, single linkage will infer a non existing relationship between a and c. Indeed, for applying single linkage clustering to the triangle *abc*, one has to check whether the fragment overlap Q_{bca}, Eq. (4), is also significant. Secondly, single linkage joins many structures that are not significantly similar, producing clusters that are not structurally consistent. These clusters may lack a common core, as it is often found applying multiple structure alignment algorithms to SCOP and even more CATH superfamilies. For the goal of modelling, it may not be convenient to join structurally dissimilar domains in the same fold, since this would increase the likelihood of selecting wrong templates. The study of structure evolution is made more difficult when structural variation is hidden inside a very diverse cluster, whereas well defined clusters connected by links expressing evolutionary relationships may represent a better framework for the study of structure divergence.

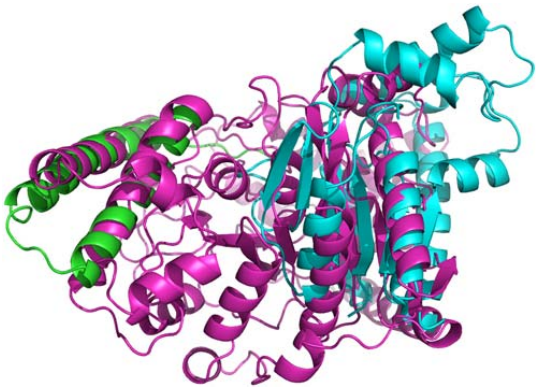


Figure 11. Example of three domains that violate transitivity with $Q=0$. They are joined after the cross-over point in the network built using similarity threshold $S_0=5$. The ASTRAL codes are d1mt5a_ (a), d1bif_1 (b) and 1b3qa1 (c). The bigger domain d1mt5a_ (red) links in the network the two smaller domains, which deviate considerably from each other as they don't share any significant part of structure between them. It holds $S(a,b)=5.75$ (red and blue), $S(a,c)=5.95$ (red and green) and $S(b,c)=2.8$ (blue and green), which violates transitivity.
doi:10.1371/journal.pcbi.1000331.g011

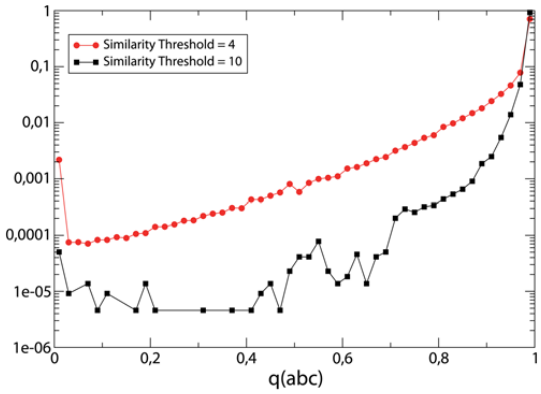


Figure 12. For networks defined through the condition $S(a,b)>S_0$, with $S_0=10$ and $S_0=4$, respectively, and for all triangles that violate the transitive property, i.e., $S(a,b)>S_0$, $S(b,c)>S_0$ and $S(a,c)<S_0$, we measured the overlap Q_{abc} between the two relevant matches of the intermediate structure b, Eq. (4). The peaks of the distribution at $Q=0$ and $Q=1$ correspond to matches over completely different and exactly the same region of protein b, respectively.
doi:10.1371/journal.pcbi.1000331.g012

Cross-Over from Discrete Sets to Continuous Space

We have observed that the transitivity violations grow while the clustering algorithm joins protein domains into clusters. Interestingly, in all instances that we studied we have found a cross-over between two regimes of slow and fast increase of transitivity violations.

1. At high similarity, transitivity violations grow slowly as the clustering algorithm proceeds, and domain size does not vary very much within a cluster. Clusters in this regime mostly correspond to subsets of SCOP superfamilies. Therefore, most domains in the same cluster are related through gene duplication and subsequent divergence, which justifies to classify related domains on a tree.
2. At low similarity, transitivity violations grow rapidly as the clustering algorithm proceeds, and domains in the same cluster differ substantially in size. Many pairs in the same cluster are related through partial substructures.

We propose that the cross-over in transitivity violations is an intrinsic point to stop the automatic classification. Lower similarity relationships should be represented as a network rather than a tree.

Influence of the Methodology

The method that we presented requires several arbitrary choices. In order to test its robustness, and the influence of the parameters, we have studied at least two alternatives for each of these choices. Qualitatively similar results were obtained for several similarity scores computed on two different alignments obtained with a local and a global version of the MAMMOTH algorithm. Both alignment algorithms were developed at our group. We did not test whether alignments obtained with algorithms developed by other groups, such as DALI, yield different conclusions, as they might do.

In all cases that we tested, we have observed a cross-over in transitivity violations, finding that most of the clusters at the cross-over point correspond to subsets of SCOP or CATH superfamilies. However, the exact location of the cross-over point and the quality of the clustering, as assessed through the clustering coefficient and through the mean value of the transitivity violations, varies for different choices.

Although we do not aim at reproducing SCOP or CATH, which we believe is impossible, we recognize that these expert classifications have important merits. It is therefore noteworthy that the highest clustering coefficients and lowest transitivity violations tend to be associated with scores that are better compatible with SCOP or CATH classifications.

The first important choice is the structure alignment algorithm. Computationally, structure alignment is an NP-complete problem, and even if it were exactly solved different algorithms would differ, since they optimize different scores. We used two versions of the algorithm MAMMOTH that are quite different, since one optimizes local superimmposition of heptamers whereas the second one, MAMMOTH-mult, otpimizes the global structure superimposition, achieving alignments with better PSI and contact overlap. Despite this important difference, the results obtained with the two methods are rather similar.

The similarity measure used is probably the most relevant choice, and we tried several of them. We obtained better results with the contact overlap than with measures that score the optimal spatial superimposition of the two structures, which are used in the standard MAMMOTH score. We conjecture that the contact overlap is a better measure than the PSI for clustering protein structures because of three reasons: (1) It does not assume that there is an optimal rigid body superimposition between the two structures. In doing so, it implicitly allows for flexible superimpositions, which might be better suited for detecting evolutionary relationships [51–54]. (2) It weights the residues in the core of the protein more than loop residues, since the former have a larger number of contacts. (3) The parameter it depends on, i.e., the threshold at which two residues are considered in contact, has a physical meaning in terms of interatomic interactions, and it is therefore less arbitrary than the tolerance parameter of the PSI,

i.e., the threshold below which two residues are considered to be superimposed.

Similarity scores based on structure superimposition typically need a tolerance threshold to decide whether two residues superimpose. We tested the TM score [42], which uses a length dependent threshold that makes this score almost independent of the size of the aligned proteins. The results obtained with this score are very similar to those obtained with the contact overlap. In contrast, the percentage of structure identity (PSI) adopts a fixed tolerance threshold, usually chosen as 4Å . To study the effect of this parameter, we repeated our numerical experiments with a more tolerant threshold of 6Å . Not surprisingly, the more tolerant similarity measure makes the space more continuous, decreasing the clustering coefficient and increasing the transitivity violations. Therefore, the cross-over from the discrete to the continuous regime occurs at higher similarity, which means that protein domains are splitted into a larger number of clusters. In this case as well, the cross-over is clear and the clusters at the cross-over are mainly subsets of superfamilies.

All measures, except the TM score, must be normalized in order to make them independent of the length of the aligned proteins. We implemented this through a length dependent Z score, as in the original MAMMOTH score. The drawback of the Z score is that not only it makes the similarity of unrelated proteins almost independent of length, but at the same time it reduces the similarity of related proteins with short length. In this way, the similarity of related proteins depend on their length and not on their evolutionary divergence, which makes the Z score an unsuitable measure for evolutionary analysis. This drawback does not occurr with the TM score, although this does not necessarily imply that it is a suitable measure for evolutionary analysis.

Last, we have to decide which clustering algorithm we use. If we adopt the definition of transitivity proposed in the present work, the average linkage algorithm has to be preferred over both single linkage and complete linkage. In fact, average linkage reduces the combination of splitting and overunification errors, whereas single linkage only eliminates splitting errors, since it joins all pairs above the similarity threshold, and the complete linkage eliminates overunification errors, since it separates all structures that are below the similarity threshold. Interestingly, from our analysis it turns out that the main difference between SCOP and CATH is that the latter uses single linkage, while the former uses some procedure effectively similar to average linkage.

As a last remark, we note that there is some analogy between our method, which uses transitivity violations to detect the point at which hierarchical clustering is not justified, and the bootstrap method that scores the significance of each cluster in a tree. Nevertheless, there are also important differences. Besides the fact that bootstrap is computationally much more cumbersome than our method, for obtaining a classification with the bootstrap method we would have to fix a threshold bootstrap probability to accept one cluster, whereas the cross-over that we obtain with our method arises in a natural way without fixing an arbitrary threshold.

Perspectives for the Automatic Classification of Proteins

The existence of two regimes of transitivity violations, and the fact that the automatic classification at the cross-over point mostly consists of sets of SCOP or CATH superfamilies are the main results of this work. They are robust with respect to changes in the clustering algorithm, the similarity measure, the set of protein domains that we automatically classify, and the accuracy of the alignment algorithm. These results suggest that it is possible to automatically and objectively define an equivalence class for protein domains up to the similarity corresponding to the cross-over point.

Clusters in the automatic classification are structurally more consistent than SCOP folds or CATH topologies, mainly because of two reasons. (1) In the automatic classification, almost 15 percent of superfamilies are split into structurally divergent clusters, indicating that there can be important structural changes in protein evolution [32,33,36]. Interestingly, domains in split superfamilies tend to have larger size difference between each other, suggesting that insertions and deletions play an important role for structural divergence, consistent with recent analysis [55,56]. (2) Only 44 percent of the pairs of domains in different SCOP superfamilies and the same SCOP fold are joined in the automatic classification. This percentage becomes much smaller for CATH (less than 11 percent), whereas 68 and 66 percent of the pairs in the same SCOP or CATH superfamily are joined in the automatic classification The similarity between most of the pairs that are not joined is significant, but it is at the level where transitivity violations are large and a network fits the data better than a classification. Our analysis thus suggests that CATH and SCOP classify proteins up to similarities that are below the cross-over of transitivity violations. The same is possibly true for the automatic FSSP classification as well, where proteins are classified in the same fold if the Z score of their similarity is above 2. This is the smallest threshold at which the structures compared are significantly related. Here we also use a Z score, but we find that the cross-over point is at $Z_0 = 6.78$ implying that the transitive property is severely violated at the similarity level $Z = 2$.

An indication that the fold defined in expert classification may not correspond to an intrinsic similarity level is that CATH and SCOP neatly agree at the level of superfamily, as assessed through the weighted kappa measure, but they disagree between each other at the level of fold even more than they disagree with the automatic classification, when the proper clustering algorithm is used. Indeed, the main difference between SCOP and CATH at fold level is that SCOP uses a procedure effectively similar to the average linkage algorithm, whereas CATH uses the single linkage algorithm, which does not penalize the joining of structurally distinct domains, resulting in clusters that are structurally very diverse.

Furthermore, we have shown that the structural diversity within a SCOP fold is larger if the fold was defined since longer time, suggesting that the criteria underlying the definition of fold may change through time. Classifications are very useful, but the present analysis supports the view that the low similarities at the fold level are better represented as a network rather than as a tree.

Possible Improvements of the Automatic Classification

The comparison between the automatic and the expert classifications also indicates that the automatic classification can be improved along three lines.

First, in the present study we considered protein domains as defined in the SCOP and CATH classifications. However, proteins are split into domains in the two schemes in a rather different way. In particular, some domains defined in the SCOP classification appear by visual inspection to consist of more than one domain. An incomplete domain partition can be an important source of transitivity violations and consequent errors in an automatic classification of protein structures. We are developing a new automatic method for decomposing proteins into domains based on their recurrence in a database of unrelated structures, similar to the method proposed by Holm and Sanders [57]. The domains obtained in this way will be subject to further

decomposition based on their structure, to obtain a set of domains to which we will apply our clustering procedure.

Secondly, our method tends to split superfamilies constituted of short domains. Some of these splitting appear to be due to the dependency of the similarity score on the protein length. The raw similarity score, either PSI or contact overlap, is transformed into a Z score in order to reduce as much as possible the dependency of the score of unrelated structures on their size. Our results show that the classification deteriorates if this normalization is not properly performed. However, due to this normalization the similarity score corresponding to identical structures decreases for decreasing domain size, which makes it more difficult to cluster together short proteins. In order to overcome this problem, it would be very helpful to define a similarity score that is independent of domain size both for unrelated and for closely related structures. This will be presented in a forthcoming work.

Third, we found 63 over 779 clusters that contain protein domains defined by SCOP curators as different folds (although 27 of these clusters are homogeneous in terms of CATH topologies). The distribution of structure similarity suggests that several of the foreign domains appearing in clusters that are mostly from another fold are characterized by low mean similarity, and that it could be possible to "clean" the clusters of the automatic classification. Preliminary results indicates that this strategy is promising.

Protein Domain Networks

Significant sequence or structure similarity below the threshold for clustering [14,15] constitutes a very valuable information for evolutionary or functional studies. In the CASP and SCOP database, these significant cross-fold similarities are not available. We present this information in the form of two networks with structure-based and sequence-based links between the clusters of the automatic classification. In this way, we can recover not only superfamily and fold relationships that are not present in the automatic classification, but also new relationships that are not reported in expert classifications.

Two Modes of Protein Evolution?

As a concluding remark, we note that the two regimes of transitivity violations that we found can be related with two modes of protein domain evolution. In the regime of large structure similarity, transitivity violations are small, related domains are similar in size, and 95 percent of them contain domains from a single CATH or SCOP fold, whereas 86 percent contain evolutionarily related domains from the same superfamily. These results indicate that most of the domains with structure similarity above the cross-over are evolutionarily related through gene duplication and divergent evolution. Moreover, domains in different superfamilies but same fold can not be excluded to be evolutionarily related, and some careful studies have been able to demonstrate this common origin also in the absence of a clear signal from sequence similarity, as in the case of the study of TIM-barrels conducted by Nagano et al. [58]. This view also agrees with the results by Deeds et al. [59], who tested models of convergent and divergent evolution using statistical properties of protein structural clusters, finding that the data support divergent evolution [60]. We summarize these findings saying that, for large similarity, protein domain evolution is mostly uniparental.

On the other hand, similarities below the cross-over of transitivity violations are often due to partial substructures, and the typical size difference between related domains raises from 20 to 40 residues, indicating the occurrence of large insertions and deletions when the related domains belong to the same superfamily. These are clues of multi-parental evolution, proceeding through the assembly of new polypeptide fragments. This hypothetical mechanism has been proposed by Lupas et al. for the evolution of early protein domains through assembly of small peptide fragments [28]. Our findings suggest that it can also be extended to more recent evolution, consistent with another recent study [15]. In this regime the domain structure space should be regarded as continuous, and significant structure similarity should be described as a network rather than a tree.

These considerations parallel recent considerations about the classification of organisms on the tree of life [61]. Speciation and evolutionary divergence generate a tree of species, which can be reconstructed by estimating the time of divergence from the molecular sequences of their genes. In order to do this, one has to use a proper sequence distance, approximately ultrametric, which makes species classification possible on a rigorous basis. Nevertheless, this view of the tree of life has been recently challenged by the discovery of the high rate of horizontal gene transfer in genome evolution. Due to horizontal gene transfer, genome evolution is multiparental, and genes that have been subject to gene transfer can not be used to reconstruct the phylogenetic tree. The extensive presence of horizontal gene transfer in evolution has led Doolittle to propose that the evolutionary relationships between organisms should be regarded as a net of life rather than a tree [61]. The present work suggests that, in the context of protein domain evolution, a tree scenario of uniparental divergent evolution is suitable to represent high similarity relationships, but a pluriparental network emerges for more remote relationships.

Methods

Datasets

We have used two non redundant sets of protein domains. The first set was obtained from the ASTRAL 40 database, in which no pair has sequence similarity larger than 40%. We used the SCOP version 1.65 and selected only domains from the four main SCOP classes, all α, all β, α/β and $\alpha+\beta$. The second set is the non redundant set of domains from the CATH classification, with sequence similarity smaller than 35%. Also in this case we excluded domains outside the four main classes. The final number of domains was 5041 for the SCOP set and 7073 for the CATH set.

Consensus Set between CATH and SCOP

In order to select a set of domains consistently defined in SCOP and CATH, we aligned with BLAST [62] the sequences of domains in the non redundant ASTRAL40 database against domains in the non redundant CATH database at 35% sequence identity. We identified two domains to be equivalent if their BLAST evalue was smaller than 10^{-3}, with sequence identity larger than 75%, and their size differed by less than 10%. In this way we have obtained a set of 2890 non redundant domains classified in 779 SCOP superfamilies, 466 SCOP folds, 885 CATH superfamilies and 473 CATH topologies.

Similarity Scores

We performed pairwise structure alignments using either the program MAMMOTH [41], which is the fastest program of protein structure alignment that we know, or its multiple alignment version MAMMOTHmult [39], which is a bit slower but much more accurate.

The MAMMOTH similarity score is based on the number of aligned residues that are closer than 4Å after optimal spatial superimposition of structures a and b, L_{ab}^{matched}. This is transformed into a percentage of structure identity (PSI) dividing

it by the length of the shortest structure,

$$\mathrm{PSI}_{ab}^{\mathrm{partial}} = \frac{L_{ab}^{\mathrm{matched}}}{\min(L_a, L_b)}. \quad (5)$$

$\mathrm{PSI}^{\mathrm{partial}}$ equals one if the two structures coincide over the length of the shorter one. There is no penalization for additional residues in the longer structure, i.e., the score is sensitive to good partial matches and we call it partial PSI. However, the fact that the score does not penalize inserted regions may lead to join domains with very large length difference. To tackle this problem, we also defined the total similarity score, which penalizes regions in the larger structure that are not matched by the short one:

$$\mathrm{PSI}_{ab}^{\mathrm{total}} = \frac{L_{ab}^{\mathrm{matched}}}{\sqrt{L_a L_b}} \quad (6)$$

$\mathrm{PSI}_{ij}^{\mathrm{total}}$ equals one only if the match completely covers the longer protein.

Third, we adopted the contact overlap, which counts the fraction of contacts in common between two aligned structures a and b. Also this score is normalized in such a way to penalize partial matches. We defined the contact matrix $C_{ij}^{(a)}$ of protein a such that $C_{ij}^{(a)}$ equals one if two heavy atoms of residues i and j are closer than 4.5Å and $|i-j| \geq l$, and zero otherwise. We considered two cases, $l=4$ and $l=6$. In this last case, intrahelical contacts are not considered. Denoting by $A(i)$ the residue in structure b aligned with residue i in structure a, the contact overlap can be written as

$$q_{ab} = \frac{\sum_{ij} C_{ij}^{(a)} C_{A(i)A(j)}^{(b)}}{\sqrt{\sum_{ij} C_{ij}^{(a)} \sum_{ij} C_{ij}^{(b)}}}. \quad (7)$$

The main qualitative difference between the contact overlap and the PSI is that in the contact overlap superimposed residues in the core of the protein, which form many contacts, receive a larger weight.

It is crucial for protein structure classification that the distribution of the similarity score used is almost independent of the length for comparisons of unrelated proteins. The MAMMOTH score takes care of this by normalizing the PSI in such a way that the distribution of the normalized PSI is almost independent of size for unrelated pairs:

$$S_{ab} = \frac{\mathrm{PSI}_{ab} - A L_{ab}^{-\alpha}}{B L_{ab}^{-\beta}} + C \quad (8)$$

where $L_{ab} = \min(L_a, L_b)$ in the case of the partial PSI, and $L_{ab} = \sqrt{L_a L_b}$ in the case of the total PSI. In the case of the overlap, we also used $L_{ab} = \sqrt{L_a L_b}$ as a normalization. The exponents α and β depend on the raw similarity score and on the alignment algorithm used, and they were determined by fitting the mean and standard deviation of the PSI of unrelated structures having L_{ab} in some given interval, using the best fit between a Gaussian fit or an Extreme Value statistics fit (see Table 4).

Using Gaussian statistics, we fit

$$\langle \mathrm{PSI} \rangle \approx A L^{-\alpha} \quad \sigma_{\mathrm{PSI}} \approx B L^{-\beta}, \quad (9)$$

Table 4. Size normalization of similarity scores.

Score	Normalization	Alignment	A	α	B	β
PSI partial	EV	Pair	5.97	0.720	0.920	0.634
PSI partial	EV	Mult	5.73	0.714	0.860	0.622
PSI total	EV	Pair	6.48	0.722	0.972	0.662
PSI total	EV	Mult	5.62	0.729	0.961	0.659
Overlap	Gauss	Pair	0.375	0.535	1.340	0.676
Overlap	Gauss	Mult	0.752	0.576	1.874	0.773

The reported parameters were used to normalize the raw scores according to Eq. (8).

doi:10.1371/journal.pcbi.1000331.t004

and using Extreme Value statistics, we fit

$$\langle \mathrm{PSI} \rangle - \frac{6 \times 0.5772}{\pi} \sigma_{\mathrm{PSI}} \approx A L^{-\alpha} \quad \sigma_{\mathrm{PSI}} \approx \frac{\pi}{6} B L^{-\beta}, \quad (10)$$

The domain similarity score of domain a in cluster A is defined as the average pairwise similarity between domain a and all other domains in the cluster,

$$S(a,A) = \frac{1}{(n_A - 1)} \sum_{b \in A, b \neq a} S_{ab} \quad (11)$$

Clustering Algorithms

We programmed and tested three hierarchical clustering algorithms: average linkage [63], single linkage and complete linkage. Starting from each element being a separate cluster, at each step t all algorithms join the two most similar clusters A and B, and compute the similarity between the new combined cluster and all other clusters in a way that depends on the clustering algorithm.

With **average linkage**, the combined similarity is computed as the average similarity with the two joined clusters,

$$S^{t+1}(AB,C) = \frac{n_A S^t(A,C) + n_B S^t(B,C)}{n_A + n_B}, \quad (12)$$

where t labels the step of the algorithm, A and B are the clusters that are joined, n_A and n_B is the number of elements they contain, AB denotes the new composite cluster, and C is any other cluster. Note that this updating rule is equivalent to computing the new similarity score as the average between the similarity between all pairs of elements from the cluster C and the cluster AB.

With **single linkage**, the combined similarity is the largest similarity in the set, so that two sets are joined if at least one pair of elements is above threshold

$$S^{t+1}(AB,C) = \max(S(A,C), S(B,C)) \quad (13)$$

With **complete linkage**, the combined similarity is the smallest similarity in the set, so that two sets are joined if all pairs of elements are above threshold

$$S^{t+1}(AB,C)=\min(S(A,C),S(B,C)) \quad (14)$$

Ultrametricity

An ultrametric set is a set X with an associated distance measure $d(a,b)\geq 0$ where every triplet of points a, b and c fulfils a property stronger than the ordinary triangular inequality: each side of a triangle is smaller than the larger between the other two sides, i.e., $d(a,c)\leq\max(d(a,b),d(b,c))$. This implies that the two longer sides must be equal. In particular, for an ultrametric set and for every threshold $\gamma>0$, it holds that if $d(a,b)\leq\gamma$ and $d(b,c)\leq\gamma$, then $d(a,c)\leq\gamma$. Consider now the cluster containing all elements within a distance γ from element a, $C_\gamma(a)=\{b\in X|d(a,b)\leq\gamma\}$. It is easy to see that, for every pair of points a and b, either $C_\gamma(a)$ and $C_\gamma(b)$ coincide, or they do not share any point. Therefore, $d(a,b)\leq\gamma$ is an equivalence relationship, since if $c\in C_\gamma(a)$ then it must also be $c\in C_\gamma(b)$, and the set of points can be considered discrete.

Clustering Coefficient

A concept related to transitivity in the context of networks is the clustering coefficient, which can be computed through the formula

$$\text{Clustering coefficient}=\frac{1}{N}\sum_i\frac{2\sum_{j<k}A_{ij}A_{ik}A_{jk}}{n_i(n_i-1)} \quad (15)$$

where N is the number of nodes in the network, labelled as i, j and k, A_{ij} is the adjacency matrix (one if i and j are joined, zero otherwise), $n_i=\sum_j A_{ij}$ is the number of neighbors of node i, and the clustering coefficient of node i is the fraction of pairs of its neighbors j and k that are neighbors between each other. If the clustering coefficient is one for all nodes, connections on the network define an equivalence relationship.

We have computed the clustering coefficient for the network obtained by joining domains with similarity $S_{ij}>S_0$, for various values of S_0. To compare different similarity measures, we have plotted the clustering coefficient versus the number of clusters obtained through single linkage clustering with the same threshold S_0.

Detecting the Cross-Over Point

For detecting the cross-over point of transitivity violations (TV), we first measure TV at each step of the clustering algorithm using Eq. (2). We then perform two-pieces exponential fits of TV versus the step t, as $\text{TV}\approx f(t,t_0)=\theta(t_0-t)\exp(a_1t+b_1)+\theta(t-t_0)\exp(a_2t+b_2)$, where $\theta(x)$ is zero for negative x and one otherwise. Fits are performed for all possible cross-over points t_0, and their quadratic error is measured as

$$\text{Error}(t_0)=\frac{\sum_t(\text{TV}(t)-f(t,t_0))^2}{\sum_t\left(\text{TV}(t)-\overline{\text{TV}(t)}\right)^2}, \quad (16)$$

where $\overline{\text{TV}(t)}$ is the mean value of TV. To find the optimum t_0 in a robust way, we perform a cubic fit of the error function in an interval I centered around the step $t_{\min}$ yielding the minimum error, and such that $\text{Error}(t_0)\leq\text{Error}(t_{\min})+0.005$ for all $t_0\in I$. The analytic minimum of this cubic fitting is then selected as the best first estimate of the cross-over point.

The last points in the $\text{TV}(t)$ curve, where the transitivity violations approach the maximum possible value, are very badly fitted through the two-pieces fit. Therefore, we refined the estimate of the cross-over point by removing the outliers of the optimal fit, with the conditions that a point is removed if its residual with respect to the optimal fit is more than three times larger than the median, which is the condition used to define type-1 outliers. We then apply the procedure described above to the reduced set of points, and we determine the cross-over point at which the clustering is stopped.

Weighted Kappa

We assessed the agreement of two classifications through the weighted kappa measure [64], which uses as reference the expected agreement for two independent classifications with the same number of relationships. We define N_A (N_B) the number of related pairs in classification A (B) of the same N objects, with $N_{\text{tot}}=N(N-1)/2$ pairs in total. If A and B are independent, the number of pairs that are either related or unrelated in both A and B is given by

$$N_e=\frac{N_AN_B+(N_p-N_A)(N_p-N_B)}{N_{\text{tot}}} \quad (17)$$

We compare this number to the observed number of pairs that agree,

$$N_o=N_{AB}+(N_{\text{tot}}-N_A-N_B+N_{AB}), \quad (18)$$

where N_{AB} is the number of pairs that are related in both classifications. >From this number, the weighted kappa is computed as

$$\kappa=\frac{N_o-N_e}{N_{\text{tot}}-N_e}. \quad (19)$$

A value of zero means that two classifications are as related as independent classifications, one means that the two classifications coincide. Using the weighted kappa, we have compared the classification obtained at every step of the clustering algorithm with the manual classifications of CATH and SCOP at the superfamily and the fold level.

Notice that the weighted kappa can be decomposed into the contributions of related and unrelated pairs as follows:

$$\kappa=w_{\text{rel}}\frac{N_{AB}-N_e^{\text{rel}}}{N_A-N_e^{\text{rel}}}+w_{\text{unrel}}\frac{(N_{\text{tot}}-N_A-N_B+N_{AB})-N_e^{\text{unrel}}}{N_{\text{tot}}-N_A-N_e^{\text{unrel}}}. \quad (20)$$

where $N_e^{\text{rel}}=N_AN_B/N_{\text{tot}}$ is the number of pairs related in both classifications expected by random, $N_e^{\text{unrel}}=N_e-N_e^{\text{rel}}$, and the weights are $w_{\text{rel}}=(N_A-N_e^{\text{rel}})/(N_{\text{tot}}-N_e)$ and $w_{\text{unrel}}=(N_{\text{tot}}-N_A-N_e^{\text{unrel}})/(N_{\text{tot}}-N_e)$ for related and unrelated pairs, respectively.

Network Analysis

For the sake of illustration, we have represented two domain similarity networks obtained before and beyond the stopping point of the automatic classification.

Two networks were constructed by considering each cluster as a node, and connecting nodes with $S>S_0$. In the first case, we used

clusters obtained before the cross-over point of the average linkage algorithm using a high similarity threshold $S=10$, and we connected them if $S_0>6.78$, which is the similarity at the cross-over point. In the second case we used clusters generated at the cross-over point and we connected them with $S_0=4$. The networks have been visualized using the Pajek software [65].

Other Methods

To visualize spatial superimpositions, we used the multiple structure allignments program MAMMOTHmult [39] in combination with the Pymol software.

Supporting Information

Figure S1 Clustering coefficient for three different similarity measures. The clustering coefficient is computed for networks in which domains with similarity above S_0 are connected, and it is plotted as a function of the number of clusters obtained with single linkage clustering of the same network.
Found at: doi:10.1371/journal.pcbi.1000331.s001 (0.02 MB PDF)

Figure S2 Transitivity violations versus the step of the clustering algorithm for three different clustering algorithms. The smallest violations are obtained with the average linkage algorithm.
Found at: doi:10.1371/journal.pcbi.1000331.s002 (0.19 MB PDF)

Figure S3 Agreement between the classifications obtained with different clustering algorithms at the same step. The best agreement is between single linkage and complete linkage.
Found at: doi:10.1371/journal.pcbi.1000331.s003 (0.05 MB PDF)

Figure S4 Network of protein clusters joining superfamilies NTH and PCK. Xie and Bourne confirmed a previously proposed evolutionary relationship between a member of SCOP superamily Phosphoenolpyruvate carboxykinase (PCK), with code 1ayl_1, and the P loop containing nucleotide triphosphate hydrolase (NTH) superfamily. PCK domain 1ayl_1 is joined in the automatic classification with domains 1knxa2 and 1ko7a2, which are classified in SCOP in the PCK superfamily but are classified in CATH in the NTH superfamily. The automatic classification supports the CATH classification. This cluster has a single significant structural link, with average similarity $S=5.0$, with an cluster containing only domains classified in the NTH superfamily in both CATH and SCOP, and through this cluster another step connects it to many other clusters in the NTH superfamily or in the NTH fold. Here we represent the relevant part of the network. The hybrid cluster containing domain 1ayl_1 is close to the upper left corner. Links denote significant structure similarity between clusters ($S>4.0$), and they are coloured red if the two joined clusters contain domains in the same superfamily according to both SCOP and CATH, green if they are in the same superfamily only according to CATH, blue if they are in the same fold according to either SCOP or CATH, and black if there is no pair in the same fold. The figure supports the view that the structurally consistent clusters joined in a network give a richer evolutionary information than a unique and structrally diverse fold.
Found at: doi:10.1371/journal.pcbi.1000331.s004 (0.02 MB PDF)

Acknowledgments

This work is dedicated to the memory of Ángel Ramírez Ortiz, who proposed and directed its starting phase. We deeply mourn his loss as an advisor and a friend. We thank Ian Sillitoe and Christine Orengo for providing the data necessary to compare our classification with the CATH database. Useful discussions with Florian Teichert and Markus Porto are gratefully acknowledged. We also thank Enrique García de Bustos, who participated in a previous phase of this project.

Author Contributions

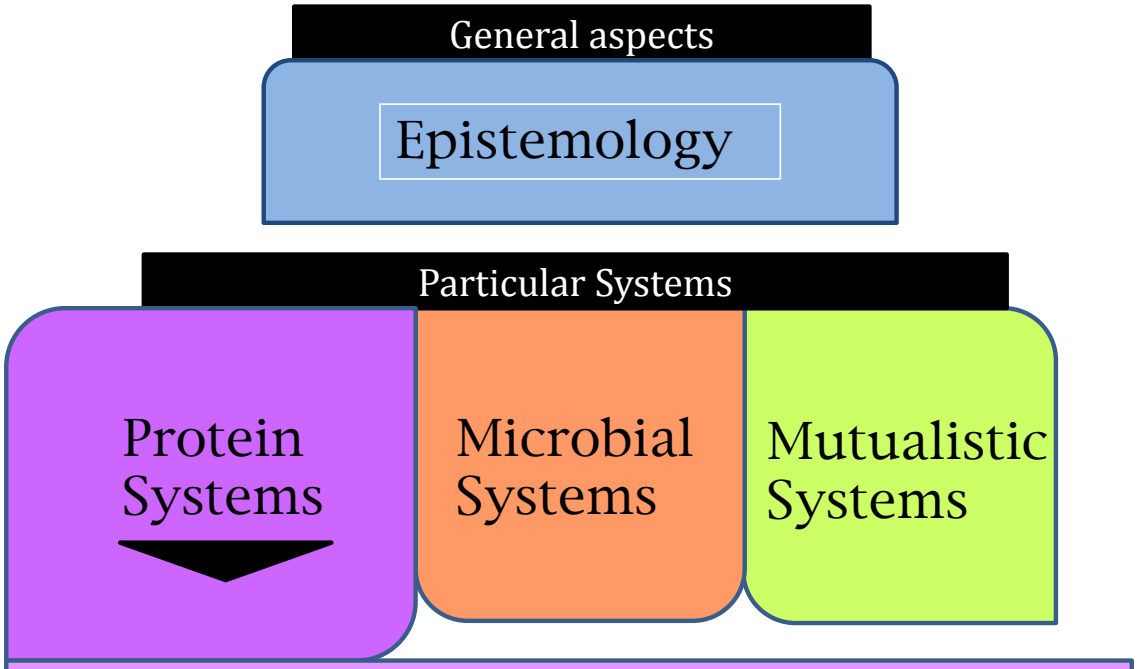
Conceived and designed the experiments: ÁRO UB. Performed the experiments: APG. Analyzed the data: APG UB. Contributed reagents/materials/analysis tools: DA. Wrote the paper: UB.

References

1. Burley SK, Bonanno JB (2002) Structuring the universe of proteins Annu Rev of Genomics. Hum Genet 3: 243–262.
2. Goldsmith-Fischman S, Honig B (2003) Structural genomics: computational methods for structure analysis. Protein Sci 12: 1813–1821.
3. Honig B (2007) Protein structure space is much more than the sum of its folds. Nat Struct Mol Biol 14: 458.
4. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247: 536–540.
5. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, et al. (1997) CATH-a hierarchic classification of protein domain structures. Structure 5: 1093–1108.
6. Brenner SE, Cothia C, Hubbard TJ (1997) Population statistics of protein structures: lessons from structural classifications. Curr Opin Struct Biol 7: 369–376.
7. Swindells MB, Orengo CA, Jones DT, Hutchinson EG, Thornton JM (1998) Contemporary approaches to protein structure classification. Bioessays 20: 884–891.
8. Holm L, Sander C (1998) Touring protein fold space with Dali/FSSP. Nucleic Acids Res 26: 316–319.
9. Holm L, Sander C (1993) Protein structure comparison by alignment of distance matrices. J Mol Biol 233: 123–138.
10. Hadley C, Jones D (1999) A systematic comparison of protein structure classifications: SCOP, CATH and FSSP. Structure 7: 1099–1112.
11. Getz G, Vendruscolo M, Sachs D, Domany E (2002) Automated assignment of SCOP and CATH protein structure classifications from FSSP. Proteins 46: 405–415.
12. Day R, Beck DAC, Armen RS, Daggett V (2003) A consensus view of fold space: combining SCOP, CATH, and the Dali Domain Dictionary. Protein Sci 12: 2150–2160.
13. Sam V, Tai C-H, Garnier J, Gibrat J-F, Lee B, et al. (2006) ROC and confusion analysis of structure comparison methods identify the main causes of divergence from manual protein classification. BMC Bioinformatics 7: 206.
14. Harrison A, Pearl F, Mott R, Thornton J, Orengo C (2002) Quantifying the similarity within fold space. J Mol Biol 323: 909–926.
15. Friedberg I, Godzik A (2005) Connecting the protein structure universe by using sparse recurring fragments. Structure 13: 1213–1224.
16. Rogen P, Fain B (2003) Automatic classification of protein structure by using Gauss integrals. Proc Nat Acad Sci U S A 100: 119–124.
17. Sam V, Tai C-H, Garnier J, Gibrat J-F, Lee B, et al. (2008) Towards an automatic classification of protein structural domains based on structural similarity. BMC Bioinformatics 9: 74.
18. Zemla A, Geisbrecht B, Smith J, Lam M, Kirkpatrick B, et al. (2007) STRALCP—structure alignment-based clustering of proteins. Nucleic Acids Res 35: e150.
19. Suhrer SJ, Wiederstein M, Sippl MJ (2007) QSCOP–SCOP quantified by structural relationships. Bioinformatics 23: 513–514.
20. Chothia C (1992) Proteins. One thousand families for the molecular biologist. Nature 357: 543–544.
21. Kolodny R, Petrey D, Honig B (2006) Protein structure comparison: implications for the nature of 'fold space', and structure and function prediction. Curr Opin Struct Biol 16: 393–398.
22. Shindyalov IN, Bourne PE (2000) An alternative view of protein fold space. Proteins 38: 247–260.
23. Tsai CJ, Maizel JV Jr, Nussinov R (2000) Anatomy of protein structures: visualizing how a one-dimensional protein chain folds into a three dimensional shape. Proc Natl Acad Sci U S A 97: 12038–12043.
24. Berezowski IN, Trifunov EN (2001) Loop fold nature of globular proteins. Protein Eng 14: 403–407.
25. Tendulkar AV, Joshi AA, Sohoni MA, Wangikar PP (2004) Clustering of protein structural fragments reveals modular building block approach of nature J Mol Biol 338: 611–629.
26. Szustakowski JD, Kasif S, Weng Z (2005) Less is more: towards an optimal universal description of protein folds. Bioinformatics 21: ii66–ii71.
27. Ohno S (1970) Evolution by Gene Duplication. New York: Springer-Verlag.
28. Lupas AN, Ponting CP, Russel RB (2001) On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? J Struct Biol 134: 191–203.

29. Efimov AV (1997) Structural trees for protein superfamilies. Proteins 28: 241–260.
30. Taylor WR (2002) A 'periodic table' for protein structure. Nature 416: 657–660.
31. Chothia C, Lesk AM (1986) The relation between the divergence of sequence and structure in proteins. EMBO J 5: 823–826.
32. Grishin NV (2001) Fold change in evolution of protein structures. J Struct Biol 134: 167–185.
33. Krishna SS, Grishin NV (2005) Structural drift: a possible path to protein fold change. Bioinformatics 21: 1308–1310.
34. Newlove T, Konieczka JH, Cordes MH (2003) Retroevolution of λ Cro toward a stable monomer. Proc Natl Acad Sci U S A 100: 2345–2350.
35. Roessler CG, Hall BM, Anderson WJ, Ingram WM, Roberts SA, et al. (2008) Transitive homology-guided structural studies lead to discovery of Cro proteins with 40% sequence identity but different folds. Proc Natl Acad Sci U S A 105: 2343–2348.
36. Viksna J, Gilbert D (2007) Assessment of the probabilities for evolutionary structural changes in protein folds. Bioinformatics 23: 832–841.
37. Rammal R, Toulouse G, Virasoro MA (1986) Ultrametricity for physicists. Rev Mod Phys 58: 765–788.
38. Toulouse G (1977) Theory of the frustration effect in spin glasses: I. Commun Phys 2: 115–119.
39. Lupyan D, Leo-Macias A, Ortiz AR (2005) A new progressive-iterative algoithm for multiple structure alignment. Bioinformatics 21: 3255–3263.
40. Teichert F, Bastolla U, Porto M (2007) SABERTOOTH: protein structural alignment based on a vectorial structure representation. BMC Bioinformatics 8: 425.
41. Ortiz AR, Strauss C, Olmea O (2002) MAMMOTH (Matching Molecular Models Obtained from Theory): an automated method for model comparison. Protein Sci 11: 2606–2621.
42. Zhang Y, Skolnick J (2004) Scoring function for automated assessment of protein structure template quality. Proteins 57: 702–710.
43. Gerstein M (1997) A structural census of genomes: comparing bacterial, eukaryotic, and archaeal genomes in terms of protein structure. J Mol Biol 274: 562–576.
44. Huynen MA, van Nimwegen E (1998) The frequency distribution of gene family sizes in complete genomes. Mol Biol Evol 15: 583–589.
45. Dokholyan NV, Shakhnovich B, Shakhnovich EI (2002) Expanding protein universe and its origin from the biological Big Bang. Proc Natl Acad Sci U S A 99: 14132–14136.
46. Lo Conte L, Brenner SE, Hubbard TJ, Chothia C, Murzin AG (2002) SCOP database in 2002: refinements accommodate structural genomics. Nucleic Acids Res 30: 264–267.
47. Islam SA, Luo J, Sternberg MJ (1995) Identification and analysis of domains in proteins. Protein Eng 8: 513–525.
48. Veretnik S, Bourne PE, Alexandrov NN, Shindyalov IN (2004) Toward consistent assignment of structural domains in proteins. J Mol Biol 339: 647–678.
49. Krishnadev O, Brinda KV, Vishveshwara S (2005) A graph spectral analysis of the structural similarity of protein chains. Proteins 61: 152–163.
50. Xie L, Bourne PE (2008) Detecting evolutionary relationships across existing fold space, using sequence order-independent profile-profile alignments. Proc Natl Acad Sci U S A 105: 5441–5446.
51. Shatsky M, Nussinov R, Wolfson HJ (2002) Flexible protein alignment and hinge detection. Proteins 48: 242–256.
52. Ye Y, Godzik A (2005) Multiple flexible structure alignment using partial order graphs. Bioinformatics 21: 2362–2369.
53. Csaba G, Birzele F, Zimmer R (2008) Protein structure alignment considering phenotypic plasticity. Bioinformatics 24: i98–i104.
54. Mosca R, Brannetti B, Schneider TR (2008) Alignment of protein structures in the presence of domain motions. BMC Bioinformatics 9: 352.
55. Reeves GA, Dallman TJ, Redfern OC, Akpor A, Orengo CA (2006) Structural diversity of domain superfamilies in the CATH database. J Mol Biol 360: 725–741.
56. Jiang H, Blouin C (2007) Insertions and the emergence of novel protein structure: a structure-based phylogenetic study of insertions. BMC Bioinformatics 8: 444.
57. Holm L, Sander C (1998) Dictionary of recurrent domains in protein structures. Proteins 33: 88–96.
58. Nagano N, Orengo CA, Thornton JM (2002) One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. J Mol Biol 321: 741–765.
59. Deeds EJ, Shakhnovich B, Shakhnovich EI (2004) Proteomic traces of speciation. J Mol Biol 336: 695–706.
60. Goldstein RA (2008) The structure of protein evolution and the evolution of protein structure. Curr Opin Struct Biol 18: 170–177.
61. Doolittle RF (1999) Phylogenetic classification and the universal tree. Science 284: 2124–2129.
62. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389–3402.
63. Sokal RR, Michener CD (1958) A statistical method for evaluating systematic relationships. Univ Kans Sci Bull 38: 1409–1438.
64. Cohen J (1968) Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 70: 213–220.
65. Batagelj V, Mrvar A. Pajek: A Program for Large Network Analysis. http://vlado.fmf.uni-lj.si/pub/networks/pajek/.

2.2. Article [PROT-2]



1) Pascual-García A, Abia D, Ortiz AR, Bastolla U. (2009) Cross-over between discrete and continuous protein structure space: insights into automatic classification and networks of protein structures. *PLoS Comput Biol.* 5(3):e1000331.

2) Pascual-García A., Abia D., Méndez R., Nido G.S., Bastolla U. (2010) Quantifying the evolutionary divergence of protein structures: The role of function change and function conservation. *Proteins*. 78(1):181-196

3) Nido G.S., Méndez R., Pascual-García A ., Abia D. and Bastolla U. (2012) Protein disorder in the centrosome correlates with complexity in cell types number. *Mol. BioSyst.* 8, 353-367

4) Sánchez-Nido, G., Romano, L., Bastolla, U. and Pascual-García , *A .* (2015) Structural Bioinformatics within an Snake Puzzle. *In preparation*

5) Pascual-García, A. (2014) Evolución de estructura de proteínas.
6) Pascual-García, A. (2014) Modelos simplificados de plegamiento de proteínas.
7) Pascual-García, A. (2014) Alineamiento de estructura de proteínas.
Published in: "Bioinformática con Ñ".
Editors: Sebastián A., and Pascual-García, A.
ISBN: 978-84-617-1976-X.

PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Quantifying the evolutionary divergence of protein structures: The role of function change and function conservation

Alberto Pascual-García, David Abia, Raúl Méndez, Gonzalo S. Nido, and Ugo Bastolla*

Centro de Biología Molecular 'Severo Ochoa' (CSIC-UAM), Cantoblanco, Madrid 28049, Spain

ABSTRACT

The molecular clock hypothesis, stating that protein sequences diverge in evolution by accumulating amino acid substitutions at an almost constant rate, played a major role in the development of molecular evolution and boosted quantitative theories of evolutionary change. These studies were extended to protein structures by the seminal paper by Chothia and Lesk, which established the approximate proportionality between structure and sequence divergence. Here we analyse how function influences the relationship between sequence and structure divergence, studying four large superfamilies of evolutionarily related proteins: globins, aldolases, P-loop and NADP-binding. We introduce the contact divergence, which is more consistent with sequence divergence than previously used structure divergence measures. Our main findings are: (1) Small structure and sequence divergences are proportional, consistent with the molecular clock. Approximate validity of the clock is also supported by the analysis of the clustering coefficient of structure similarity networks. (2) Functional constraints strongly limit the structure divergence of proteins performing the same function and may allow to identify incomplete or wrong functional annotations. (3) The rate of structure versus sequence divergence is larger for proteins performing different functions than for proteins performing the same function. We conjecture that this acceleration is due to positive selection for new functions. Accelerations in structure divergence are also suggested by the analysis of the clustering coefficient. (4) For low sequence identity, structural diversity explodes. We conjecture that this explosion is related to functional diversification. (5) Large indels are almost always associated with function changes.

Proteins 2009; 00:000–000.

Key words: protein structure evoluation; molecular clock; protein function; protein structure classification.

INTRODUCTION

The molecular clock hypothesis[1] played a fundamental role in the early days of molecular evolution studies after Zuckerkandl and Pauling recognized that protein sequences accumulate amino acid substitutions almost linearly in time, with a rate that varies with the protein family but is almost constant in different lineages.[2] The neutral theory, proposed almost simultaneously by Kimura[3] and King and Jukes,[4] interprets the constancy of the evolutionary rates as the result of neutral substitutions,[5,6] i.e., substitutions that have very little effect on fitness and are fixed in natural populations through random genetic drift instead of positive selection. This theory was subsequently generalized by Ohta to include nearly neutral substitutions for which either the selective effect or the effective population size is small.[7] The nearly neutral theory can be derived from standard population genetics models[8] and it is formally equivalent to equilibrium statistical mechanics, since molecular properties arise from a balance between mutational entropy in sequence space and fitness, where population size plays the role of inverse temperature.[9] In particular, protein folding stabilities in bacterial genomes are predicted to be smaller for bacteria with low effective population size.[10]

Though controversial in a first time,[11] the neutral and nearly neutral hypothesis had the great merit to give theoretical support to the molecular clock hypothesis, which is still of fundamental importance for methods that reconstruct evolutionary trees from molecular data.[12] Moreover, the neutral hypothesis also predicts that we should find violations of the molecular clock in the interesting cases when adaptive evolution takes place, for instance when new molecular functions emerge.

The quantitative study of the rate of protein structure evolution has received comparatively less attention. A milestone was the 1980 paper by Chothia and Lesk, who showed that the Root Mean Square Deviation (RMSD) between different globins diverges regularly with the number of amino acid substitutions,

Additional Supporting Information may be found in the online version of this article.
Grant sponsors: Spanish Ministery of Science and Innovation (Ramón y Cajal fellowship), Grant numbers: BIO2008-04384, CSD200623.
*Correspondence to: Dr. Ugo Bastolla, Centro de Biologia Molecular 'Severo Ochoa', (CSIC-UAM), Cantoblanco, Madrid 28049, Spain. E-mail: ubastolla@cbm.uam.es.
Received 8 April 2009; Revised 9 September 2009; Accepted 10 September 2009
Published online 22 September 2009 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/prot.22616

up to a limit of low sequence identity where the RMSD explodes.[13] Although this result suggests a generalization of the molecular clock hypothesis to the evolution of protein structure, it is limited by the fact that the RMSD can be used as a measure of structure divergence only for aligned residues that have a good spatial superimposition. We will propose here a measure of structure divergence based on evolutionary considerations, which is more suitable for such a quantification.

Together with the clock-like divergence of protein structures, the results by Chothia and Lesk also suggested that protein evolution conserves the fold, an equivalence class of protein structures defined as a spatial arrangement with "the same major number and direction of secondary structures with a same connectivity".[14] This view strongly influenced the classification of protein structures in databases such as SCOP[14] and CATH,[15] where proteins recognized as evolutionarily related (i.e., homologous) are classified in the same structural fold. However, the accumulation of protein structure data has revealed that "fold change" is relatively frequent in the evolution of proteins[16–18] and that folds or topologies as defined in SCOP and CATH fail to pass tests of consistency with respect to structure similarity measures.[19]

Protein classification and molecular clocks are intimately related. The very possibility to objectively classify protein structures requires that the structure similarity measure is transitive, i.e., similarity between *a* and *b* and between *b* and *c* must imply similarity between *a* and *c*. This property is guaranteed by the phylogenetic trees underlying the gene duplication process. Therefore, if protein sequences or structures diverge regularly in evolution (the molecular clock hypothesis), their divergence can be used for objective and consistent classification. However, if divergence is accelerated for instance through positive selection, function diversification or large insertions and deletions (which are not mutually exclusive processes), we expect that the transitive property is violated and consistent classification is not possible. We will test here the molecular clock through a quantitative study of structural and functional divergence in the evolution of four large superfamilies: Globins, Aldolases, P-loop containing nucleotide triphosphate hydrolases, and NADP-binding Rossmann-like domains.

The relationship between protein function on one hand, and sequence and structure on the other hand, has been subject to intense investigation. For instance, Devos and Valencia[20] and Wilson *et al.*[21] independently concluded that protein function, assessed through the Enzyme Commission (EC) classification, is generally conserved above 40 percent sequence identity. Using the CATH classification of proteins, Todd, Orengo and Thornton[22] found that function divergence is common in homologous superfamilies, although the extent of this divergence varies from one superfamily to the other. Lecomte *et al.*[23] studied the divergence of protein sequences, structures and functions in the globins superfamily, and Sangar *et al.*[24] found that, for proteins with more than 50% sequence identity, function assigned through homology is correct in 94% of the cases. It has been found through these studies that structure similarity at the fold level is compatible with a multiplicity of functions. It has been proposed that these multiple functions originated from divergent evolution followed by structure and function diversification,[25] a view that we adopt in this analysis, examining proteins in the same superfamily that are believed to share a common ancestor. This multiplicity of functions makes function prediction from sequence and structure a difficult problem, because homologous proteins often have different functions.[22,26] And yet it is a more and more urgent problem, due to the accumulation of huge sequence data waiting for annotation.[27] Despite the ambiguity of the structure-function relationship, it has been found that structural information provides added value for function prediction with respect to plain sequence information.[28,29] We can shed light on the structure-function uncertainty[30] using evolutionary information, since phylogenetic, structural and functional distance are correlated.[31] These considerations motivated us to undertake a study of how function change and function conservation influence the evolutionary divergence of protein structures.

RESULTS

Contact divergence: a new measure of structure divergence

In their seminal paper, Chothia and Lesk quantified protein structure divergence through the RMSD. However, this measure can be computed only for aligned residues that are well superimposed in space. In practice, it is necessary to fix a cut-off distance that specifies which residues are well superimposed, and the RMSD increases with the cut-off. A more robust measure of structure similarity is the number of superimposed residues within this cut-off, called percentage of structure identity (PSI). This and other measures of structure similarity have to be normalized in such a way that the comparison between two unrelated proteins is not trivially correlated with their size. To achieve this normalization, one typically uses the mean and standard deviation of the similarity of unrelated proteins of similar length, assuming either Gaussian statistics (the *Z* score) as in the Dali program,[32] or extreme value statistics, as in the significance score of the program Mammoth.[33] However, this normalization has the drawback that the similarity of related proteins becomes strongly dependent on their length. For instance, the *Z* score of 100 percent PSI increases as a power law of protein length. Therefore, this significance can not measure the evolutionary divergence. A possible solution to this problem is a new type of normalization,

such as the TM-score proposed by Zhang and Skolnick.[34] Proteins with 100 percent structure identity have TM score equal to one and unrelated proteins have TM score that uncorrelated with their length.

We have recently observed that the contact overlap (see Materials and Methods) performs better than the number of superimposed residues for the sake of classifying protein structures based on their similarity.[19] There are two reasons for this: (1) The contact overlap weights more the aligned residues in the core of the protein, where the number of contacts is large or, equivalently, it penalizes less the non-superimposed residues with few contacts, such as those in loops; (2) Relative motions of two subdomains, such as hinge motions, are much less penalized by the contact overlap since intra-subdomain contacts are conserved in the two conformations. Therefore, we look here for a way to normalize the contact overlap making it independent of protein length both for related and unrelated proteins.

To this aim, we will use the analogy with an evolutionarily motivated measure of protein sequence divergence. Consider two proteins related by gene duplication that diverged during t years. We assume that the probability that no substitution happens in the time t decays exponentially with rate $1/\tau$ as $exp(-t/\tau)$. The conditional probability that two amino acids are equal given that at least one change happened at their common position is $p = \sum_a f(a)^2$, where $f(a)$ is the frequency of amino acid a. Using the frequencies $f(a)$ measured by Jones *et al.*[35] on the SwissProt database, we get $p = 0.058$. Therefore, the probability to observe the same amino acid at an aligned position i in two proteins that diverged for a time t is

$$P\{A_i^1 = A_i^2\} = \mathrm{e}^{-t/\tau} + p(1 - e^{-t/\tau}). \quad (1)$$

We can estimate the probability that two amino acids are equal as the sequence identity between the two proteins, SI. This estimate is only rigorous if the substitution process is independent at each protein position, which is clearly not true, but this is an almost unavoidable assumption. Using $P\{A_i^1 = A_i^2\} \approx \mathrm{SI}$, we can solve Eq. (1) finding the evolutionary divergence time t as

$$t/\tau = -\log\left(\frac{\mathrm{SI} - p}{1 - p}\right). \quad (2)$$

(from here on, log indicates the Neperian logarithm). When $\mathrm{SI} \gg p$, this formula coincides with the standard Poisson formula used to estimate evolutionary distances.[36] Equation (2) is also in fair agreement with simulations of protein sequence evolution subject to the global constraint of folding stability,[37] provided that SI is not close to p, in which limit the evolutionary information is wiped out. The formula is not defined if $\mathrm{SI} \leq p$.

We generalize Eq. (2) to the evolutionary divergence of the inter-residue contacts in a protein structure. Given two proteins with contact overlap q (see Materials and Methods), we define their contact divergence as

$$D_{\mathrm{cont}}(q, L) = \begin{cases} -\log\left(\frac{q - q_\infty(L)}{1 - q_\infty(L)}\right) & \text{if } q > \epsilon(L) \\ D_0 - (q - \overline{q}(L))/\sigma_q(L) & \text{otherwise} \end{cases} \quad (3)$$

The upper line of the aforementioned equation defines the contact divergence of related proteins, in analogy to how sequence identity is tranformed to estimate evolutionary divergence in Eq. (2), so that $D_{\mathrm{cont}} = 0$ for proteins having identical contact matrices and $D_{\mathrm{cont}} \to \infty$ for $q \to q_\infty(L)$. Therefore, the parameter $q_\infty(L)$, which is the analogous of p for protein structures, represents the asymptotic limit of the contact overlap after a very long evolutionary time. For $q \leq q_\infty(L)$ the logarithm in the upper line is not defined, and we define in the lower line the contact divergence of unrelated and distantly related proteins. The cross-over takes place at $q \leq \epsilon(L) > q_\infty(L)$, and after this point contact divergence is given by a linear function of the Z score of the overlap, $Z = (q - \overline{q})/\sigma_q$. We have tested in previous work that the Z score is a convenient similarity measure for unrelated proteins. As for other structure similarity measures, the mean and standard deviation of the overlap of unrelated proteins, $\overline{q}(L)$ and $\sigma_q(L)$, depend on protein length. To simplify this dependence, we parameterize the size of the protein pair as the geometric mean of the length of the two proteins, $L = \sqrt{L_1 L_2}$, and we measure $\overline{q}(L)$ and $\sigma_q(L)$ for unrelated protein pairs of length L in the representative set of structural domains ASTRAL40 (see Materials and Methods).

The formula (3) depends on the parameters $q_\infty(L)$ (asymptotic overlap), $\epsilon(L)$ (threshold overlap) and D_0. To reduce the number of free parameters, we make the following assumptions. First, we assume that the asymptotic overlap $q_\infty(L)$ is a linear function of the mean and standard deviation of the overlap of unrelated proteins:

$$q_\infty(L) = \overline{q}(L) + A\sigma_q(L). \quad (4)$$

Since $\overline{q}(L)$ and $\sigma_q(L)$ depend on length, so does $q_\infty(L)$ as well. A is a free parameter whose positive value means that the asymptotic overlap of homologous proteins separated by a very long evolutionary distance is larger than the mean overlap of unrelated proteins, i.e., the memory of the relatedness is never lost. Second, we fix the parameter $\epsilon(L)$ by imposing continuity of Eq. (3) at $q = \epsilon(L)$

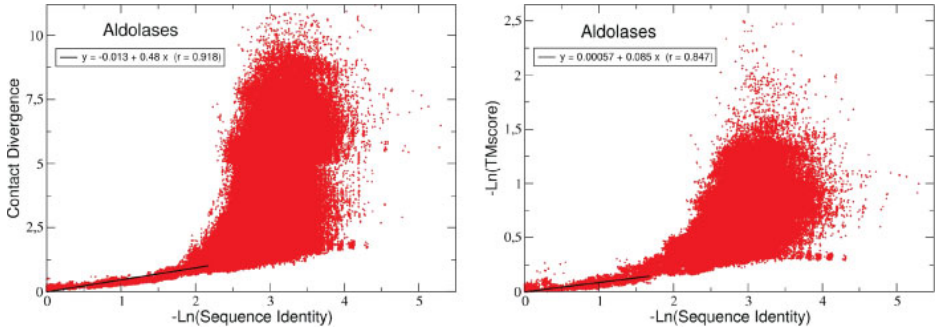


Figure 1

Structure divergence versus sequence divergence for proteins in the aldolase superfamily. Left plot: Contact divergence. Right plot: natural logarithm of the TM score. The linear fits are restricted to the largest region in which the intercept of the fit does not differ significantly from zero. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

(see Materials and Methods). This continuity condition can be imposed only if the parameter D_0, which is independent of length, is large enough. We therefore decided to take the smallest value of D_0 such that the continuity condition is met for all protein pairs in our representative data set of single domain proteins (see Materials and Methods). With these choices, the only free parameter in the definition of the contact divergence is the parameter A in Eq. (4). We chose A by testing the consistency between the new measure and evolutionarily grounded classifications of protein structures. We clustered 2890 nonredundant protein structures with less than 40 percent pairwise sequence identity using the average linkage algorithm applied to different similarity scores, and compared the corresponding classifications with the SCOP and CATH classifications at superfamily level using the weighted kappa measure.[38] This level was chosen because superfamily relationships reflect common evolutionary origin, and because SCOP and CATH agree with each other at the superfamily level much more than at the fold level.[19] We also compared our structural clusters to the ones obtained using as similarity the sequence identity after optimal structure alignment (see Supporting Information Fig. 1). Notice that, since protein structure is used for the alignment, this measure is much more reliable than the sequence identity obtained through sequence alignment. At large identity, corresponding to the initial steps of the clustering algorithm, sequence identity is believed to yield reliable phylogenetic trees. Therefore, this comparison tests the ability of the structural score to yield trees that are consistent with the process of evolutionary divergence for closely related proteins, whereas the superfamily comparison addresses farther evolutionary relationships. For each comparison, we selected the maximum weighted kappa for all threshold structure similarities.

The results of these tests (see Table I) show that the contact divergence score outperforms both the Z score of the contact overlap and the TM score regarding its consistency with evolutionary based classifications, such as SCOP superfamilies, CATH superfamilies, and sequence identity based trees. All three evolutionary classifications give very similar rankings, despite the sequence identity measure has a low agreement with the superfamily classifications. This is not surprising, since most pairs have sequence identities below 25 percent (the so-called twilight zone) that would not be significant in the absence of structure information, which is used for superfamily assignment in both CATH and SCOP. We found the worst agreement with sequence identity using the Z score of the overlap. The latter measure reduces as much as possible the length dependence for unrelated protein pairs but it is strongly length dependent for closely

Table I

Consistency Between the Clusters Obtained Through Different Similarity Measures and Evolutionary Based Classifications

Score	Parameter	WK SCOP S.F.	WK CATH S.F.	WK Seq. Id.
Seq. Id.	—	0.48	0.48	—
Z-Score	—	0.63	0.61	0.562
TM-Score	—	0.59	0.58	0.720
Cont. Divergence	$A = 0$	0.56	0.58	0.723
Cont. Divergence	$A = 2$	0.58	0.58	0.745
Cont. Divergence	$A = 3$	0.62	0.60	0.749
Cont. Divergence	$A = 4$	0.64	0.62	0.753
Cont. Divergence	$A = 5$	0.66	0.64	0.754
Cont. Divergence	$A = 6$	0.64	0.62	0.750
Cont. Divergence	$A = 8$	0.63	0.61	0.692

As a test set, we used a consensus set of 2890 nonredundant domains classified in 779 SCOP superfamilies and 885 CATH superfamilies. Consistency was assessed through the maximum weighted kappa measure[38] obtained for all threshold similarities. We did not perform computations for $A = 1$ since, interpolating results with $A = 0$ and $A = 2$, it is clear that this value is suboptimal. The same holds for $A = 7$.

related proteins. For instance for $q = 1$, corresponding to $D_{\mathrm{cont}} = 0$, we have $Z = (1 - \bar{q}(L))/\sigma_q(L)$. The worst agreement with the superfamily classifications was found for the TM score, confirming that scores based on the number of superimposed residues perform worse than scores based on contacts for detecting distant evolutionary relationships. The best consistency with all evolutionary classifications was found for the contact divergence measure with $A = 5$. In the following, when we mention contact divergence we will mean this choice of parameters.

Molecular clock for structure divergence

We now analyse four large superfamilies, each containing more than thousand crystallized structures: Globins, Aldolases, P-loop containing nucleoside triphosphate hydrolases and NADP-binding Rossman-fold. The list of domains and their definition were taken from the CATH database.[15] The list of the corresponding SCOP domains is very similar, but their definition is somewhat different, since SCOP domains are typically larger than CATH. We eliminated NMR structures, chains with more than one domain, for which function assignment is problematic, and redundant domains almost identical both in sequence and in structure. Identical sequences with slightly diverged structures were retained in order to have a glimpse at conformation changes. For each pair of domains in the same superfamily we measured pairwise dissimilarities in structure, sequence, function and length (see Materials and Methods). In particular, structure divergence was measured through the contact divergence score D_{cont} defined earlier, sequence divergence was measured as $-\log(\mathrm{SI})$, where SI is the sequence identity obtained through structure alignment, and function similarity was defined to be one if all GO terms[39] of the two proteins coincide, zero otherwise. For globins we also used InterPro signatures[40] to complement GO terms.

First, we examined the relationship between sequence and structure divergence. One can see from Figure 1 that structure divergence increases almost linearly with sequence divergence when this is not too large. If the sequence diverges in a clock-like manner, this result is consistent with the extended molecular clock hypothesis that structure divergence accumulates linearly with time. Figure 1 represents the Aldolase superfamily. In the left plot we measure structure divergence through the contact divergence measure. We linearly fitted contact divergence versus sequence divergence up to the point where the intercept of the fit differs significantly from zero (i.e., where the intercept becomes larger than its standard error). This point corresponds to SI = 0.115, and the correlation coefficient of the fit is $r = 0.918$. We repeated the same procedure using the TM score, measuring TM score divergence as $-\log(\text{TM score})$. Also in this case the molecular clock hypothesis holds, but its range of validity is narrower (it is SI $\geq$ 0.187) and the correlation coefficient is smaller, $r = 0.847$. If we assume that the sequence divergence $-\log(\mathrm{SI})$ evolves approximately clockwise, the fact that contact divergence is approximately linearly related to sequence divergence over a broader range suggests that this measure evolves more clockwise than the TM score and it is more convenient for quantifying the evolutionary divergence of protein structures.

The other superfamilies yielded similar results, except for the Globin superfamily for which several proteins with conformation changes and unchanged sequences are present. In this case, the intercept of the linear fit is significantly different from zero also for very small divergence, and we could not apply the aforementioned method to determine the range of validity of the molecular clock. These results confirm that contact divergence is a convenient measure for quantifying protein structure change in evolution.

Structure diversity explosion

For small sequence identity (large divergence), the approximate proportionality between structure divergence and sequence divergence disappears and one can see an explosion of structure diversity. One possible explanation to this spectacular explosion, observed for all structure divergence measures and in all four superfamilies with very similar characteristics, is the attenuation of functional constraints, since almost all of the strongly diverged pairs have different function (see below). Strongly diverged pairs also tend to have large insertions and deletions, which may be responsible for the increased structure divergence. As we will see in the following, our analysis supports both interpretations. However, an even simpler interpretation is also possible.

As expressed in Eq. 1, after a very long divergence time multiple substitutions have occurred at most sites, and the sequence identity of two homologous proteins reaches an asymptotic distribution where aligned residues may become identical by chance rather than by common origin and all evolutionary information is lost. This situation can be studied by simulating protein sequence divergence through random mutations that are fixed if they do not appreciably modify the stability of the target protein structure, assessed through an effective free energy function.[37] Using these simulations, the mean number of attempted mutations, which is related with the evolutionary divergence time, may be represented versus $-\log(\mathrm{SI})$ as in Figure 2. We can see from this plot that the sequence divergence $-\log(\mathrm{SI})$ is a reliable estimate of the divergence time only for large enough sequence identity (small divergence), whereas large sequence divergences tend to strongly underestimate the divergence time. After a very long time all evolutionary information is lost, and sequence identity reaches an asymptoyic distribution centered around the small value

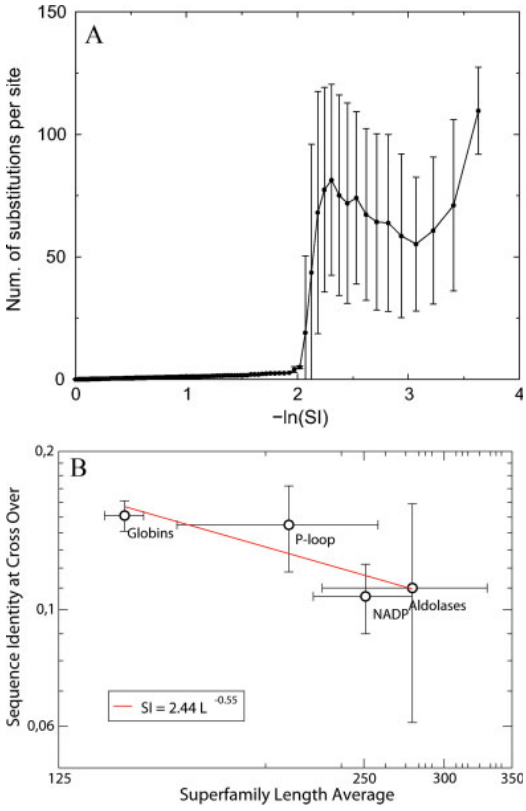


Figure 2

(**A**) Results from a simulation of protein sequence evolution with conservation of the folding stability of the target structure. The mean number of substitutions, measuring the divergence time, is plotted versus the sequence divergence −log(SI). Data modified from.[37] (**B**) For the four superfamilies studied, we plot the sequence similarity at which the structural explosion occurs versus the average length of the superfamily. The error bars indicate the uncertainty on the cross-over point and the standard deviation of protein length. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

SI $= p \approx 0.058$. The largest sequence identity found with non negligible probability in this asymptotic ensemble determine the cross-over, since smaller identities do not allow to estimate the divergence time. Both probabilistic arguments and simulations[37] suggest that the sequence identity at the cross-over decreases with protein length L approximately as $L^{-1/2}$. For the four superfamilies studied in Figure 3, we estimated the sequence identity at the cross-over by plotting the standard deviation of contact divergence versus sequence identity. This quantity makes a jump at the cross-over that allows to identify it with reasonable accuracy (data not shown). For the Aldolases and NADP superfamilies, which do not present important conformation changes, the cross-over estimated in this way is in very good agreement with the limit of validity of the molecular clock estimated in the previous section through the condition that the intercept of the linear fit should be zero. We found that the cross-over identity decreases as $L^{-0.55}$ when the mean length L of the superfamily increases (see Fig. 2, right plot), consistent with the aforementioned interpretation. Therefore, the apparent explosion of structure divergence at the cross-over might be a simple consequence of the fact that sequence identity below the cross-over strongly underestimates the divergence time, coupled with the relaxation of functional constraints on protein structure that will be discussed below.

Functional constraints on protein evolution

We represent in Figure 3 structure divergence versus sequence identity, distinguishing protein pairs that perform the same function (i.e., all their GO terms regarding the Molecular function are equal) from those with different functions. We only consider in this analysis proteins whose GO terms have been manually curated, as indicated by their evidence code. As one can see from this figure, proteins sharing the same function are more conserved in sequence and in particular in structure with respect to pairs with different functions. This result is expected, since protein function is known to constraint sequence and structure. Nevertheless, the strength of these constraints is surprising, since we found very few pairs with different functions having contact divergence larger than 2, and almost all of them can be attributed to conformation changes rather than evolutionary divergence (see below). Structure divergence is very limited even for pairs with sequence identity lower than the cross-over of structural explosion, for which the evolutionary divergence time may be very large. Moreover, as we will see later, several pairs with very large structure divergence have been electronically annotated as having the same GO terms. Both our results and the InterPro annotations suggest that these electronic annotations may be incomplete. Therefore, using the knowledge of the strength of functional constraints on protein structure would have avoided these incomplete annotations.

Conversely, we observed many pairs of proteins with different function and very similar structure, confirming the known fact that even small structure divergence may be sufficient to change protein function. In other words, structure divergence is a very strong indication of functional change, but structure conservation does not always imply function conservation.

Electronic annotations

We now plot in Figure 4 structure divergence versus sequence identity also for proteins that have been electronically annotated, according to the evidence codes in

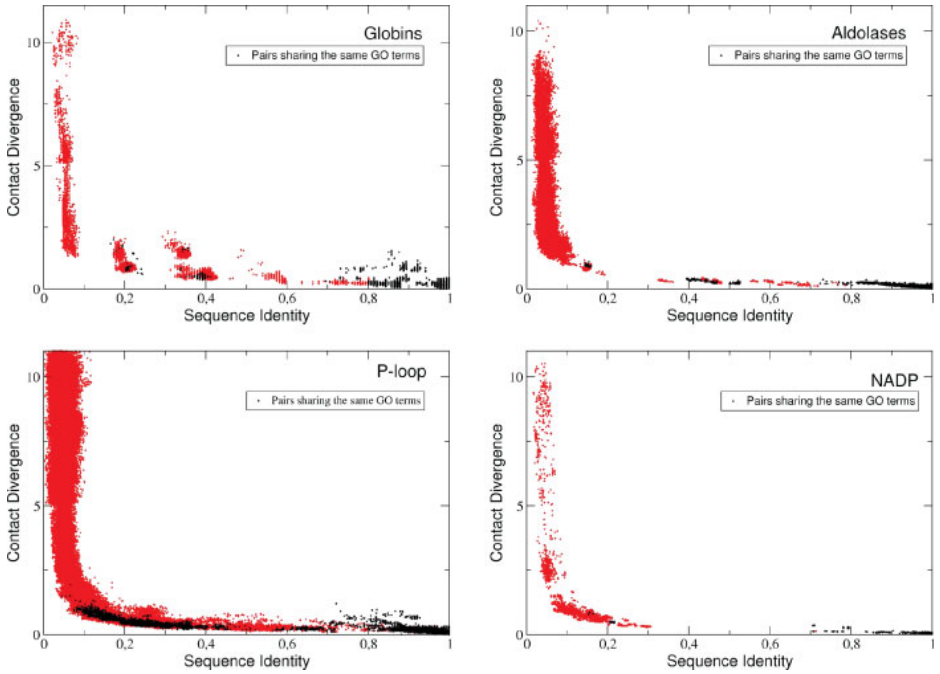


Figure 3

For each of the four superfamilies we plot contact divergence versus sequence identity, distinguishing protein pairs performing the same function according to all of their GO terms (dark points). Only proteins with manually annotated GO terms are represented. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

the GO. For the Aldolases and the NADP superfamilies the plots are very similar to Figure 3, and they are not shown. However, for the P-loop and in particular the Globins superfamilies, we found a very large number of pairs annotated as having the same function but with very large contact divergence. Most globins that are electronically annotated are classified as having the heme binding, oxygen binding, and oxygen transporter function. We then adopted the InterPro classification (see Materials and Methods), which distinguishes different types of Hemoglobin chains (alpha, beta, zeta, pi), and lamprey and annelid globins. Although all of them are involved in oxygen transport, they may have rather different affinity for oxygen and regulation mechanisms[41] and may perform secondary functions,[42] which makes it reductive to classify all of them under the same functional class. Besides, paralog genes are believed to perform different functions in order to be retained in evolution, so that classifying all hemoglobin types under the same function is likely to reduce too much the resolution at which we can look at protein function. We found the surprising results that no protein pair with the same InterPro signature has contact divergence larger than 2 (see Fig. 4), except for a pair involved in a large conformation change, in perfect agreement with the result that we obtained for manually annotated GO terms. This result suggests that proteins with contact divergence larger than 2 with respect to manually annotated proteins with the same function may be incompletely or wrongly annotated. For the P-loop superfamily all such outliers, i.e., the dark points in Figure 4 with contact divergence larger than 2, are explained by only 5 domains (PDB codes 1xjcA, 1gvnB, 1gvnD, 1y63A and 1ghhA), for the NADP superfamily we identified two proteins that may be incompletely annotated (1jax and 1jay), and no one for the Aldolases superfamily, whereas most globins are insufficiently annotated electronically, as discussed earlier.

Global structure conservation and function prediction

As expected, the results presented in the previous section show that large sequence and structure divergence are strong predictors of function change, and sequence and structure conservation are (weaker) predictors of function conservation. To quantitatively assess the performances of

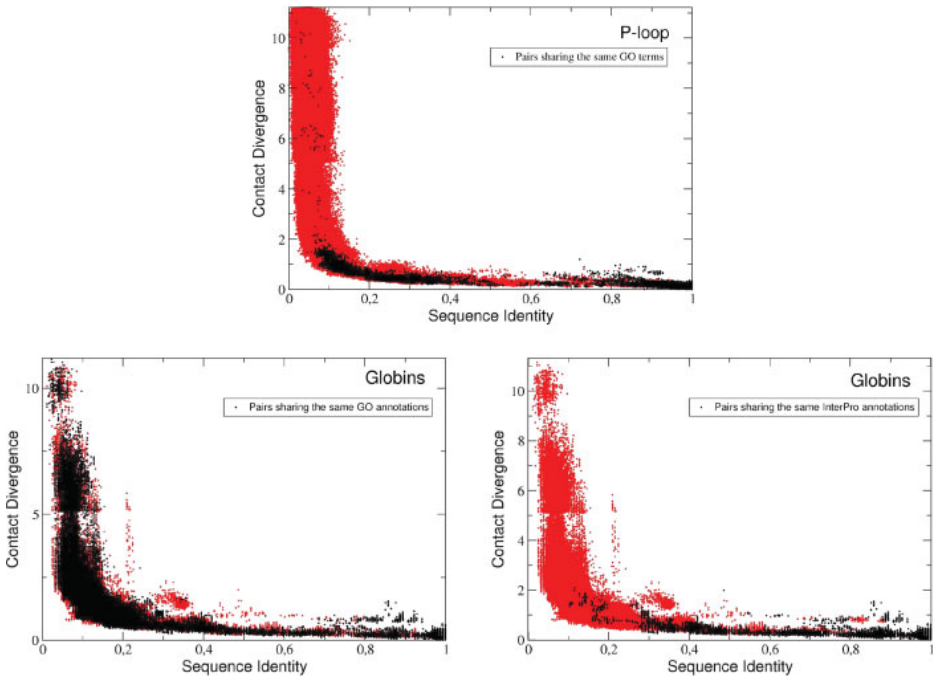


Figure 4

For the Globins and P-loop superfamilies we plot contact divergence versus sequence identity, distinguishing protein pairs sharing the same function according to the GO terms. Also electronically annotated proteins are considered in these plots. For the Globins superfamily we represent on the lower right panel the same plot where we identify proteins with the same function as those with the same InterPro signatures. Notice that we find several pairs with the same electronically annotated function but very large contact divergence. Such pairs are not found in case of manual annotations and InterPro signatures. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

contact divergence and other sequence and structure similarity measures under this respect, we measured the sensitivity and selectivity (see Materials and Methods) for predicting function conservation using different thresholds on structure similarity. The corresponding ROC plots show almost perfect Area Under the Curve (AUC), tabulated in Table II (see Supporting Information Fig. 2). AUC of one means perfect prediction, 0.5 indicates a random prediction. All scores perform very similarly but, surprisingly, the sequence identity score is an even better predictor than actual structure similarity measures. Notice, however, that we measure sequence identity after optimal structure alignment, so that the performances of this measure would not be possible if we did not dispose of structural information.

Structure evolution is accelerated upon function change

To characterize more quantitatively the effect of function on structure divergence, we quantified the relationship between sequence and structure divergence. For sequence identities above the cross-over, we can estimate the divergence time either through sequence divergence as $t \approx -\log(\mathrm{SI})$ or through structure divergence as $t \approx D_{\mathrm{cont}}$. These two estimates are proportional, which means that the molecular clock based on sequence and the one based on structure are consistent.

However, a closer look shows that the two molecular clocks present discrepancies when functional changes occur. Through a linear fit, we computed the slope of D_{cont} versus $-\log(\mathrm{SI})$ before the cross-over, distinguishing protein pairs with the same function (see Table III). One can see that all of these slopes are smaller than one, confirming that protein structure diverges more slowly than sequence, and they are all in a relatively limited range, from 0.25 for P-loops to 0.48 for Aldolases.

For all four families, protein pairs with different functions present significantly larger slopes (in the range 0.29 to 0.48) than those with the same function (from 0.25 to 0.37). Although not unexpected, this is a rather interesting result, since it demonstrates a quantitative

Table II
AUC (Area Under the Curve) of the ROC Plots of Function Conservation Predictions Using Different Structure Similarity Measures

Superfamily	Seq. Id.	Cont.Div.	Z-*Score*	TM-Score
Aldolases	0.988	0.980	0.979	0.988
Globins	0.977	0.984	0.979	0.982
P-loop	0.978	0.973	0.977	0.974
NADP	0.840	0.812	0.809	0.833

influence of protein function on the sequence to structure relationship. Moreover, it suggests possible improvements to protein function prediction. In fact, it is known that very small changes in sequence and structure are sufficient to modify protein function, so that sequence and structure conservation are not a sufficient indication of function conservation. Our observation that function change modifies quantitatively the sequence to structure relationship suggests that this information could be used in order to predict function conservation more reliably.

This influence of function change on the sequence to structure relationship can be interpreted either as due to the fact that function change relaxes the constraints on protein structure (negative selection) or due to positive selection for modified structure to perform the new function. We think that the latter mechanism is more relevant. In fact, we observed this behavior while structure divergence is still within the range $D_{\text{cont}} < 2$ typical for proteins with the same function, so that structural constraints imposed by function conservation would be fulfilled.

Evolutionary rates and clustering

We have seen earlier that, for sequence divergence below the cross-over, sequence divergence and structure divergence are approximately proportional. This implies that the divergence times estimated through sequence divergence and through structure divergence are proportional, so that the molecular clock based on sequence and the one based on structure are consistent. This approximate clockwise evolution of protein structures has an important implication for protein structure classification. In fact, if the molecular clock approximately holds, structure divergence is expected to be able to reconstruct the phylogenetic tree underlying protein evolution, similar to how this is done with sequence divergence. Given a phylogenetic tree, the time distance from the leaves of the tree passing through their closest common ancestor is ultrametric,[43] i.e., if C is the outgroup of the triple A,B,C it holds $t_{AC} = t_{BC} > t_{AB}$. This relationship is valid for all triples, and it guarantees that the transitive property holds for all distance thresholds, i.e., if A and B are related and B and C are related also A and C must be related. Therefore, relatedness along the tree induces an equivalence relationship whose equivalence classes are the phylogenetic groups. If the molecular clock approximately holds, the divergence $D_{AB} \approx kt_{AB}$ can be used to estimate the divergence time and to reconstruct the tree.

Table III
Slope of Contact Divergence Versus Sequence Divergence for Protein Pairs Sharing the Same Function and for all Possible Protein Pairs

Superfamily	Slope (all pairs)	Slope (same function)
Aldolases	0.4830 ± 0.0007	0.3733 ± 0.0006
Globins	0.3912 ± 0.0003	0.3572 ± 0.0007
P-loop	0.2888 ± 0.0008	0.2529 ± 0.0011
NADP	0.3914 ± 0.0005	0.3245 ± 0.0007

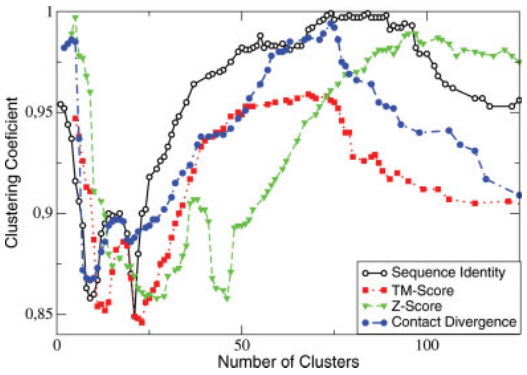


Figure 5
For the NADP superfamily and for each different divergence measures and distance thresholds, we constructed a network by joining all pairs of proteins closer than the threshold. For each network, we plot the clustering coefficient versus the number of connected components, i.e., the number of clusters obtained with single linkage, which decreases with increasing distance threshold. One can see that there is a range of optimal similarity threshold such that the clustering coefficients are close to one, as expected from the molecular clock hypothesis. Also notice that the clustering coefficients present dips that suggests that the evolutionary rate is not constant in this region.

To test the clustering properties of the divergence measures studied here, we measured the clustering coefficient (see Materials and Methods) of the networks constructed by joining together proteins with D_{AB} smaller than some threshold. If the clustering coefficient is one, all related proteins share all their neighbors and transitivity exactly holds. Ultrametricity implies clustering coefficient equal to one for all thresholds. The validity of the molecular clock hypothesis therefore implies that the clustering coefficient is close to one for all thresholds.

Figure 5 shows the clustering coefficient of the networks obtained with a given distance measure and given threshold versus the number of connected components of the same network (i.e., the number of clusters obtained with single linkage clustering). The larger this number, the smaller the distance threshold. There is a range of distance thresholds such that the clustering coefficient is close to one, consistent with the molecular clock hypothesis. The clustering coefficient is larger using

sequence identity measured with the optimal structure alignment than using structure similarity measures. This suggests that protein sequences evolve in a more clock-like fashion than protein structures. Among structure similarity measures, the contact divergence yields the best clustering coefficient for NADP and P-loop whereas the TM score is the most clock-like for globins and aldolases.

Notice that in Figure 5 the clustering coefficients present dips suggesting that the evolutionary rate is not constant for some values of the thresholds, corresponding to some typical evolutionary distance. We conjecture that this phenomenon is related to the rate acceleration when the protein function changes. This interpretation is supported by the fact that we measured a significantly larger rate of structure divergence for proteins with different function. We will analyse this issue in future work.

Conformation changes

Figures 3 and 4 show some outliers, i.e., protein pairs with large sequence identity whose structures diverge much more than expected. These are often examples of conformation changes, i.e., proteins that change conformations while performing their biological activity. We discuss in this section the examples that produce the most severe outliers in Figure 4, i.e., pairs whose structure divergence is much larger than expected based on their sequence identity. The conformation changes discussed below are represented in Supporting Information Figure 4.

Globins

Many proteins in this superfamily have been crystallized with different co-factors (mostly oxydized and reduced Heme) that give raise to small scale conformation change. Engineered mutants as well are associated to small conformation changes. The strongest conformation change involves hemoglobin crystallized together with the alpha-haemoglobin-stabilizing protein (AHSP), which inhibits its capacity to react with oxygen (PDB code 1z8u). "The structure of AHSP bound to ferrous alpha-HB is thought to represent a transitional complex through which alpha-Hb is converted to a nonreactive, hexacoordinate ferric form (...) The structure of the complex shows significant conformational changes involving translocation of main chain atoms by as much as 10 Å".[44] This structure is responsible of the most serious outliers in Figure 3, in particular an almost vertical line of outliers at sequence identity ≈ 0.22, a large blob of outliers at sequence identities between 0.30 and 0.40, and a long horizontal line of outliers with SI > 0.4.

Aldolase

The most relevant conformation change in this superfamily involves a mutant (Y24F) of the protein glicolate oxidase, PDB code 1gylB. This mutant could not be crystallized with its natural cofactor FMN. According to the authors, "the absence of the cofactor FMN and differences in packing of the subunits give raise to much larger differences in the structure than the mutation per se".[45] This structure is involved in most outliers with sequence identity larger than 0.5.

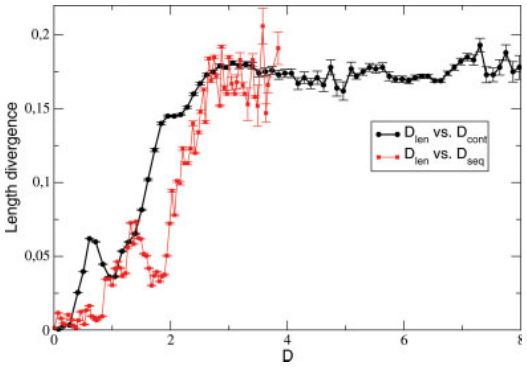


Figure 6

For the case of globins, we show the effect of sequence and structure divergence on the average length divergence. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

NADP

In this superfamily, the largest conformation change involves the protein Abeta-binding alcohol dehydrogenase (ABAD), PDB code 1so8A which displays substantial distortion of the NAD-binding pocket and the catalytic triad.[46] Other smaller conformational changes involve the Enoyl-acyl carrier reductase of *Plasmodium falciparum* crystallized with different ligands.[47]

Ploop

In this superfamily, the most severe outliers (in particular, those with SI > 0.6) involve three structures of the protein p21(H-ras) studied at different time points along the GTPase reaction with the synchrotron Laue method[48] (PDB codes 1plj, 1plk and 1pll).

Relationship between length divergence and structure divergence

Finally, we studied how length divergence (defined in Materials and Methods) influences sequence and structure divergence and is influenced by it. Large length differences between two proteins are an indication that large insertions and deletions have occurred in their evolution. However the contrary does not hold, i.e., even proteins of the same length may undergo multiple insertions and dele-

tions in their evolution. We show in Figure 6 results for the case of Globins. Other superfamilies yield qualitatively similar pictures. As expected, length divergence increases more or less gradually with sequence and structure divergence. It then reaches a plateau, more or less corresponding to the cross-over of the structural divergence explosion. Beyond the cross-over, most protein pairs differ significantly in length. Interestingly, large length divergence strongly predicts function change (see Supporting Information Fig. 3). On the other hand, similarity in length is a weak predictor of sequence, structure and function conservation (data not shown). In this case, the four superfamilies yield different pictures: whereas for Globins proteins with similar length tend to be similar in sequence, structure and function, this is less true for the other superfamilies.

DISCUSSION

In this study, we examined how protein function change and protein function conservation quantitatively influence the relationship between sequence and structure divergence in evolution. We quantified structure divergence through a novel measure, the contact divergence, which is based on the similarity of contact matrices. This measure is evolutionarily motivated, since it is constructed in analogy with a sequence divergence measure grounded in molecular evolution studies, and it is properly normalized both for related and for unrelated pairs, in such a way that it is suitable both for evolutionary analysis and for protein structure classification. We tested that this measure is more consistent with evolution based classifications than other previously proposed measures of structure divergence, and that it allows to better represent the molecular clock of protein structure divergence.

Our first qualitative conclusion confirms that, for small divergence, structure divergence and sequence divergence are proportional, as previously shown by Chothia and Lesk using as structure divergence measure the RMSD.[13] This implies that the molecular clock hypothesis approximately holds also for protein structure divergence if it holds for sequence divergence. The approximate validity of the molecular clock is also supported by our finding that networks constructed using structure similarity have clustering coefficient close to one, so that they are consistent with phylogenetic trees. Therefore, we can use structure similarity to reconstruct evolutionary trees for protein structures.

Secondly, our results show that proteins that perform exactly the same molecular function are limited in their sequence and, even more, structure divergence. Although this result is expected as a consequence of functional constraints on protein structure, the strength of these constraints is somewhat surprising. Notice that conservation seems to act on global structure similarity measures, not only on the active site. This is at first surprising, but it is consistent with the idea that allosteric effects at the level of the whole structure are important for protein function. This finding may have important consequences both for protein structure and for protein function prediction. Concerning structure prediction, if two proteins perform the same function they will have very similar structures even if their sequence identity is below the twilight zone, and the known structure of one of them will be a good template for homology or threading based modeling of the other one even at very low identity. Concerning function prediction, we have seen that structure divergence larger than a threshold is an almost certain indication of some (possibly subtle) function change. We have also shown that this observation can be used to identify electronically annotated functions that are likely to be incomplete or wrong. This observation can be therefore used to improve automatic annotation methods. The complete linkage clustering method, which forbids to join in the same cluster any two proteins with divergence larger than a threshold, should be the natural way to exploit structural information for automatic function prediction. We found that the ROC plot for predicting protein function from structure similarity have an area under the curve very close to one, meaning that it is possible to achieve very good prediction accuracy if the structure is known, or if it can be predicted through threading methods.

Third, we have found that the rate of structure versus sequence divergence is larger for proteins performing different functions than for proteins performing the same function. This acceleration may be attributed either to positive selection for new function or to relaxation of negative selection for structure conservation upon function change. We prefer the first interpretation, since the acceleration is also observed for low structure divergence, which is compatible with function conservation. The accelerations of the rate of structure divergence are also supported by the observation that the clustering coefficient of networks constructed with measures of sequence and structure divergence present significant dips, indicating violations of the molecular clock hypothesis, and by the finding that protein sequence evolution is more clock-like than structure evolution, also based on the analysis of the clustering coefficient. We conjecture that this is due to the acceleration of the rate of structural evolution in the presence of positive selection for functional changes. We will test this hypothesis in future work. In any case, this finding also suggests a way to improve protein function prediction when structure information is available. In fact, sequence and structure conservation is not sufficient to unambiguously decide that two proteins perform exactly the same function. Complementing structure similarity with a test of the constancy of the evolutionary rate may improve the accuracy of function prediction.

Fourth, we have observed that, below a cross-over value of sequence identity, there is an explosion of structural di-

versity, which may increase much faster than linearly with sequence divergence for proteins with different functions. This finding extends the previous finding of Chothia and Lesk based on the RMSD as divergence measure. The simplest explanation for such an explosion is that, below the cross-over, sequence identity does not allow to estimate the evolutionary divergence time, so that protein pairs with identity below the cross-over may have diverged for a time much longer than what is inferred from their sequence identity, allowing them to reach very different conformations. Despite this simple explanation is supported by the observed relationship between the cross-over values and the protein length, it is interesting that a qualitatively similar explosion of structural diversity has been found in a recent study of protein sequence design.[49] In this study, protein sequences were designed by optimizing the folding stability of a target structure. It was found that, when the target structure and the reference structure in the PDB are very similar, the designed sequence has a rather large identity with the reference sequence. However, when the target and the reference structure become more different, as it would be in case of selection for new function, designed and reference sequence only share very low identity, of the order of twenty percent, i.e., slightly more than the average identity of unrelated protein pairs. This phenomenon has the appearance of a cross-over in the relationship between sequence divergence and structure divergence, very much reminiscent of the one that we observed, and it may provide an alternative explanation for it: When two proteins perform the same function, natural selection targets very similar structures, determining sequence and structure conservation, whereas for proteins with significantly different function natural selection targets different structures, whose typical sequence identities are below the cross-over region. This interpretation is consistent with the findings, here reported, that protein function influences evolution by limiting the extent of sequence and structure divergence in case of function conservation, and by accelerating structure divergence with respect to sequence divergence in case of function change.

Finally, we observed that large length divergence, which is an indication of insertions and deletions, are almost always associated with functional changes (see Supporting Information Fig. 3), but length conservation is not an indicator of functional conservation. In other words, large differences of length of homologous proteins are a strong hint of functional change, i.e., large length differences are hardly neutral under a functional point of view.

MATERIALS AND METHODS

Protein sets

In this work, we used five protein domain sets. (1) A representative set of protein domains having less than 40 percent sequence identity, which are decomposed almost identically in the CATH and SCOP database (consensus set available at the URL: http://ub.cbm.uam.es/research/ProtNet.php). (2) Four superfamilies: Globins, Aldolases (TIM barrel fold), P-loop containing nucleoside triphosphate hydrolases and NADP-binding Rossmann-fold domains. The list and the definition of domains in each superfamily were taken from the CATH database.[15] From all sets, we eliminated NMR structures, domains extracted from multi-domain chains, for which the function assignment is problematic, and domains with both very high structure and sequence identity (the product of sequence identity times contact overlap must be smaller than 0.98). The sequence identity and contact overlap, respectively, took values in the ranges (0.01, 1.00) and (0.13, 1.00) for Globins, (0.01, 1.00) and (0.08, 1.00) for Aldolases, (0.00, 1.00) and (0.04, 1.00) for P-loop, (0.00, 1.00) and (0.00, 1.00) for NADP.

Function characterization

We retrieved Gene Ontology (GO)[39] terms for PDB chains from the web page of the Structure integration with function, taxonomy and sequence (SIFTS) initiative (http://www.ebi.ac.uk/msd/sifts/). To avoid wrong assignments of GO terms to CATH domains, we removed those cases were more than one CATH domain correspond to the same PDB chain. From GO terms, we used only the molecular function annotation and we removed annotations contained in paths already assigned to the same PDB chain.

For globins, GO terms were not specific enough, so we also used InterPro Signatures.[40] Notice that InterPro signatures do not necessarily yield a classification, but we verified that they do in the case of Globins, i.e., that in this set having the same InterPro signature is an equivalence relationship. To retrieve these signatures, we used the SSMap tool[50] that relating PDB chains with UniProt accessions, which also include InterPro Signatures.

We considered GO terms to be manually assigned if their evidence code was EXP (Inferred from Experiment), IDA (Inferred from Direct Assay), IPI (Inferred from Physical Interaction), IMP (Inferred from Mutant Phenotype), IGI (Inferred from Genetic Interaction), IEP (Inferred from Expression Pattern) or TAS (Traceable Author Statement). All other evidence codes, such as for instance ISS (Inferred from Sequence or Structural Similarity), were attributed to computational analysis. The number of manually annotated domains is dramatically reduced: 92 over 676 for NADP, 533 over 1209 for P-loop, 272 over 1341 for Aldolases, 702 over 1313 for Globins.

Divergence measures

For each pair of domains in the same superfamily structure alignments were computed using a new version

of the program Mammoth[33] which was improved performing the same two steps dynamic programming procedure implemented in the multiple alignment version Mammoth-mult,[51] and optimizing the corresponding parameters (UB, APG, Florian Teichert and Markus Porto, unpublished). We measured pairwise dissimilarities in structure, sequence, function and length.

1. Sequence divergence D_{seq} was computed from the sequence identity SI $\in [0, 1]$ measured from the optimal structure alignment as

$$D_{\mathrm{seq}} = -\log(\mathrm{SI}). \tag{5}$$

   Here and in the following, log indicates Neperian logarithms.
2. Function similarity was based on GO terms[52] or on InterPro signatures[40] in the case of Globins. Two proteins where considered to perform the same function ($D_{\mathrm{fun}} = 0$) if all of their GO terms or InterPro signatures coicided, otherwise they were regarded as performing different functions ($D_{\mathrm{fun}} = 1$).
3. For two proteins A and B, we measure their length difference and define the dimension-less variable $d_{\mathrm{len}}(A, B)$ as

$$d_{\mathrm{len}}(A, B) = \frac{|L_{\mathrm{A}} - L_{\mathrm{B}}|}{\sqrt{L_{\mathrm{A}} L_{\mathrm{B}}}} \tag{6}$$

   We observed that $d_{\mathrm{len}} < 1$ for all pairs of proteins in the superfamilies that we examined. We therefore defined the corresponding length divergence as $D_{\mathrm{len}} = -\log(1 - d_{\mathrm{len}})$, in analogy with sequence or structure divergence (notice that this variable is not defined if $d_{\mathrm{len}} \geq 1$, i.e., if $L_{\mathrm{A}}/L_{\mathrm{B}} > 2.6$).
4. The contact overlap is a convenient measure of protein structure similarity, which counts the fraction of contacts in common between two aligned protein structures A and B. The contact matrix of protein A, $C_{ij}^{(A)}$, is defined such that $C_{ij}^{(A)}$ equals one if two heavy atoms of residues i and j are closer than 4.5 Å and $|i - j| \geq 5$, and zero otherwise, so that we do not consider short range contacts. As the same short range contacts are formed with higher probability in unrelated structures, eliminating them has the effect to reduce the mean overlap of unrelated structures. We expect in this way to increase the signal to noise ratio of the contact overlap. Denoting by $a(i)$ the residue in structure B aligned with residue i in structure A, the contact overlap can be written as

$$q_{\mathrm{AB}} = \frac{\sum_{ij} C_{ij}^{(A)} C_{a(i)a(j)}^{(B)}}{\sqrt{\sum_{ij} C_{ij}^{(A)} \sum_{ij} C_{ij}^{(B)}}}. \tag{7}$$

   where summation runs over all pairs of residues in protein A.
5. The contact overlap of unrelated proteins depends on their length. We characterize the length of the protein pair as the geometric mean of the two lengths,

$$L = \sqrt{L_{\mathrm{A}} L_{\mathrm{B}}}. \tag{8}$$

   The mean $\overline{q}(L)$ and standard deviation $\sigma_q(L)$ were computed by performing pairwise alignments for the ASTRAL40 set of domains having less than 40% sequence identity, using the program MAMMOTH[33] and considering only pairs in different SCOP folds. In this case, only short regions of the two proteins superimpose in space, typically consisting of one or few secondary structure elements. For each length in the range 40 to 800 residues, $\overline{q}(L)$ and $\sigma_q(L)$ were well fitted by the power laws

$$\overline{q}(L) = 0.386 L^{-0.547} \tag{9}$$

$$\sigma_q(L) = 1.327 L^{-0.673} \tag{10}$$

   To eliminate the length dependence, we used the Z score of the overlap, subtracting the average value of the overlap of unrelated protein pairs with the same length, $\overline{q}(L)$, and dividing times the corresponding standard deviation, $\sigma_q(L)$, to obtain

$$Z = \frac{(q - \overline{q}(L))}{\sigma_q(L)} \tag{11}$$

6. As explained in the main text, the overlap q was transformed to obtain a measure of contact divergence D_{cont}, defined as

$$D_{\mathrm{cont}}(q, L) = \begin{cases} -\log\left(\frac{q - q_\infty(L)}{1 - q_\infty(L)}\right) & \text{if } q > \epsilon(L) \\ D_0 - (q - \overline{q}(L))/\sigma_q(L) & \text{otherwise} \end{cases} \tag{12}$$

   The upper line of the aforementioned equation defines the contact divergence of related proteins, in analogy to how sequence identity is tranformed to estimate evolutionary divergence, Eq. (2). It is such that $D_{\mathrm{cont}} = 0$ for proteins having identical contact matrices and $D_{\mathrm{cont}} \to \infty$ for $q \to q_\infty(L)$. The lower line defines the contact divergence of unrelated or distantly related proteins as $D_{\mathrm{cont}} = D_0 - Z$, where Z is defined in Eq. (11). The aforementioned equation depends on three parameters, the asymptotic overlap $q_\infty(L)$, the cross-over overlap $\epsilon(L)$ and the parameter D_0. They are fixed as follows. First, we make the ansatz

$$q_\infty(L) = \overline{q}(L) + A\sigma_q(L), \quad (13)$$

which means that the asymptotic overlap of distantly related proteins is larger than the mean overlap $\overline{q}(L)$ of unrelated proteins Since both $\overline{q}(L)$ and $\sigma_q(L)$ depend on protein length, so does $q_\infty(L)$. The parameter A in the aforementioned equation was fixed to the value $A = 5$ by assessing the contact divergence measure through the clustering experiments described in the main text.

The cross-over $\epsilon(L)$ is fixed imposing that Eq. (12) is continuous for $q = \epsilon(L)$. To this end, we introduce the variable $z = (\epsilon(L) - q_\infty(L))/\sigma_q(L)$. The continuity condition reads

$$z - \log(z) = D_0 - A + \log \sigma_q(L) - \log(1 - q_\infty(L)), \quad (14)$$

The function $z - \log(z)$ takes values between one, for $z = 1$, and infinite, for z tending to zero and to infinite. Therefore, two solutions of the aforementioned equations exist if and only if the right hand side is larger than one, i.e., $D_0 - A + \log \sigma_q(L) - log(1 - q_\infty(L)) > 1$. We decided to take the smallest value of D_0 such that solutions exist for all protein domains contained in our set, i.e.,

$$D_0 = 1 + A - \log \sigma_q(L_{\max}) + \log(1 - q_\infty(L_{\max})) = 10.2 \quad (15)$$

where $L_{\max} = 880$ is the length of the longest domain in all sets that we used and $A = 5$. We then numerically solved Eq. (14) for each L, taking the solution with $z < 1$, which corresponds to an ϵ with small Z score, and we obtained $\epsilon(L) = q_\infty(L) + \sigma_q(L)z(L)$. In this way, the only free parameter in the definition of the contact divergence is the parameter A that expresses the extent to which homologous proteins keep memory of their evolutionary relatedness. This parameter was fixed to the value $A = 5$ by performing the clustering tests described in the main text.

Classification analysis

We assessed the agreement of two classifications through the weighted kappa measure,[38] which uses as reference the expected agreement for two independent classifications with the same number of relationships. We define N_A (N_B) the number of related pairs in classification A (B) of the same N objects, with $N_{tot} = N(N-1)/2$ pairs in total. If A and B are independent, the number of pairs that are either related or unrelated in both A and B is given by

$$N_e = \frac{N_A N_B + (N_{tot} - N_A)(N_{tot} - N_B)}{N_{tot}} \quad (16)$$

We compare this number to the observed number of pairs that agree,

$$N_o = N_{AB} + (N_{tot} - N_A - N_B + N_{AB}), \quad (17)$$

where N_{AB} is the number of pairs that are related in both classifications From this number, the weighted kappa is computed as

$$\kappa = \frac{N_o - N_e}{N_{tot} - N_e}. \quad (18)$$

A value of zero means that two classifications are as related as independent classifications, one means that the two classifications coincide.

Clustering coefficient

The clustering coefficient of node i in a network is defined as the fraction of pairs of its neighbors j and k that are neighbors between each other, and the clustering coefficient of the network is the average clustering coefficient of its nodes. Formally, this is defined as

$$\text{Clustering coefficient} = \frac{1}{N}\sum_i \frac{2\sum_{j<k} A_{ij}A_{ik}A_{jk}}{n_i(n_i-1)} \quad (19)$$

where N is the number of nodes, A_{ij} is the adjacency matrix (one if nodes i and j are joined, zero otherwise), $n_i = \sum_j A_{ij}$ is the number of neighbors or degree of node i If the clustering coefficient is one for all nodes, connections on the network define an equivalence relationship.

We have computed the clustering coefficient for the network obtained by joining domains with similarity $S_{ij} > S_0$, for various values of S_0. To compare different similarity measures, we have plotted the clustering coefficient versus the number of disjoint components found in the network.

ROC plots

Given a binary classifier (predictor) assigning positive and negative values, and a test set of examples whose positive and negative values are considered true, the receiver operating characteristic (ROC) plots the sensitivity, or true positive rate, defined as sensitivity = TP/P versus the false positive rate or 1-specificity, FPR = FP/N, for different thresholds used for classification. The performance of the classifier is evaluated through the area under the curve (AUC) of the ROC plot, which is 0.5 for ran-

dom classifiers and 1 for a perfect classifier having sensitivity one for all thresholds.

Conditional averages

For studying the relationship between different types of divergence measures, we measured the conditional average of one variable conditioned to values of the other variable in a given interval.

ACKNOWLEDGMENTS

The authors is gratefully dedicate this article to Ángel Ramirez Ortiz, who guided their studies of protein structure evolution and classification and has been a great mentor, colleague, and friend. The authors acknowledge interesting discussions with Julián Echave, Florian Teichert, and Markus Porto. This work has been financed through a Ramón y Cajal fellowship to UB and through projects BIO2008-04384 and CSD200623 of the Spanish Ministery of Science and Innovation.

REFERENCES

1. Bromham L, Penny D. The modern molecular clock. Nature Reviews Genetics 2003;4:216–224.
2. Zuckerkandl E, Pauling L. In: Kasha M, Pullman B, editors. Horizons in biochemistry. New York: Academic Press; 1962.
3. Kimura M. Evolutionary rate at the molecular level. Nature 1968; 217:624–626.
4. King J-L, Jukes TH. Non-Darwinian evolution. Science 1969;164: 788–798.
5. Ohta T, Kimura M. On the constancy of the evolutionary rate of cistrons. J Mol Evol 1971;1:18–25.
6. Kimura M. The neutral theory of molecular evolution. Cambridge: Cambridge University Press; 1983.
7. Ohta T. Role of very slightly deleterious mutations in molecular evolution and polymorphism. Theor Pop Biol 1976;10:254–275.
8. Durrett R. Probability models for DNA sequence evolution. New York: Springer; 2002.
9. Sella G, Hirsch AE. The application of statistical physics to evolutionary biology. Proc Natl Acad Sci USA 2005;102:9541–9546.
10. Bastolla U, Moya A, Viguera E, van Ham RCHJ. Genomic determinants of protein folding thermodynamics. J Mol Biol 2004;343: 1451–1466.
11. Gillespie JH. The causes of molecular evolution. Oxford: Oxford University Press; 1991.
12. Graur D, Li WH. Fundamentals of molecular evolution. Sinauer, Sunderland: Vagaries of the molecular clock; 2000.
13. Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. EMBO J 1986;5:823–826.
14. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
15. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH-a hierarchic classification of protein domain structures. Structure 1997;5:1093–1108.
16. Grishin NV. Fold change in evolution of protein structures. J Struct Biol 2001;134:167–185.
17. Krishna SS, Grishin NV. Structural drift: a possible path to protein fold change. Bioinformatics 2005;21:1308–1310.
18. Viksna J, Gilbert D. Assessment of the probabilities for evolutionary structural changes in protein folds Bioinformatics 2007;23:832–841.
19. Pascual-García A, Abia D, Ortiz AR, Bastolla U. Cross-over between discrete and continuous protein structure space: Insights into automatic classification and networks of protein structures. PLoS Comput Biol 2009;5:e1000331.
20. Devos D, Valencia A. Practical limits of function prediction. Proteins 2000;41:98–107.
21. Wilson CA, Kreychman J, Gerstein M. Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. J Mol Biol 2000;297:233–249.
22. Todd AE, Orengo CA, Thornton JM. Evolution of function in protein superfamilies, from a structural perspective. J Mol Biol 2001; 307:1113–1143.
23. Lecomte JT, Vuletich DA, Lesk AM. Structural divergence and distant relationships in proteins: evolution of the globins. Curr Opin Struct Biol 2005;15:290–301.
24. Sangar V, Blankenberg DJ, Altman N, Lesk AM. Quantitative sequence-function relationships in proteins based on gene ontology. BMC Bioinformatics 2007;8:294.
25. Shakhnovich BE, Dokholyan NV, DeLisi C, Shakhnovich EI. Functional fingerprints of folds: evidence for correlated structure-function evolution. J Mol Biol 2003;326:1–9.
26. Whisstock JC, Lesk AM. Prediction of protein function from protein sequence and structure. Q Rev Biophys 2003;36:307340.
27. Friedberg I. Automated protein function predictionthe genomic challenge. Brief Bioinf 2006;7:225–242.
28. Ponomarenko JV, Bourne PE, Shindyalov IN. Assigning new GO annotations to protein data bank sequences by combining structure and sequence homology. Proteins. 2005;58:855–865.
29. Wang K, Horst JA, Cheng G, Nickle DC, Samudrala1 R. Protein meta-functional signatures from combining sequence, structure, evolution, and amino acid property information. PLoS Comput Biol 2008;4:e1000181.
30. Shakhnovich BE, Max Harvey J. Quantifying structure-function uncertainty: a graph theoretical exploration into the origins and limitations of protein annotation. J Mol Biol 2004;337: 933–949.
31. Shakhnovich BE. Improving the precision of the structure-function relationship by considering phylogenetic context. PLoS Comput Biol 2005;1:e9.
32. Holm L, Sander C. Protein structure comparison by alignment of distance matrices. J Mol Biol 1993;233:123–138.
33. Ortiz AR, Strauss C, Olmea O. MAMMOTH (Matching Molecular Models Obtained from Theory): an automated method for model comparison. Protein Sci 2002;11:2606–2621.
34. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins 2004;57:702–710.
35. Jones DT, Taylor WR, Thornton JM. The rapid generation of mutation data matrices from protein sequences. CABIOS 1992;8:275–282.
36. Nei M, Kumar S. Molecular evolution and phylogenetics. Oxford: Oxford University Press; 2000.
37. Bastolla U, Porto M, Roman HE, Vendruscolo M. Statistical properties of neutral evolution. J Mol Evol 2003;57 (Suppl 1):S103–S119.
38. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 1968;70:213–220.
39. Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. Nature Genet 2000;25:25–29.
40. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, Finn RD, Gough J, Haft D, Hulo N, Kahn D, Kelly E, Laugraud A, Letunic I, Lonsdale D, Lopez R, Madera M, Maslen J, McAnulla C, McDowall J, Mistry J, Mitchell A, Mulder N, Natale D, Orengo C, Quinn AF, Selengut JD, Sigrist CJ, Thimma M, Thomas PD, Valentin F, Wilson D, Wu CH, Yeats C. InterPro: the integrative protein signature database. Nucleic Acids Res 2009;37 (Database Issue):D211–D215.

41. Clegg JB, Gagnon J. Structure of the zeta chain of human embryonic hemoglobin. Proc Natl Acad Sci USA 1981;78:6076–6080.
42. Giardina B, Messana I, Scatena R, Castagnola M. The multiple functions of hemoglobin. Crit Rev Biochem Mol Biol 1995;30: 165–196.
43. Rammal R, Toulouse G, Virasoro MA. Ultrametricity for physicists Rev Mod Phys 1986;58:765–788.
44. Feng L, Zhou S, Gu L, Gell DA, Mackay JP, Weiss MJ, Gow AJ, Shi Y. Structure of oxidized alpha-haemoglobin bound to AHSP reveals a protective mechanism for haem. Nature 2005;435:697–701.
45. Stenberg K, Clausen T, Lindqvist Y, Macheroux P. Involvement of Tyr24 and Trp108 in substrate binding and substrate specificity of glycolate oxidase. Eur J Biochem 1995;228:408–416.
46. Lustbader JW, Cirilli M, Lin C, Xu HW, Takuma K, Wang N, Caspersen C, Chen X, Pollak S, Chaney M, Trinchese F, Liu S, Gunn-Moore F, Lue LF, Walker DG, Kuppusamy P, Zewier ZL, Arancio O, Stern D, Yan SS, Wu H. ABAD directly links Abeta to mitochondrial toxicity in Alzheimer's disease. Science 2004;304: 448–452.
47. Perozzo R, Kuo M, Sidhu AS, Valiyaveettil JT, Bittman R, Jacobs WR Jr, Fidock DA, Sacchettini JC. Structural elucidation of the specificity of the antibacterial agent triclosan for malarial enoyl acyl carrier protein reductase. J Biol Chem 2002;277:13106–13114.
48. Scheidig AJ, Sanchez-Llorente A, Lautwein A, Pai EF, Corrie JE, Reid GP, Wittinghofer A, Goody RS. Crystallographic studies on p21(H-ras) using the synchrotron Laue method: improvement of crystal quality and monitoring of the GTPase reaction at different time points. Acta Cryst 1994;D50:512–520.
49. Ding F, Dokholyan NV. Emergence of protein fold families through rational design. PLoS Comp Biol 2006;2:e85.
50. David FP, Yip YL. SSMap: A new UniProt-PDB mapping resource for the curation of structural-related information in the UniProt/ Swiss-Prot Knowledgebase. BMC Bioinformatics 2008;9:391.
51. Lupyan D, Leo-Macias A, Ortiz AR. A new progressive-iterative algoithm for multiple structure alignment. Bioinformatics 2005;21: 3255–3263.
52. Wang JD, Du Z, Payattakool R, Yu PS, Chen CF. A new method to measure the semantic similarity of GO terms. Bioinformatics 2007; 23:1274–1281.

Chapter 3

Microbial systems

A great truth is a truth whose opposite is a great truth.

Niels Bohr

Summary

In this chapter we present the results found for bacterial systems, where we investigate the determinants influencing the distribution of bacterial taxa observed in different samples. As it was outlined in the introduction, we aim to infer whether there is any role for the ecological interactions to explain this distribution.

It is interesting to infer interactions from this kind of data, because we consider many different environments and it is known that interactions between bacteria may change under different environmental conditions [Klitgord and Segrè (2010)]. Therefore, if it is possible to justify that inferred relationships are actual ecological interactions, these may be interactions conserved during the evolutionary process, similarly to the structural cores found for proteins.

The global strategy relies also in the comparison of microstates (bacterial samples, whereas we compared protein structures in the previous section), searching for similarities that may be a consequence of underlying ecological interactions. Given the nature of our data, patterns of co-occurrence (aggregation) and exclusion (segregation) between the different taxa may arise from these comparisons, from which we must address their significance, and then further interpret the results in terms of ecological interactions.

The procedure we designed to assess the significance of these comparisons consists on the application of pattern-based (null) models aiming to reproduce the observed distribution of taxa, which is a characteristic approach in classical Biogeography. We consider as a null hypothesis that species do

not interact, and therefore the patterns obtained from the random procedure do not contain information about species interactions. In this way, patterns identified in the observed data departing significantly from those obtained for random data, could be interpreted as a signature of an interaction. We generated these random patterns through generalized linear models [Bolker et al. (2009)]. In these methods, the average of a statistical distribution selected from the exponential family is estimated through a regression, where the predictors are related to the constraints found.

In the first work presented here [Pascual-García et al. (2014b)] we adapted the proposal of Navarro-Alberto and Manly [Navarro-Alberto and Manly (2009)] to consider environmental information. We dealt with a large compositional matrix, recovered from experiments sequencing the 16S rRNA gene of bacterial taxa –a gene which is *in principle* characteristic for each specie, and thus it is often used to build Operational Taxonomic Units–, which result in more than two thousand samples from almost fifty diverse environments. Moreover, we proposed a measure that allowed us to estimate the probability that an observed aggregation or segregation of two taxa could be obtained by chance –although we must remember that this is not a completely random process, given that several constraints are implicitly considered–.

The results revealed a surprisingly large number of significant relationships, being most of them aggregations. Taking into account the coarse grained nature of our data, we cannot conclusively reject the null hypothesis, given that we cannot include explicitly in the null model all the environmental features that may influence the outcome of these communities. Whether the patterns found actually reflect ecological interactions or a more simple hypothesis such as habitat filtering, must be further verified through experiments handling indirect evidences to support one or the other hypothesis.

We found that many aggregations take place in a wide variety of environments, which makes unlike any underlying habitat preference driving their appearance. Many of them are cosmopolitan taxa without well defined habitat preferences. Indeed, cosmopolitan taxa have stronger propensity to aggregate than specialists, a result opposite to what is found following the null model. In addition, we have built a network considering the significant aggregations, and it has a marked structure with high clustering coefficient and nestedness, in analogy with mutualistic networks of plants and pollinators. Last, but not least, we verified that several of the significant aggregations found are indeed known cooperative relationships.

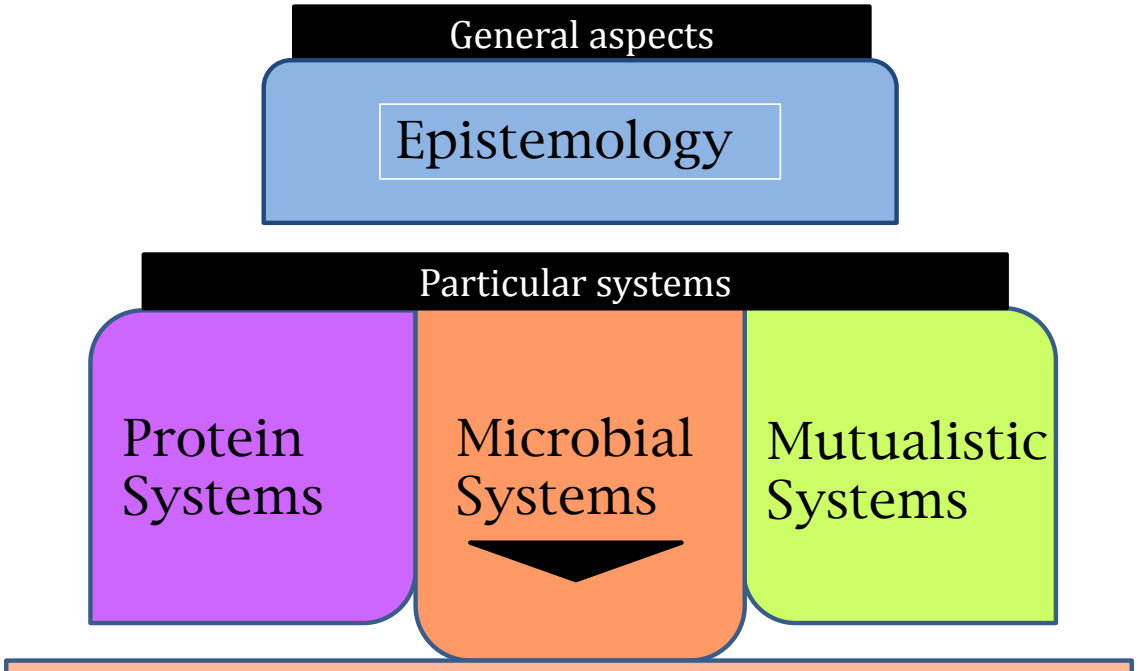
Altogether, these results suggest that mutualistic associations are common in the bacterial world, and they are key to explain the remarkable cosmopolitanism of bacteria and the diversity of their communities. Moreover, we find that phylogenetically related taxa have a strong tendency to aggregate and avoid segregation, and this signal is stronger for closer taxa. This observation is consistent with an evolutionary scenario where metabo-

lic syntrophy and interchange of genetic material would have a relevant role in the adaptation of taxa to the different environmental conditions. Under this scenario, phylogenetic similarity would facilitate the success of events such as horizontal gene transfer, challenging a paradigm where competitive exclusion would be the main driver in bacterial speciation.

In the second work [Vallès et al. (2014)] we applied a generalized linear model to deal with matrices of taxa and samples where also their abundances are considered, and we measure taxa aggregation and segregation considering classical ressemblance measures [Legendre and Legendre (2012)]. The experimental data was obtained from metagenomics experiments where faeces from a group of infants were sampled during their first year of life, together with those of their mothers before and after their children's birth. From these samples microbial genes were sequenced and associated to bacterial taxa in order to estimate the putative bacterial diversity in the gut, from which we inferred significant relationships at the different stages.

Two clearly differentiated trends were found, mainly determined by the diet of the infants, being the first trend related with milk intake and the second one with the introduction of solid food. Nevertheless, the bacterial assemblage reflected a systematic increase in the ressemblance of the infants'communities with respect to those observed in the mothers. Interestingly, the information obtained allowed us to understand which relationships are more important for the gut bacterial development, identifying a core of bacterial taxa with significantly positive relationships. This core may play a critical role in the assemblage of other taxa and, in turn, in the structural outcome of the community. Indeed, the comparison of these relationships with those obtained from the previously discussed work are consistent [Vallès et al. (2014)], further facilitating the identification of the environmental preferences of the different microbes. Moreover, an important question to start building a picture of the ecological dynamics in gut is whether each microbe can be identified as an obligatory or facultative guest. As a result, the core inferred from the infants'samples seems to be embedded within a larger core of obligatory bacteria, providing further support to its interpretation as a scaffold in gut development. On the other hand, most of the bacterial species that dissapeared during the infants'development were rare species, and its environmental affinities revealed a rather facultative role.

3.1. Article [MIC-1]



1) Pascual-García, A., Tamames, J., and Bastolla, U. (2014) Bacteria dialog with Santa Rosalia: Are aggregations of cosmopolitan bacteria mainly explained by habitat filtering or by ecological interactions? *BMC microbiology*, *14*(1), 1-16.

2) Vallés Y., Artacho A., Pascual-García A., Ferrús M.L., Gosalbes M.J., Abellán J.J. and Francino M.P. (2014) Microbial succession in the gut: directional trends of taxonomic and functional change in a birth cohort of Spanish infants. *PLoS Genet.* 10(6): e1004406

Pascual-García *et al. BMC Microbiology* 2014, **14**:284
http://www.biomedcentral.com/1471-2180/14/284

BMC Microbiology

RESEARCH ARTICLE Open Access

Bacteria dialog with Santa Rosalia: Are aggregations of cosmopolitan bacteria mainly explained by habitat filtering or by ecological interactions?

Alberto Pascual-García[1], Javier Tamames[2] and Ugo Bastolla[1*]

Abstract

Background: Since the landmark Santa Rosalia paper by Hutchinson, niche theory addresses the determinants of biodiversity in terms of both environmental and biological aspects. Disentangling the role of habitat filtering and interactions with other species is critical for understanding microbial ecology. Macroscopic biogeography explores hypothetical ecological interactions through the analysis of species associations. These methods have started to be incorporated into microbial ecology relatively recently, due to the inherent experimental difficulties and the coarse grained nature of the data.

Results: Here we investigate the influence of environmental preferences and ecological interactions in the tendency of bacterial taxa to either aggregate or segregate, using a comprehensive dataset of bacterial taxa observed in a wide variety of environments. We assess significance of taxa associations through a null model that takes into account habitat preferences and the global distribution of taxa across samples. The analysis of these associations reveals a surprisingly large number of significant aggregations between taxa, with a marked community structure and a strong propensity to aggregate for cosmopolitan taxa. Due to the coarse grained nature of our data we cannot conclusively reject the hypothesis that many of these aggregations are due to environmental preferences that the null model fails to reproduce. Nevertheless, some observations are better explained by ecological interactions than by habitat filtering. In particular, most pairs of aggregating taxa co-occur in very different environments, which makes it unlikely that these associations are due to habitat preferences, and many are formed by cosmopolitan taxa without well defined habitat preferences. Moreover, known cooperative interactions are retrieved as aggregating pairs of taxa. As observed in similar studies, we also found that phylogenetically related taxa are much more prone to aggregate than to segregate, an observation that may play a role in bacterial speciation.

Conclusions: We hope that these results stimulate experimental verification of the putative cooperative interactions between cosmopolitan bacteria, and we suggest several groups of aggregated cosmopolitan bacteria that are interesting candidates for such an investigation.

Keywords: Bacterial ecology, Habitat filtering, Biodiversity, Cooperation, Syntrophy, Bacterial speciation, Black queen hypothesis, Ecological null-models, Ecological networks

*Correspondence: ubastolla@cbm.csic.es
[1]Centro de Biología Molecular Severo Ochoa (CSIC-UAM), c. Nicolás Cabrera 1, campus UAM, E-28049 Madrid, Spain
Full list of author information is available at the end of the article

Pascual-García *et al. BMC Microbiology* 2014, **14**:284
http://www.biomedcentral.com/1471-2180/14/284

Background

In his seminal paper *Homage to Santa Rosalia* or *Why are there so many kinds of animals?* [1], George E. Hutchinson addressed the determinants of biodiversity in light of a renewed concept of niche. Hutchinson's question has an interesting challenge in the microbial world. Understanding the determinants of bacterial niches can open new perspectives on plant and animal evolution as well, since bacteria co-evolved with multicellular eukaryotes for hundreds of millions of years mutually influencing each other [2], and it may have important biomedical applications. An increasing quantity of data from high-throughput experiments is now available for large scale ecological studies of bacterial communities, in particular for the human microbiome [3-6]. Early analysis suggested that ecological patterns are qualitatively similar for macro- and microorganisms [7] and allowed identifying taxa-area and distance decay relationships [8,9] and the influence of environmental variables such as depth [10] or salinity [11], stimulating the emergence of prokaryotic biogeography [12,13].

Data on presence or absence of species in different locations are used by biogeographists to infer ecological processes [14]. Similarly, presence-absence matrices obtained by sequencing environmental samples or by mining abstracts of scientific papers [15] offer new opportunities to shed light son bacterial ecology. Recently, several groups used large-scale data to study bacterial associations [16-24], reviewed in [25]. In the present work, we analyse bacterial species associations for a comprehensive collection of samples from a large variety of environments classified at the three hierarchical levels of environmental subtype, type and supertype [26]. We assess the significance of their associations by means of a recently proposed null-model [27] that optimally reproduces the global distribution of taxa across samples, and that we modified to take into account environmental preferences. To this end, we exploited the hierarchical classification of samples into environmental groups developed by Tamames et al. [26] (see Additional file 1: Figure S1 and Table S1) and developed a new analytical pairwise score.

We are interested in pairs of taxa that aggregate and segregate, which means that they co-occurr significantly more often and less often than expected based on the null model. Aggregations and segregations can be attributed to habitat preferences, to direct ecological interactions (cooperative interactions, commensalism and parasitism for aggregation, competitive interactions for segregation) or to indirect interactions with another species or group of species. We consider environmental preferences in our null model, with the aim to reduce the number of aggregations that are due to common preferences. We try to estimate how many aggregations may be due to environmental preferences that are not removed by the null model by differentiating associations that occur in a specific environment from those that are not specific. Another way to perform this analysis consists in focusing on cosmopolitan taxa that do not show apparent environmental preferences but are found in many diverse environments. The previous work of Tamames et al. [26] found that, at the genus level that we consider here, cosmopolitanism is not rare among bacteria. We aim at investigating in which way ecological associations may contribute to this property.

In this work, we assess all possible segregations and aggregations of 1187 bacterial taxa corresponding to the genus level, observed in 2322 samples from different environments, and we analyse the relationship between these environmental associations on the one hand and cosmopolitanism and known ecological associations on the other hand.

Results

Constructing networks of bacterial associations

Our first aim is to construct a null model that optimally represents the global distribution of taxa across samples, considering their habitat preferences at the level of environmental subtypes, but assuming that taxa do not interact between themselves. This approach is different from most current approaches to microbial community studies in that it explicitly considers habitat preferences, and in that the association score of a pair of taxa depends on the observed distribution of all other taxa.

Since for many samples abundance information is not present, the 2322 samples were transformed into the binary presence-absence matrix $X_{ia} \in \{0,1\}$, where i labels one of the N taxa and a labels one of the M samples. To limit the bias caused by the choice of primers in sequencing experiments, we excluded experiments targeted at detecting specific taxa (see Methods). We adopt the probabilistic null model proposed by Navarro-Alberto and Manly [27], in which the probability π_{ia} that taxon i is observed at sample a in the absence of taxa interactions is parametrized as $\pi_{ia} = 1 - \exp(-p_i q_a)$ where the parameter p_i is related with the abundance of taxon i and q_a is related with the biodiversity supported by sample a, respectively.

The $M+N$ parameters p_i and q_a are determined by maximum likelihood, so that the resulting null model is most difficult to reject. We take into account that each taxon has a preference for some habitat by assuming that the taxa parameters $p_i(A)$ are specific for each environmental subtype A (see Methods). If taxon i is never found in environment A, then $p_i(A) = 0$, implying that $\pi_{ia} = 0$ if $a \in A$, i.e. the taxon is never found in samples of environment A simulated through the null model either.

The significance of the observed co-occurrences is analytically assessed through the *aggregation score* $S_{ij}^{\mathrm{A}} = -\log \mathrm{P}_{ij}\left(n \geq n_{ij}\right)$, where n_{ij} is the observed number of samples where taxa i and j co-occur and $\mathrm{P}_{ij}(n)$ is the null-model probability that taxa i and j co-occur at n locations (see Association scores in Methods). Similarly, we compute the *segregation score* as $S_{ij}^{\mathrm{S}} = -\log \mathrm{P}_{ij}\left(n \leq n_{ij}\right)$. These computations are performed analytically and they last few minutes on an ordinary computer even for large systems.

We compute the significance of $N(N-1)/2 = 703891$ potential associations between all pairs of taxa for the observed matrix as well as for 100 random realizations generated through the null model. To correct for multiple testing and reduce the dependence on the number of samples where i and j are present, scores are transformed into Z scores over random realizations of the null model. (see Additional file 1: Figure S2). Large Z scores are found with the observed matrix but not with realizations.

Number of aggregations and segregations for comparable significance thresholds

The reconstructed association network depends on the threshold above which associations are considered significant. In order to choose these thresholds in a comparable way for aggregations and segregations, we generate random realizations of presence-absence matrices using the null model π_{ia}, and we treat them in the same way as the observed matrix, computing their null model π'_{ia}, the association scores and the number of inferred associations for given threshold. Since in the null model taxa do not interact, these inferred associations represent false positives. In this way, we estimate the false positive rate (FPR) and the positive predictive value (PPV) as a function of the threshold.

We plot in Figure 1 the number of inferred associations versus the PPV. For equal PPV, the number of aggregations is larger than the number of segregations. The same qualitative result is found using the FPR as control variable (see Additional file 1: Figure S3).

A possible artefact that can produce this result is that our method does not allow to detect significant associations for all pairs of taxa. For instance, if two taxa never co-occur in the same environmental subtype, they never co-occur in the null model as well and their segregation score is zero. In general, two taxa can have a significant aggregation (segregation) score only if the probability that they always (never) co-occur is smaller than the chosen threshold. To take this into account, for each threshold we consider only pairs of taxa for which both segregation and aggregation can be detected (consensus set). Also in this case aggregation prevails over segregation: for PPV=0.96, which represents a good compromise between completeness and accuracy, we find 2313 aggregations and 628 segregations (see Table 1). The results shown in the following are obtained using these thresholds but considering also pairs for which only one association type can be detected.

We compared our predicted associations with those obtained by Freilich et al. [28]. These authors predicted potential cooperative and competitive interactions of bacterial species based on the simulation of their metabolic networks in different environments, and complemented

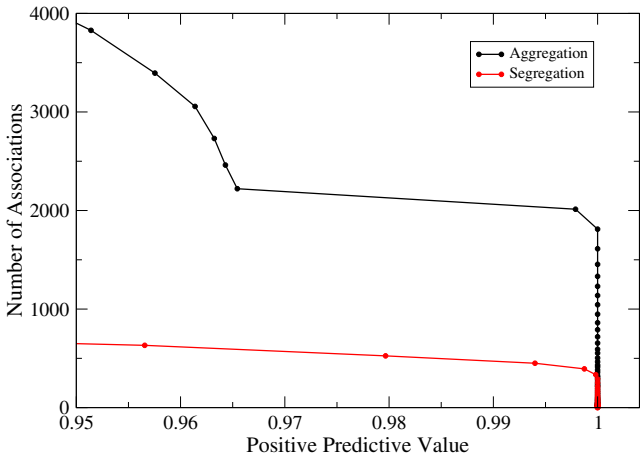


Figure 1 Number of significant aggregations (black) and segregations (red) found in the observed presence-absence matrix versus the positive predictive value (see Methods).

Pascual-García *et al. BMC Microbiology* 2014, **14**:284
http://www.biomedcentral.com/1471-2180/14/284

Table 1 Properties of networks obtained with Positive Predictive Value 0.96

Type	Pairs	Threshold	FPR	Associations
Aggregation	All	5.75	$4.1 \cdot 10^{-4}$	3394
Segregation	All	5.75	$2.4 \cdot 10^{-4}$	632
Aggregation	Cons.	4.75	$2.0 \cdot 10^{-4}$	2313
Segregation	Cons.	5.75	$6.2 \cdot 10^{-5}$	628

these associations through the analysis of pairs of species that co-occurr in metagenomic experiments more and less often than expected by chance. This study differs from ours in three important ways: (1) It assesses the significance of aggregations and segregations through the hypergeometric distribution, which only depends on the pair of taxa examined, instead of using a global null model that also accounts for the absence/presence of other taxa; (2) More importantly, it only examines pairs of species that show the same environmental preferences, while our method takes care of removing environmental preferences through the null model (3) Finally, it considers the species level, whereas we detect associations between genera.

Despite these differences, the results are strikingly similar. Out of 62 genera for which we could identify a correspondence with 73 species studied by Freilich et al., we found 26 aggregations and 16 segregations over 1891 possible associations (1.4 % and 0.8%, respectively) and they found 49 aggregations and 16 segregations over 2628 possible associations (1.9% and 0.6%, respectively). Nine of our aggregating taxa were associated with different environments, so that their co-occurrence was not tested by Freilich et al. Of the remaining 17, 11 were co-occurring also in Freilich et al. study (65%). This is a very significant overlap, since the overlap expected by chance is $17 \times 49/2628 = 0.32$. Note that 38 out of 49 pairs co-occurring in their study were not significantly aggregated in ours, presumably because they are associated to the same environment and were filtered out by our null model. None of the segregating pairs coincided in the two studies. We conclude that our method is effective in filtering out pairs that aggregate because of environmental preferences, and that most of the aggregating pairs that it identifies agree with Freilich et al.'s method. On the contrary, segregations do not agree between the two methods (the overlap expected by chance is 0.1, and we find zero), perhaps because they are more difficult to detect.

Control network

To take into account possible biases caused by our computational procedures, we constructed a control network. We used as the starting point a random presence/absence matrix extracted with the probabilities computed with the null model for each combination of samples and taxa. We computed association scores for all pairs of taxa exactly as for the observed matrix and we assigned associations using thresholds lower than for the observed network ($T = 3.34$ instead of 5.75 for aggregation, and $T = 4.60$ instead of 5.75 for segregation) in such a way that the number of associations is the same for networks obtained from the two matrices. One should keep in mind that this control network is not a random network, since its construction produces correlations. For instance, since taxa appearing in many samples tend to co-occur with many other taxa, when we decrease the significance threshold they tend to form many aggregations, which produce aggregation propensity (see below).

Community structure

Association propensity. We investigate the community structure of the observed and the control network by measuring the propensity (see Methods) that two taxa associate given that they both associate with a third taxon k. This measure is analogous to the clustering coefficient, but it is clearer to interpret since it is negative if the association with k disfavors the association between i and j. There are two types of associations, aggregation (A) and segregation (S), and three conditioning associations: both i and j aggregate with k (AA), both segregate (SS), and one aggregates and the other segregates (AS). We obtain six propensities, which are reported in Table 2. Even the control network generates significant propensities, since taxa present in many samples tend to form many associations and produce positive propensities. However, propensities are much stronger for the observed network, suggesting that they provide non-trivial information on the community structure. The favored triangles are AAA and ASS, whereas it is disfavored that two segregating taxa aggregate with the same taxon (triangle AAS). These patterns are compatible both with the ecological and with the environmental interpretation of aggregations, and they suggest the existence of separate communities such that taxa of the same community aggregate between themselves and segregate from taxa in other communities.

Table 2 Association propensities for the observed and a random network

Propensity	Observed network	Control network
(A | AA)	3.59 ± 0.04	1.82 ± 0.10
(S | AA)	-1.83 ± 0.27	-0.71 ± 0.13
(A | AS)	-1.91 ± 0.12	-0.42 ± 0.12
(S | AS)	4.45 ± 0.09	1.32 ± 0.17
(A | SS)	3.24 ± 0.04	0.42 ± 0.14
(S | SS)	1.93 ± 0.11	0.62 ± 0.42

(A | AA) represents propensity of aggregation given two aggregations, and so on (see text).

Nestedness ν (see Methods) is a property related with the aggregation propensity that has been shown to be enhanced in mutualistic networks [29], although it may also arise from habitat filtering. Strongly nested pairs share many common aggregations, and they are more frequently observed in the observed than in the control network, see Additional file 1: Figure S5. The medians of the two distributions are different at the 1% significance level (Wilcoxon rank sum test).

Habitat filtering or ecological interactions?

Significant associations may be attributed either to ecological interactions or to habitat preferences. The null model reduces the second possibility by taking habitat preferences into account, since the taxon-specific parameter $p_i(A)$ vary for each subtype A of the environmental classification so that preference for the same subtype would not necessarily result in significant aggregation. Nevertheless, the environmental classification is necessarily coarse, and we cannot exclude that aggregations are due to habitat preferences that the null model fails to reproduce, such as for instance pH, oxygen or light. Disentangling environmental and ecological preferences is very difficult, since interactions in large natural communities of bacteria cannot be directly observed on a large scale. Therefore, in the following we examine indirect evidences that support one or the other interpretation.

Environmental and phylogenetic relatedness favor aggregation and disfavor segregation

Firstly, we examined the propensity (see Methods) between aggregation and shared habitat preferences. A positive propensity means that pairs of taxa that share the same habitat preference tend to aggregate more often than generic pairs of taxa or, conversely, that aggregated taxa tend to share habitat preferences. This relationship is expected even for a random presence-absence matrix. Therefore, we compared the observed aggregation network with the control network described above.

Figure 2 (top panel) shows the propensity for aggregation versus the environmental relatedness at the level of subtype, type and supertype. We consider a taxon associated with an environment if more than 50% and at least 3 of the samples in which it is observed belong to this environment. We distinguish three types of environmental relatedness for each level, in decreasing order of similarity: *Same*, if the two taxa are associated with the same environment, *Und*, if one or both of them are not associated with any specific environment, and *Diff* if they are associated with different environments. For instance, (Same, Diff, Und) means that the preference is the same at the supertype level, different at the type level and undefined at the subtype level. We represent in the plot only points for which there are at least 10 pairs, for instance no point is shown for same family and same order and (Same, Same, Diff) habitat preferences.

As expected, one can see from Figure 2 that environmentally related taxa have a strong propensity to aggregate. Nevertheless, in the control network (black curve) the maximum environmental relatedness (same subtype, type and supertype) does not produce significant propensity for aggregation, indicating that the null model is effective in reducing aggregations caused by environmental preferences at the subtype level. At the type and supertype level, small but significant propensities arise even in the control network. Similarly, different supertypes generate a small but negative propensity to aggregate.

These propensities are much stronger for the observed network (red curve) than for the control network, in particular taxa with the same habitat preferences at the supertype level are more prone to aggregate. The most parsimonious interpretation of this observation is that these aggregations are caused by habitat filtering, through environmental preferences that the null model does not take into account. Under this respect, habitat filtering is the preferred explanation for the two points corresponding to share subtype and shared type. Nevertheless, the possibility that some aggregations come from ecological interactions cannot be conclusively rejected either, as discussed in the following sections.

Furthermore, the aggregation propensity is significantly larger for pairs of taxa that are both environmentally and phylogenetically related. This relationship between aggregations and phylogeny goes beyond shared habitat preferences, since pairs of taxa belonging to the same order (blue curve) and family (orange curve) are prone to aggregate even in the absence of a common environmental preference. The propensity for aggregation increases with the phylogenetic relatedness (root, phylum, class, order and family) and the propensity for segregation decreases (see Figure 2, bottom panel), in agreement with the results by Chaffron et al. [17], and also reminiscent of the results by Tamames et al. [26], who found that the environments can be classified based on the affinities that different phyla have with different environments.

The propensity for segregation gives little information, since the number of significant segregations is small. Although we require that significantly segregating taxa co-occur in at least one subtype, we do not find any pair of segregating taxa with the same habitat preference at subtype level, which suggests that the data that we use cannot effectively identify competing taxa. However, most of the segregating pairs that we identify coexist in at least one sample, and the fact that their preferences are different can be also due to

Pascual-García *et al. BMC Microbiology* 2014, **14**:284
http://www.biomedcentral.com/1471-2180/14/284

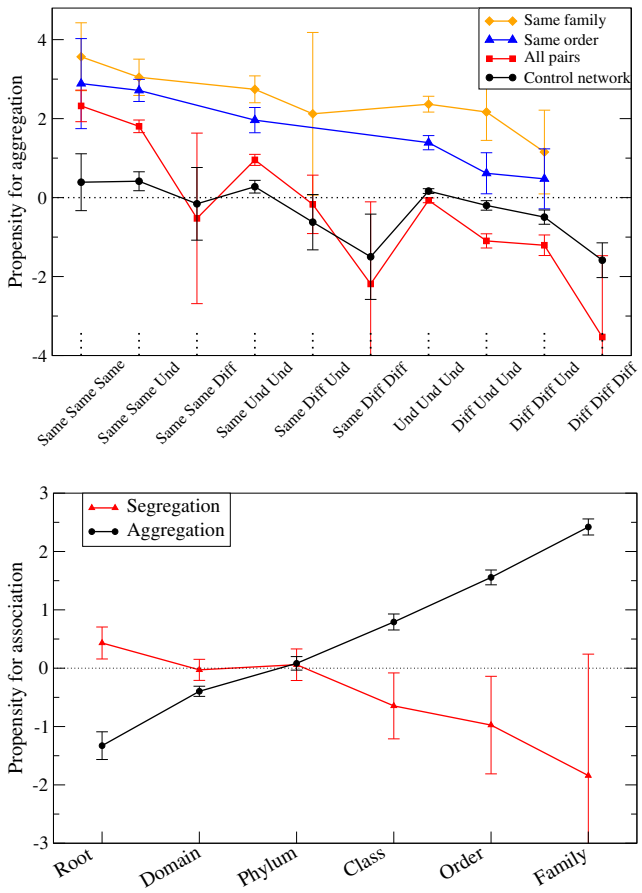


Figure 2 Propensity to association versus environmental relatedness (top) and versus phylogenetic relatedness (bottom). In the top figure the x axis labels environmental relatedness at the three levels of subtype, type and supertype. For each level, three values of relatedness are possible: 'Same' if the preferred habitat is the same, 'Und' if it is undetermined for one or both taxa, and 'Diff' if it is different.

the high threshold that we choose to assign habitat preferences.

Aggregated taxa co-occur in very different environments

To distinguishing habitat filtering from ecological interactions, we envisage two scenarios in which an association due to environmental preferences is not recognized by our null model. The first scenario is that aggregated taxa share a preference for a habitat that occurs in different environmental subtypes, such as nitrite rich habitats found in wastewater treatments and agricultural samples classified in different subtypes, so that the preference is underestimated by the null model. We would expect that this scenario is more likely if the habitat occurs in similar subtypes belonging to the same type (for instance, human gut and mouse gut), rather than in different supertypes (for instance, forest and hydrothermal). The second scenario is that the same sample contains many micro-habitats (for instance, the human gut hosts different environments that cannot be resolved in the most common experimental settings), and the apparent aggregation stems from specialization to different habitats found in the same sample. If this is the case, most of the samples where the taxa co-occur should contain the same micro-habitats, which is more likely if the samples belong to the same subtype or the same type (for instance, human and mouse gut), but not if they come from different supertypes (for instance, open sea and rhizosphere). In both cases, taxa that co-occur in similar samples are more likely to aggregate because of habitat filtering.

To test these scenarios, for each pair of significantly aggregated taxa we measured the number of different subtypes, types and supertypes to which the samples where they co-occur belong. This number was measured as the exponential of the Shannon entropy, $-\sum_i f_i \log f_i$, in order to reduce the impact of unfrequent environments. We found that 77% of the significantly aggregated pairs co-occur in samples from more than two different subtypes, 60% from more than two types, and 57% from more than one supertype. These data support the view that most aggregations cannot be explained by habitat preferences. The distribution of the number of different environments shared by each pair of significantly aggregated taxa is shown in Figure 3.

Cosmopolitan taxa are prone to aggregate

The study of Tamames et al. [26] found that cosmopolitanism, i.e. the fact that some taxa occurr in very diverse environments, is relatively common in the bacterial world, in particular if higher order taxonomic groups are considered. We set up to further investigate the relationship between cosmopolitanism and aggregations because of two reasons: first, since cosmopolitan taxa do not possess environmental specificity, they may allow distinguishing between habitat filtering and ecological interactions; second, this investigation may give hints on whether aggregations play a role in the cosmopolitanism of some bacterial taxa.

We measured taxa cosmopolitanism in two ways: (1) As environmental cosmopolitanism, i.e. the number of different environmental subtypes in which a taxon is present, and (2) As community cosmopolitanism, i.e. the number of different communities in which a taxon is present (see Eq.(1) in Methods). To investigate possible methodological artefacts, we compared the observed aggregation network and the control network.

The number of aggregations of a taxon is positively correlated with its cosmopolitanism both for the control and for the observed network, but in the latter case the correlation is much stronger ($r = 0.64$ instead of $r = 0.35$). If we normalize the number of aggregations dividing it by the number of samples in which the taxon is present, called prevalence, the relationship with cosmopolitanism remains positive for the observed network whereas it becomes negative for the control network (see Figure 4 for community cosmopolitanism and Additional file 1: Figure S4 for environmental cosmopolitanism). This qualitative difference suggests that the observed relation between aggregations and cosmopolitanism goes beyond the trivial effect that more common taxa are more likely to co-occur. Since cosmopolitan taxa do not present well-defined preferences, it seems unlikely that the excess aggregation is due to habitat filtering. For instance, Flavobacterium and Pseudomonas are present in 36 different subtypes such as Arctic, Mouse Gut, Food Treatment or Mines among others. The hypothesis that their cooccurrence is explained by habitat preferences would imply that these hypothetical preferred properties co-occur in such a wide variety of environments. A more economical hypothesis is that the excess of co-occurrence is explained by cooperative interactions. Another possible hypothesis is that two cosmopolitan have an indirect

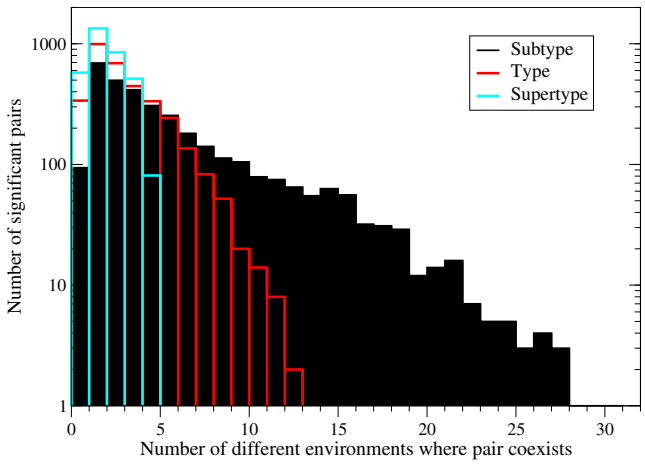


Figure 3 Distribution of the number of significantly aggregated pairs of taxa that coexist in *n* different subtypes, types and supertypes. The number of environments is computed as the exponential of the Shannon entropy. The maximum possible number of environments is 5 at supertype level, 20 at type level and 46 at subtype level.

Pascual-García *et al. BMC Microbiology* 2014, **14**:284
http://www.biomedcentral.com/1471-2180/14/284

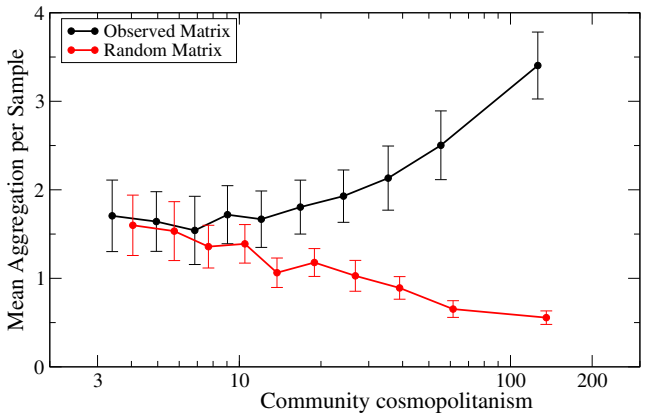


Figure 4 Number of aggregations of bacterial taxa divided by number of samples in which the taxon is present versus community cosmopolitanism Eq.(1). The black lines are obtained for the observed network, the red lines are obtained for a control network. Each point represents the average of all taxa in the same bin of cosmopolitanism.

relationship, due to the fact that there are specialist taxa that, if present, exclude both of them. Our data do not allow distinguishing between direct and indirect relationships, therefore we cannot judge how likely is this hypothesis.

Known cooperative pairs are found to aggregate

The hypothesis that some of the aggregations that we find are due to cooperative interactions can be tested examining pairs of taxa for which such interactions are known. They fall into three main cathegories: Syntrophy [30], in which one taxon is metabolically dependent on reactions carried out by a different taxon; Biofilms [31], in particular those formed by pathogenic bacteria, which cooperate to promote the chronic nature of the infection [32]; Mutualistic interactions with a shared host [33]. Many pairs of taxa for which there are hints of a cooperative relationship show significant aggregation, as described below.

An important example of syntrophy are methanogenic environments in which organic acids are degraded by syntrophic associations of acetogenic bacteria and methanogenic archaea. Hydrogen consumption by methanogens allows acetogenic bacteria to convert organic acids to acetate and hydrogen [34]. Consistently, we find significant aggregations between Acetobacterium and the methanogen archaea Methanolobus and Methanocalculus. In a similar context, an experimental study of methanogenesis from ethanol identified a three species mutualistic coculture with Desulfovibrio as the ethanol-degrading species producing acetic acid and hydrogen, which was converted to methane by a Methanobacterium sp. while the pH was maintained by the acetate-utilizing Methanosarcina mazei [35]. The two latter taxa show significant aggregation. In another study it was demonstrated that "the coexistence of two types of methanogens, i.e. hydrogenotrophic (Methanoculleus receptaculi) and acetoclastic (Methanosarcina thermophila) methanogens is necessary to respond successfully to perturbation and leads to stable process performance" [36]. These taxa are significantly aggregated. Similarly, the persistence of Pseudomonas putida in an environment with benzyl alcohol as the sole carbon source is dependent on the presence of Acinetobacter. Experimental evolution of this community in a biofilm lead to establish a structured community in which interactions between the two species evolved, enhancing productivity and stability [37]. This is the strongest association that we detect. Nitrosomonas and Nitrospira are two significantly aggregated nitrifying bacteria frequently found in wastewater treatment plants, where they oxidize ammonia and nitrite, respectively [38]. In this case, the aggregation seems to result from habitat preferences and specialization rather than syntrophy.

Another very interesting example of syntrophy are chemolithotrophic bacterial communities that oxidize iron and sulfur leading to the formation of metal-rich acidic water. In these peculiar ecosystems, whose best known example is the acidic river Rio Tinto in Spain, the energetic cycle is characterized by several types of bacteria that act cooperatively [39,40]: sulfur- and iron-oxidizing bacteria, such as Acidithiobacillus ferrooxidans and Leptospirillum ferrooxidans, Acidiphilium, which removes organic compounds toxic for Leptospirillum and reduces iron even in the presence of oxygen, and Acidithiobacillus spp. and members of the Acidimicrobiacea family, which can facultatively reduce iron under anoxic conditions. These taxa have large aggregation scores.

Other interesting examples are related to the ability of bacteria to adapt to environmental conditions that change due to human activity. For instance, Sphingomonas sp. TFEE and Burkholderia sp. MN1, isolated in soils treated with the pesticide fenitrothion, were shown to be able to degrade the pesticide jointly but not alone [41], and they aggregate.

Another important class of examples concern biofilms of pathogenic bacteria. Pseudomonas aeruginosa and Burkholderia, the main pathogens in cystic fibrosis, form mixed biofilms in the lungs of patients. They have frequently exchange genetic material and communicate through a common quorum-sensing system [42]. Their aggregation score is lower than the conservative threshold that we adopt, but it is large ($Z = 3.8$). Other associations between pathogenic bacteria are frequently observed in chronic wounds biofilms, in which bacteria cooperate to promote the chronic nature of the infection [32]. The seven taxa most frequently observed in these biofilms break down in two communities, one in which Pseudomonas and Enterobacter are significantly aggregated with Serratia although they are not aggregated between themselves ($Z = 1.5$) and another one in which Staphylococcus and Stenotrophomonas are significantly aggregated with Finegoldia and marginally aggregated between themselves ($Z = 4.9$), while Peptoniphilus is marginally aggregated with both Finegoldia and Streptococcus.

Finally, an important type of indirect interactions are mutualistic interactions with a common host. Such aggregations can be viewed both as an example of habitat filtering and as an indirect cooperative interaction over evolutionary time scales mediated by the host. Gut and root microbiota constitute the most studied examples and present interesting common features [33]. For instance, Rhizobium tropici and Devosia form a symbiosis with the same aquatic legume host, they have been shown to have interchanged symbiotic genes by horizontal transfer [43] and they are significantly aggregated. Photobacterium and Vibrio, two taxa present in the light organs of some fishs [44], are significantly aggregated.

Summarizing, excluding nitrifying bacteria that are a likely example of habitat filtering, we have examined 26 pairs of experimentally known associations that have different feature that suggest a synergistic relationship, finding that 18 of them (69%) show significant aggregations. Note that two more pairs have large aggregation scores but smaller than our chosen threshold, which indicates that the threshold that we have chosen is strict.

Network analysis

We now concentrate our attention on a portion of the aggregation and segregation network (the full network of 1187 taxa is too large to be visualized) related with animal Guts, given the major interest on understanding the ecological determinants underlying the assemblage of Human Gut communities. Indeed, it has been suggested that a better understanding of Gut communities may be achieved considering samples from environments different than Human Gut in order to identify the facultative or obligatory nature of the different taxa [4]. With this motivation, we selected a subnetwork containing taxa that has been observed not just in the Human Gut, but also in other guts such as Cattle or Mouse (see Methods). In addition, and for the sake of comparison, we also selected two more subnetworks related with the Saline and Plants environments (see Methods), which are shown in Additional file 1: Figure S8 and briefly commented in Additional file 1: Supplementary text S2.

The gut related network shown in Figure 5 comprises 5 subtypes, and we require taxa to be present in at least 3 of them. This condition selects 141 taxa, which are not necessarily preferentially associated with the gut environment. For most of them the association is strong, since 87% are found in at least 5 gut-related samples and 58% in at least 10 samples. Note that, selecting taxa that are observed in at least three Gut subtypes, we underscore ecological relations that may prevail in the Gut independently of the host at the expense of losing some taxa only found in the Human Gut. The 141 selected taxa are related through 468 aggregations and 146 segregations that are computed from the entire data set, so they may co-occur in environments different from the Gut. We can visually distinguish two large groups of strongly aggregated taxa (solid lines) and an intermediate group that links them through transitive aggregations. One of the large groups is constituted by taxa preferentially found in the supertype "host" (red circles), and the other is constituted by generalist taxa (white circles: no supertype accounts for more than 50% of the samples). The two groups are mostly related through segregations (dashed lines) in Figure 5. To quantitatively confirm this structure, we have analysed the modularity of the aggregation network with the modularity algorithm proposed in Ref. [45] implemented in the program Gephi [46]. This algorithm subdivided the Gut related network into five communities: the two large groups clearly seen in the main figure, two intermediate communities connected to both of them and between themselves, and a small community (Enterobacter, Citrobacter and Klebsiella) only connected to the generalist community. They are represented in Additional file 1: Figure S7. There are two possible interpretations of this pattern: associations may be mainly attributed to habitat filtering, which would explain why generalist taxa tend to segregate from gut taxa. Alternatively, associations may be attributed to ecological interactions, and in this case the observed pattern would suggest antagonistic interactions between the gut community and opportunistic invaders.

Pascual-García *et al. BMC Microbiology* 2014, **14**:284
http://www.biomedcentral.com/1471-2180/14/284

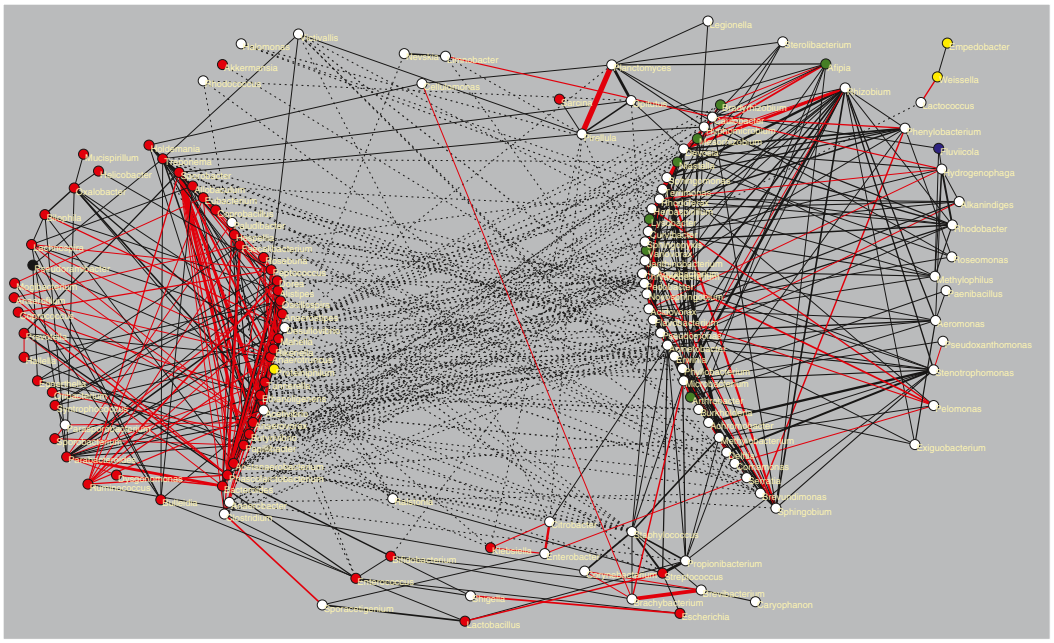


Figure 5 Networks obtained for taxa present in a group of samples related with animal guts. Solid lines represent aggregations, dashed lines represent segregations. Circles represent taxa, coloured according to the supertype to which at least 50% of the samples belong. (red = host, green = terrestrial, blue = aquatic, magenta = thermal, yellow = other, white = undefined). Red lines connect taxa belonging to the same family. The graphs have been plotted with the program Pajek [47].

Finally, we have also compared the network in Figure 5 with independent data obtained sampling the gut microbiota at 5 time points during the first year of life of 13 infants [48]. Interestingly, members of the host-related group appear and remain until the last time point, whereas generalist taxa are intermittently observed at different time points, supporting their interpretation as opportunistic invaders.

Discussion

Bacterial communities can be very diverse. Hundreds of species have been found in animal gut [49], vagina, mouth, and other organs. However, the global extent of bacterial diversity is currently debated. Estimates based on species-area curves [50] and on the depth of bacterial divisions on the rRNA tree of life [51] anticipated a huge number of bacterial species, but extrapolations from higher taxonomic levels were much lower than expected [52,53]. This discrepancy is due at least in part to the fact that the definition of bacterial species is artificial [54], and very important differences in gene content may exist between individuals classified as the same species, so that the concept of ecotype may be more relevant for bacteria than the rRNA-based species definition [55]. Unfortunately, the resolution of the data does not allow us to address the ecotype level, and we had to conduct this study at the somehow artificial genus level (98% identity in rRNA). It is therefore reassuring that a recent study argued that the bacterial genus level is ecologically coherent [56], supporting the approach undertaken here.

We performed a large scale survey of significant aggregations and segregations of bacterial taxa, adopting a maximum likelihood null model that takes into account environmental preferences at the environmental subtype level. We found a large number of significant aggregations, which may be attributed either to shared habitat preferences that are not taken into account by the null model or to cooperative ecological interactions. Both explanations are at least partially valid. The null model almost eliminates the aggregation propensity of pairs of taxa that share habitat preferences at the subtype level, but not at the type and supertype level (see Figure 2). On the other hand, 18 out of 26 (69%) known examples of cooperative interactions are recovered by our analysis as significant aggregations, and some others have large scores that fall below our chosen threshold, suggesting that the threshold that we use is strict.

In order to quantitatively assess the two kinds of explanation, we examined two main mechanisms that may lead to environmentally driven aggregations: (1) The preferred habitat may be distributed between several environmental subtypes so that the null model does not detect this preference; (2) The same sample may contain several microniches, so that the taxa aggregation is only apparent. Both mechanisms are plausible if the aggregated taxa co-occur in very similar environments. Therefore, we conservatively attribute to habitat filtering the aggregation or pairs of taxa with shared environmental preferences. However, we found that most significantly aggregated pairs coexist in more than two different subtypes (77%) and types (60%) and more than one supertype (57%) of the environmental classification, and in these cases habitat filtering appears a less likely explanation of the aggregation.

Cosmopolitanism offers another indirect evidence of the mechanism underlying aggregation. Cosmopolitan taxa, which live in very diverse environmental conditions and communities, present many more aggregations than specialist taxa. The number of aggregations increases with cosmopolitanism faster in the real network than in the network that we use to control methodological artefacts. If shared habitat preferences are the main source of aggregation, we would expect fewer aggregations for cosmopolitan taxa, which lack well defined preferences. Thus this result is consistent with the view that many aggregations are due to ecological interactions.

Cosmopolitanism is apparently at odds with the view that biodiversity is maintained by distinct ecological niches that avoid the competitive exclusion of species. The fact that cosmopolitan taxa tend to aggregate suggests the interesting possibility that cooperative interactions may favor the remarkable cosmopolitanism of some bacterial taxa. Of course, this hypothesis needs to be tested experimentally. We hope that the statistical signal presented here will stimulate such a test. To this end, we provide here examples of groups of cosmopolitan taxa that show strong aggregations between themselves in many diverse environments and can be interesting candidates for experimental studies. (1) The four taxa Pseudomonas, Acinetobacter, Stenotrophomonas and Sphingobium strongly aggregate; interestingly, cooperative interactions between Pseudomonas and Acinetobacter have been observed in experimental evolution [37]. (2) The group of the plant-associated taxa Rhizobium, Arthrobacter, Sphingomonas and Nocardioides, which also co-occur within several scientific papers; (3) The group Devosia, Rhizobium, Lysobacter and Sphingopyxis, the first two associated with plant symbiosis; (4) Streptococcus, Staphylococcus and Propionibacterium, associated with many infectious processes; (4) The aquatic genera Flavobacterium, Acidovorax, Rhodoferax and Polaromonas; (5) The soil bacteria Bradyrhizobium, Rhodoplanes, Conexibacter, Gemmata, Isosphaera and Stella. Some of these taxa, like Nocardioides, Conexibacter, Rhizobium or Byssimonas, are highly promiscuous, forming more than 30 aggregations each.

The aggregation network identified in this work has a marked community structure, in particular it is significantly clustered and nested. The triangles where all three taxa aggregate (AAA) and those where two aggregated taxa segregate from another one (ASS) are statistically favored, suggesting the existence of different communities characterized by intra-community aggregation and inter-community segregation. These data are compatible both with ecological interactions and habitat filtering as the basis of the community structure. Similarly, pairs of aggregated taxa share more common aggregations than expected at random (nestedness). Although habitat filtering can explain this property, it is interesting to note that nestedness is also observed in mutualistic networks of plants and pollinators [29], and it has been suggested to reduce effective competition and favor structural stability and biodiversity [57].

When we compared significant aggregations and segregations, taking care that the comparison is performed at equal false positive rate, we obtained the surprising result that aggregations are more frequent than segregations in the bacterial world. This comparison may be biased by the fact that sparse binary data are little effective at detecting segregation, or that the broad phylogenetic range of bacteria genera makes it difficult to detect competitive exclusion, as it has been recently suggested [58]. Moreover, many of the aggregations that we find must be attributed to habitat filtering. Nevertheless, in most cases this explanation does not appear as the most likely, which opens the way to the surprising ecological interpretation that cooperative interactions between bacteria may be very widespread.

This interpretation is worth investigation. Several other works studied bacterial communities on a large scale, emphasizing the role of either habitat filtering or competitive exclusion or cooperative interactions.

Chaffron et al. [17] performed a study very similar to ours, detecting numerous significant aggregations from bacterial co-occurrence in environmental samples. The main differences with respect to our study are that their null model does not take into account environmental preferences, so that these aggregations must be conservatively attributed to habitat filtering, and segregation was not assessed. The study of Faust et al. examined the microbiome of several body sites with high spatial resolution, finding a comparable number of positive and negative associations within body site [24], but another recent study of the human gut microbiome did not find significant negative associations [23]. Arumugam et al. provided evidence for

Pascual-García *et al. BMC Microbiology* 2014, **14**:284
http://www.biomedcentral.com/1471-2180/14/284

the existence of three distinct types of composition of the gut microbiome called enterotypes [18]. However, even in distinct enterotypes one does not observe complete exclusion of any common bacterial taxon, except *Prevotella*.

In an interesting study, Freilich et al. studied the ability of pairs of bacterial species to interact competitively or synergistically by predicting their metabolic growth in isolation and in the presence of the other species on different media [28]. Interestingly, in most cases both outcomes are possible depending on the growth medium. They also performed a co-occurrence analysis similar to ours, but with the important difference that it did not adopt any null model and it did not attempt to eliminate pairs that are associated due to common environmental preferences, but instead focused on such pairs. Nevertheless, the qualitative incidence of aggregations and segregations is similar to the one that we found, and 65% of the significantly aggregated genera that we could compare were also co-occurring in their analysis. The co-coccurrence analysis shows that pairs of taxa that are ecologically related through co-occurrence or exclusion tend to have larger competition and cooperation scores than unrelated taxa, which supports the idea that ecological interactions lie behind many aggregation and segregation events. Horner-Devine and coworkers examined 86 matrices of presence-absence of bacterial taxa and computed their C-score [59], finding that all but one significant C-scores were positive, which suggests prevalence of segregation over aggregation [16]. However, the C-score is a global measure that may be positive even in the absence of significant segregations if the majority of pairs co-occur less than expected, which is very likely due to the discretization of presence-absence matrices. Gotelli and Ulrich found that the C-score may be highly significant even if the number of significantly aggregated pairs is larger than the number of significantly segregated pairs [60]. Levy and Borenstein recently studied through metabolic models the complementarity and competition of pairs of bacterial taxa, predicting that taxa that co-occur in the gut microbioma tend to compete more than those that exclude themselves [61]. This prediction suggests that microbiome assembly is dominated by habitat filtering. We also consider habitat filtering as the most economic explanation for the aggregation of taxa that co-occur in one or few environmental subtype, but not for those that co-occur in a wide range of environments. One should be cautious in using metabolic predictions, since the difference between metabolic competition and syntrophy may depend on a small number of key enzymes: The introduction of just one engineered gene in strains of the same bacterial species can turn their competition into a strong synergistic interaction [62]. Moreover, using metabolic predictions it has been shown that it is possible to identify putative media that induce commensalism or mutualism for all the examined pairs of seven bacterial species[63].

There is an increasing number of experiments that attempt to investigate ecological interactions between bacteria on a large scale. A recent experiment measured the overall respiration of assemblies of species and attributed competitive interactions to assemblies in which the total respiration was less than the sum of the respiration of individual species, concluding that competition, not cooperation, dominates interactions among culturable bacteria [64]. However, respiration does not measure biomass production but production plus dissipation, which is expected to increase in the absence of ecological partners [65]. In contrast, another recent experiment found the seemingly opposite result that bacterial taxa have lower growth rate when assayed in the absence of other taxa in their natural community [66], suggesting that cooperative interactions are common. Moreover, a recent experiment found that environmental bacteria are organized into socially cohesive units in which cooperation mediated by antibiotic resistance tends to occur within each ecologically defined population while antibiotic-mediated antagonism occurs between populations [67].

In addition, it is relatively easy to set up experiments in which cooperative interactions evolve or are maintained [68-72], or to find growth media compositions such that the two species are predicted to grow synergistically [28]. A recent work has realized synthetic communities of engineered strains of the same bacterial species linked through the metabolic exchange of amino acids, finding that biosynthetically costly amino acids tend to promote strong cooperative interactions and presenting genomic evidence that suggests that amino acid crossfeeding and synergistic growth are common in bacteria [62].

Last, we discuss the interesting observation that phylogenetically related taxa have large aggregation propensity. This result was also found in Ref. [17,61], where it was attributed to habitat filtering. Nevertheless, this tendency exist also for cosmopolitan bacteria and for pairs that co-occur in many different environments, which suggests that some of these aggregations may be due to cooperative interactions. This hypothesis is puzzling. Since closely related taxa are expected to have large metabolic overlap and to compete strongly, as predicted by the metabolic models of Ref. [61], specialization into different niches or physical separation as in allopatric speciation may be expected to be a likely outcome of a speciation event, leading to segregation between related taxa, which is the contrary of what we observe here. This interpretation is consistent with the recent experiment by Mee et al., who turned strains of the same bacterial species from competitors to cooperators by engineering metabolic dependencies[62].

The recently proposed Black Queen Hypothesis [73] postulates a process in which the evolutionary loss of a gene whose product leaks out of other cells is selectively advantageous for the acceptor strain, which looses the gene and reduces its genome, and neutral for the donor strain, which disposes the gene without additional costs. This model has been proposed as a general mechanism for the establishment of cooperative bacterial communities [74], and its paradigmatic example is thought to be the evolution of genome reduction in several strains of the marine cyanobacteria Prochlorococcus [75].

The observation that phylogenetically related taxa are prone to aggregate may suggest that cooperative interactions played a role in their differentiation. A possible scenario, consistent with the Black Queen Hypothesis and the experiment of Mee et al., who turned strains of the same bacterial species from competitors to cooperators by engineering metabolic dependencies [62], is that one strain lost some genes not needed in its new dominant environment and established an enviroment-dependent metabolic dependency on a sister strain that disposes the products of these genes. This scenario may be testable. In the absence of a direct test, it is just a speculation, and the interpretation that the aggregation between related taxa is mainly due to habitat filtering should be preferred as more economic.

Conclusions

In conclusion, our results show that aggregations are frequent in the bacterial world, and they occur more frequently for cosmopolitan taxa and for phylogenetically related taxa. Our data support the view that a large number of these aggregations may be due to cooperative interactions. 57% of the aggregations occur in at least two different supertypes, and in our view they are more likely explained by cooperative interactions than habitat filtering, although the latter cannot be ruled out and indirect interactions with a third taxon can offer another possible explanation. Aggregations are particularly common for cosmopolitan taxa that are found in very different environments and for phylogenetically related taxa, which leads us to conjecture that cooperative interactions may be key for the remarkable cosmopolitanism of some bacterial taxa, and they may influence the mechanisms of bacterial differentiation.

Methods

Data set

The taxa presence-absence matrix was derived from the data presented in Ref. [26]. Briefly, 3,502 samples of 16S rDNA sequencing experiments were classified into environmental subtypes, types and supertypes and 1187 taxa were identified from the 16S rDNA sequence clustered at 98% sequence identity. Restricted samples, analysed with specific primers with the objective of studying the presence and/or abundance of particular taxa, were identified either from the presence of taxonomic names in the title of the article or identifying samples that contain a single taxon and eliminated from the data set, leaving us with 2322 samples.

Null model

We implemented the null model proposed in [27], summarized here for completeness. Our data consist of N taxa $i = 1 \dots N$ observed at M locations $a = 1 \dots M$, stored in the binary presence-absence matrix $X_{ia} \in \{0, 1\}$. We want to determine probabilities π_{ia} that generate random presence-absence matrices $\tilde{X}_{ia}$ as similar as possible to the observed one under the assumption that species do not interact and all $\tilde{X}_{ia} \in \{0, 1\}$ are independent. We assume that there is no preferential association between taxa and locations, an assumption that we will relax later.

We parametrize $\pi_{ia} = f(p_i q_a)$ so that the probabilities depend on N taxon-specific parameters p_i and M location-specific parameters q_a. Gilpin and Diamond [76] proposed the ansatz $\pi_{ia} = p_i q_a$ and determined p_i and q_a such that the mean of the sum of rows and columns is the same in random matrices as in the observed one. However, as they noted themselves, their model can give probabilities $\pi_{ia} \geq 1$. To avoid this problem, Navarro-Alberto and Manly proposed the ansatz $\pi_{ia} = 1 - \exp(-p_i q_a)$, justified assuming Poisson distributed species abundances, and determined the parameters that maximize the likelihood of the observed matrix given the model. The resulting log-likelihood function is $\mathcal{L} = \sum_{ia} \left[X_{ia} \log(\pi_{ia}) + (1 - X_{ia}) \log(1 - \pi_{ia}) \right]$. Maximizing this function, we obtain $N + M$ equations that we solve with a globally convergent Newton method with analytically computed gradients.

An important drawback of this model is the assumption that taxa do not have habitat preferences. We relax this assumption grouping locations into environmental subtypes and allowing the taxon-specific parameters $p_i(A)$ to depend on the subtype A to which the sample belong. We then solve the maximum likelihood equations separately for samples of each subtype A. If taxon i is never seen in subtype A, then $p_i(A) = 0$ and $\pi_{ia} = 0$ for all $a \in A$.

Association scores

The null model allows us to iteratively compute the probability that two taxa i and j co-occur at n locations over m, $P_{ij}(n|m)$:

$$\begin{aligned} P_{ij}(n|m) = {} & P_{ij}(n|m-1)(1 - \pi_{im}\pi_{jm}) \\ & + P_{ij}(n-1|m-1)(\pi_{im}\pi_{jm}) \end{aligned}$$

This equation, with initial conditions $P_{ij}(0|0) = 1$ and $P_{ij}(0|1) = 0$, yields the probability $P_{ij}(n|M)$ that the two taxa co-occur at n over M samples under the null model.

We then define the taxon aggregation (TA) and the taxon segregation (TS) scores as

$$S_{ij}^{\mathrm{TA}} = -\log\left(P_{ij}(n \geq n_{ij}|M)\right)$$
$$S_{ij}^{\mathrm{TS}} = -\log\left(P_{ij}(n \leq n_{ij}|M)\right)$$

where n_{ij} is the observed number of co-occurrences. Sample aggregation (SA) and segregation (SS) are defined in a similar way from the probability that two samples share n taxa. These scores are correlated with the number of samples in which individual taxa are present. To eliminate this correlation, we transform them into Z scores as follows. We extract 100 random matrices with the null model of the observed matrix, we compute their null model and, through it, we compute the scores S_{ij} for all pairs in the random matrix. Finally, we obtain mean and standard deviation of the observed S_{ij} over the random matrices, and we normalize the observed score subtracting the mean and dividing by the standard deviation.

Thresholds

In order to choose the significance threshold in an objective way, we estimate the false positive rate FPR (ratio between false positives and total number of pairs), and the positive predictive value PPV (true positives divided by total positives) by generating random association networks with the null model. Namely, we extract a random presence-absence matrix, determine its associated null model and compute aggregation and segregation scores for all pairs. The associations detected in the random network are considered as false positives, and their number is recorded versus the threshold.

Cosmopolitanism

The environmental cosmopolitanism of a taxon is the number of different environmental subtypes in which it is present, according to the hierarchical classification of Tamames et al. [26]. The community cosmopolitanism is defined as the number of samples in which the taxon is present counting only samples with significantly different communities. We adopt for such a purpose the sample aggregation score $S^{\mathrm{SA}})_{ab}$ that characterizes pairs of samples *ab* that contain more common species than expected by chance, defined similarly as the taxa aggregation score $S^{\mathrm{TA}})_{ij}$. We perform a similar analysis to choose the significance threshold $S_0^{\mathrm{SA}} = 4.92$ such that the PPV is 0.96. The community cosmopolitanism of a taxon i is defined by counting all pairs of samples in which the taxon is present that are below the significance threshold and dividing by all the samples in which the taxon is present:

$$(\mathrm{Comm.Cosm.})_i = 1 + \frac{2\sum_{a<b} X_{ia}X_{ib}\vartheta\left(S_0^{\mathrm{SA}} - S_{ab}^{\mathrm{SA}}\right)}{\sum_a X_{ia}} \quad (1)$$

The sum in the numerator runs over all pairs of samples where taxon i is present, and the theta function selects only significantly different pairs ($S_{ab}^{\mathrm{SA}} < S_0^{\mathrm{SA}}$). Eq.(1) equals one if all of the communities in which taxon i is present are significantly similar, and it equals $m_i = \sum_a X_{ia}$ if they are all different.

Association between taxa and environments

We associate a taxon with its favored environment at subtype, type or supertype level if more than 50%, and at least 3 of the samples where the taxon is found belong to that environment.

With these criteria, we could assign the dominant environment of 10% of the taxa at subtype level, 30% at type level and 51% at supertype level.

Propensity

The propensity that two random variables A and B assume the values a and b is defined as the logarithm of the ratio between the conditional probability of a given b and the probability without any condition: $\mathrm{Prop}(a,b) = \log\left[\mathrm{P}\{A=a|B=b\}/\mathrm{P}\{A=a\}\right] = \log\mathrm{P}\{A=a,B=b\} - \log\mathrm{P}\{A=a\} - \log\mathrm{P}\{B=b\}$. The propensity is symmetric exchanging a and b, it is positive when property b favors a or the other way round, and negative if the contrary holds.

Nestedness

In analogy with the definition in [57], we define the nestedness of two nodes i and j in a network with adjacency matrix A_{ij} as the fraction of links that they share:

$$\nu_{ij} = \frac{\sum_k A_{ik}A_{jk}}{\sqrt{\sum_k A_{ik}\sum_k A_{jk}}}\,. \quad (2)$$

The nestedness is one if i and j share all of their links, which implies that the clustering coefficient is also one.

Ethics

The research conducted in this paper did not require ethical approval, since it used previously published data.

Additional file

Additional file 1: Supplementary material. Figure S1: Number of samples and number of taxa present in each subtype of the environmental classification of Ref. [26]. **Figure S2**: Distribution of the aggregation and segregation scores for the observed matrix and a random realization. **Figure S3**: Predicted aggregations and segregations as a function of the false positive rate. **Figure S4**: Environmental cosmopolitanism versus the normalized number of aggregations for the observed matrix and for a random matrix. **Figure S5**: Distribution of nestedness between pairs of taxa for the observed matrix and for a random matrix. **Text S1**: describing the clustering of environmental subtypes. **Figure S6**: Hierarchical clustering of the environmental subtypes of Ref. [26]. **Text S2**: Description of networks restricted to subclusters of environmental subtypes related to Plants and Marine. **Figure S7**: Figures for the above networks. **Table S1**: Properties of nets represented in **Figure S8**. **Figure S8**: Propensity to share

environmental preferences conditioned to the phylogenetic relatedness. **File Aggregations.txt**: List of 3362 significant aggregations of bacterial taxa and their properties (text file). **File Segregations.txt**: List of 632 significant segregations of bacterial taxa and their properties (text file).

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
All authors conceived this study. JT provided the data on which the study was based. APG and UB designed the null model, the association scores and the statistical tests and performed the corresponding analytic calculations. APG performed the numerical computations and analyzed the results, with the help of UB and JT. All authors wrote and approved the final manuscript.

Acknowledgments
We gratefully acknowledge interesting discussions with Daniel Aguirre. APG acknowledges hospitality in the Andrés Moya lab. at Centro Superior de Salud Pública (CSISP) in Valencia, and would like to thank Juanjo Abellán, Pilar Francino, Yvone Vallés and Ana Durban for helpful discussions. This work was supported by the Spanish Ministry of Economy and Competitiveness (FPI grant BES-2009-013072 to APG and grants BFU2011-24595 and BFU2012-40020 to UB) and by the Comunidad de Madrid (Amarauto program to UB). Research at the CBMSO is facilitated by the Fundación Ramón Areces. We acknowledge support of the publication fee by the CSIC Open Access Publication Support Initiative through its Unit of Information Resources for Research (URICI).

Author details
[1]Centro de Biología Molecular Severo Ochoa (CSIC-UAM), c. Nicolás Cabrera 1, campus UAM, E-28049 Madrid, Spain. [2]Centro Nacional de Biotecnologí a (CSIC) c. Darwin 3, campus UAM, E-28049 Madrid, Spain.

Received: 23 June 2014 Accepted: 4 November 2014
Published online: 04 December 2014

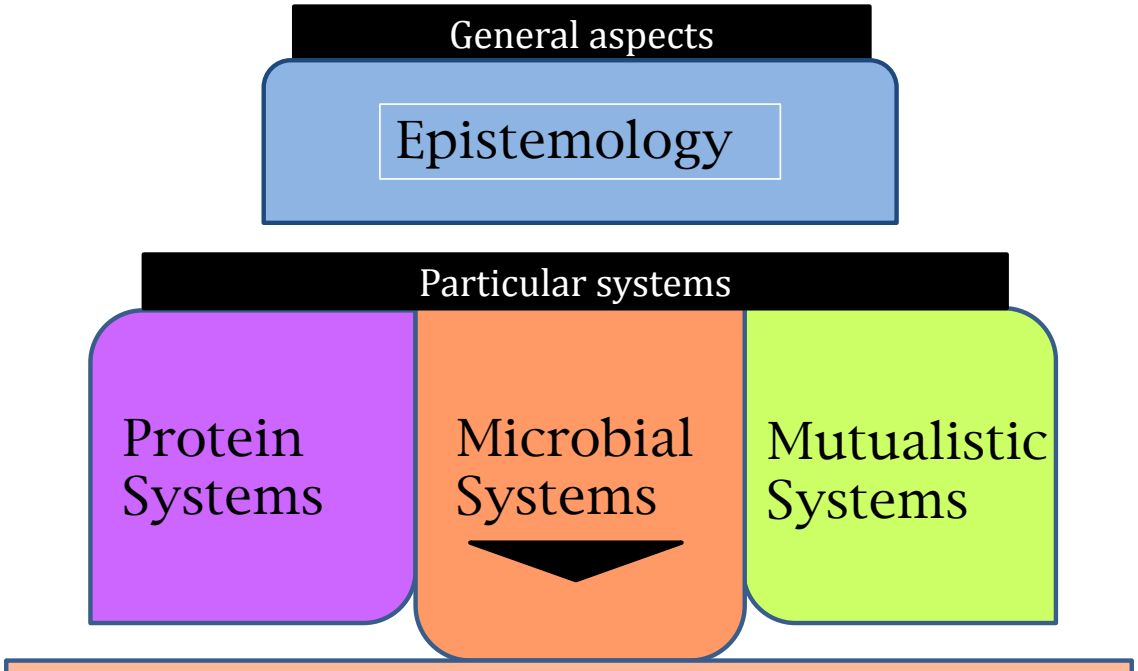
References

1. Hutchinson GE: **Homage to Santa Rosalia or why are there so many kinds of animals?** *The American Naturalist* 1959, **93:**145-159.
2. McFall-Ngai M, Hadfield MG, Bosch TC, Carey HV, Domazet-LoÅąo T, Douglas AE, Dubilier N, Eberl G, Fukami T, Gilbert SF, Hentschel U, King N, Kjelleberg S, Knoll AH, Kremer N, Mazmanian SK, Metcalf JL, Nealson K, Pierce NE, Rawls JF, Reid A, Ruby EG, Rumpho M, Sanders JG, Tautz D, Wernegreen JJ: **Animals in a bacterial world, a new imperative for the life sciences.** *Proc Natl Acad Sci USA* 2013, **110:**3229–3236.
3. Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, Sargent M, Gill SR, Nelson KE, Relman DA: **Diversity of the human intestinal microbial flora.** *Science* 2005, **308:**1635–1638.
4. Ley RE, Lozupone CA, Hamady M, Knight R, Gordon JI: **Worlds within worlds: evolution of the vertebrate gut microbiota.** *Nat Rev Microbiol* 2008, **10:**776–788.
5. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, Mende DR, Li J, Xu J, Li S, Li D, Cao J, Wang B, Liang H, Zheng H, Xie Y, Tap J, Lepage P, Bertalan M, Batto JM, Hansen T, Le Paslier D, Linneberg A, Nielsen HB, Pelletier E, Renault P, *et al.*: **Human gut microbial gene catalogue established by metagenomic sequencing.** *Nature* 2010, **464:**59–65.
6. The Human MicrobiomeProjectConsortium: **Structure, function and diversity of the healthy human microbiome.** *Nature* 2012, **486:**207–214.
7. Horner-Devine MC, Carney KM, Bohannan BJM: **An ecological perspective on bacterial biodiversity.** *Proc R Soc Lond B* 2003, **271:**113–122.
8. Horner-Devine MC, Lage M, Hughes JB, Bohannan BJM: **A taxa-area relationship for bacteria.** *Nature* 2004, **432:**750–753.
9. Green J, Bohannan BJM: **Spatial scaling of microbial biodiversity.** *Trends Ecol Evol* 2006, **21:**501–507.
10. Rocap G, Larimer FW, Lamerdin J, Malfatti S, Chain P, Ahlgren NA, Arellano A, Coleman M, Hauser L, Hess WR, Johnson ZI, Land M, Lindell D, Post AF, Regala W, Shah M, Shaw SL, Steglich C, Sullivan MB, Ting CS, Tolonen A, Webb EA, Zinser ER, Chisholm SW: **Genome divergence in two prochlorococcus ecotypes reflects oceanic niche differentiation.** *Nature* 2003, **424:**1042–1047.
11. Hopkinson CS, Sogin ML, Hobbie JE, Crump BC: **Microbial biogeography along an estuarine salinity gradient: combined influences of bacterial growth and residence time.** *Appl Environ Microbiol* 2004, **70:**1494–1505.
12. Ramette A, Tiedje JM: **Biogeography: An emerging cornerstone for understanding prokaryotic diversity, ecology and evolution.** *Microb Ecol* 2006, **53:**197–207.
13. Martiny JB, Bohannan BJ, Brown JH, Colwell RK, Fuhrman JA, Green JL, Horner-Devine MC, Kane M, Krumins JA, Kuske CR, Morin PJ, Naeem S, OvreÅès L, Reysenbach AL, Smith VH, Staley JT: **Microbial biogeography: putting microorganisms on the map.** *Nat Rev Microbiol* 2006, **4:**102–112.
14. Vellend M: **Conceptual synthesis in community ecology.** *Q Rev Biol* 2010, **85:**183–206.
15. Freilich S, Kreimer A, Meilijson I, Gophna U, Sharan R, Ruppin E: **The large-scale organization of the bacterial network of ecological co-occurrence interactions.** *Nucl Ac Res* 2010, **38:**3857–68.
16. Horner-Devine MC, Silver JM, Leibold MA, Bohannan BJ, Colwell RK, Fuhrman JA, Green JL, Kuske CR, Martiny JB, Muyzer G, OvreÅès L, Reysenbach AL, Smith VH: **A comparison of taxon co-occurrence patterns for macro- and microorganisms.** *Ecology* 2007, **88:**1345–1353.
17. Chaffron S, Rehrauer H, Pernthaler J, von Mering C: **A global network of coexisting microbes from environmental and whole-genome sequence data.** *Genome Res* 2010, **20:**947–959.
18. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, Fernandes GR, Tap J, Bruls T, Batto JM, Bertalan M, Borruel N, Casellas F, Fernandez L, Gautier L, Hansen T, Hattori M, Hayashi T, Kleerebezem M, Kurokawa K, Leclerc M, Levenez F, Manichanh C, Nielsen HB, Nielsen T, Pons N, Poulain J, Qin J, Sicheritz-Ponten T, Tims S, *et al.*: **Enteroptypes of the human gut microbiome.** *Nature* 2011, **473:**174–180.
19. Barberán A, Bates ST, Casamayor EO, Fierer N: **Using network analysis to explore co-occurrence patterns in soil microbial communities.** *ISME J* 2012, **6:**343–51.
20. Deng Y, Jiang YH, Yang Y, He Z, Luo F, Zhou J: **Molecular ecological network analyses.** *BMC Bioinformatics* 2012, **13:**11.
21. Gilbert JA, Steele JA, Caporaso JG, SteinbrÃijck L, Reeder J, Temperton B, Huse S, McHardy AC, Knight R, Joint I, Somerfield P, Fuhrman JA, Field D: **Defining seasonal marine microbial community dynamics.** *ISME J* 2012, **6:**298–308.
22. Larsen PE, Field D, Gilbert JA: **Predicting bacterial community assemblages using an artificial neural network apprach.** *Nat Methods* 2012, **9:**621–625.
23. Lozupone C, Faust K, Raes J, Faith JJ, Frank DN, Zaneveld J, Gordon JI, Knight R: **Identifying genomic and metabolic features that can underlie early successional and opportunistic lifestyles of human gut symbionts.** *Genome Res* 2012, **22:**1974–1984.
24. Faust K, Sathirapongsasuti JF, Izard J, Segata N, Gevers D, Raes J, Huttenhower C: **Microbial co-occurrence relationships in the human microbiome.** *PLoS Comp Biol* 2012, **8:**e1002606.
25. Faust K, Raes J: **Microbial interactions: from networks to models.** *Nat Rev Microbiol* 2012, **10:**538–550.
26. Tamames J, Abellán JJ, Pignatelli M, Camacho A, Moya A: **Environmental distribution of prokaryotic taxa.** *BMC Microbiol* 2010, **10:**85.
27. Navarro-Alberto JA, Manly BFJ: **Null model analyses of presence-absence matrices need a definition of independence.** *Popul Ecol* 2009, **51:**505–512.
28. Freilich S, Zarecki R, Eilam O, Segal ES, Henry CS, Kupiec M, Gophna U, Sharan R, Ruppin E: **Competitive and cooperative metabolic interactions in bacterial communities.** *Nat Commun* 2011, **2:**589.
29. Bascompte J, Jordano P, Melián CJ, Olesen JM: **The nested assembly of plantanimal mutualistic networks.** *Proc Natl Acad Sci USA* 2003, **100:**9383–9387.
30. Schink B, Stams AJM: **Syntrophism among prokaryotes.** In *The Prokaryotes*. Edited by Dworkin M. vol. 2. New York: Springer; 2006:309–335.
31. Boyle KE, Heilmann S, van Ditmarsch D, Xavier JB: **Exploiting social evolution in biofilms.** *Curr Opin Microbiol* 2013, **16**(2):207–212.
32. Dowd SE, Sun Y, Secor PR, Rhoads DD, Wolcott BM, James GA, Wolcott RD: **Survey of bacterial diversity in chronic wounds using**

pyrosequencing, DGGE, and full ribosome shotgun sequencing. *BMC Microbiol* 2008, **8:**43.
33. Ramirez-Puebla ST, Servin-Garcidueñas LE, Jimenez-Marin B, Bolaños LM, Rosenblueth M, Martinez J, Rogel MA, Ormeño-Orrillo E, Martinez-Romero E: **Gut and root microbiota commonalities.** *Appl Environ Microbiol* 2013, **79:**2–9.
34. Stams AJ, Plugge CM, de Bok FA, van Houten BH, Lens P, Dijkman H, Weijma J: **Metabolic interactions in methanogenic and sulfate-reducing bioreactors.** *Water Sci Technol* 2005, **52:**13–20.
35. Tatton MJ, Archer DB, Powell GE, Parker ML: **Methanogenesis from ethanol by defined mixed continuous cultures.** *Appl Environ Microbiol* 1989, **55:**440–445.
36. Lerm S, Kleyböcker A, Miethling-Graff R, Alawi M, Kasina M, Liebrich M, Würdemann H: **Archaeal community composition affects the function of anaerobic co-digesters in response to organic overload.** *Waste Manag* 2012, **32:**389–399.
37. Hansen SK, Rainey PB, Haagensen JA, Molin S: **Evolution of species interactions in a biofilm community.** *Nature* 2007, **445:**533–536.
38. Siripong S, Rittmann BE: **Diversity study of nitrifying bacteria in full-scale municipal wastewater treatment plants.** *Water Res* 2007, **41:**1110–1120.
39. González-Toril E, Llobet-Brossa E, Casamayor EO, Amann R, Amils R: **Microbial ecology of an extreme acidic environment, the Tinto River.** *Appl Environ Microbiol* 2003, **69:**48534865.
40. Santofimia E, Gonzàlez-Toril E, López-Pamo E, Gomariz M, Amils R, Aguilera A: **Microbial diversity and its relationship to physicochemical characteristics of the water in two extreme acidic pit lakes from the iberian pyrite belt (SW Spain).** *PLoS ONE* 2013, **8**(6):e66746.
41. Katsuyama C, Nakaoka S, Takeuchi Y, Tago K, Hayatsu M, Kato K: **Complementary cooperation between two syntrophic bacteria in pesticide degradation.** *J Theor Biol* 2009, **256:**644–654.
42. Eberl L, Tümmler B: **Pseudomonas aeruginosa and Burkholderia cepacia in cystic fibrosis: genome evolution, interactions and adaptation.** *Int J Med Microbiol* 2004, **294:**123–131.
43. Rivas R, Velázquez E, Willems A, Vizcaíno N, Subba-Rao NS, Mateos PF, Gillis M, Dazzo FB, Martínez-Molina E: **A new species of Devosia that forms a unique nitrogen-fixing root-nodule symbiosis with the aquatic legume Neptunia natans (L.f.) druce.** *Appl Environ Microbiol* 2002, **68:**5217–5222.
44. Douglas AE, Smith DC: **Are endosymbioses mutualistic?** *Trends Ecol Evol* 1989, **4:**350–352.
45. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E: **Fast unfolding of communities in large networks.** *J Stat Mech Theor Exp* 2008, **10:**1000.
46. Bastian M, Heymann S, Jacomy M: **Gephi: an open source software for exploring and manipulating networks.** In *International AAAI Conference on Weblogs and Social Media*; 2009. Association for the Advancement of Artificial Intelligence (www.aaai.org).
47. Batagelj V, Mrvar A: **Pajek: A Program for Large Network Analysis.** Home page: http://vlado.fmf.uni-lj.si/pub/networks/pajek/.
48. Vallès Y, Artacho A, Pascual-García A, Ferrús ML, Gosalbes MJ, Abellán JJ, Francino MP: **Successional patterns of community assembly and functional ecology during gut microbiota development in an infant birth cohort.** *PLoS Genetics* 2014, **10:**e1004406. Aceptado en PloS Genetics
49. Lozupone CA, Stombaugh JI, Gordon JI, Jansson JK, Knight R: **Diversity, stability and resilience of the human gut microbiota.** *Nature* 2012, **489:**220–230.
50. Curtis TP, Sloan WT, Scannell JW: **Estimating prokaryotic diversity and its limits.** *Proc Natl Acad Sci USA* 2002, **99:**10494–10499.
51. Dykhuizen DE: **Santa Rosalia revisited: Why are there so many species of bacteria?** *Antonie Leeuwenhaek* 1998, **73:**25–33.
52. Schloss PD, Handelsman J: **Status of the microbial census.** *Microbiol Mol Biol Rev* 2004, **68:**686–691.
53. Mora C, Tittensor DP, Adl S, Simpson AGB, Worm B: **How Many Species Are There on Earth and in the Ocean?** *PLoS Biol* 2011, **9**(8):e1001127.
54. Staley JT: **The bacterial species dilemma and the genomic-phylogenetic species concept.** *Philos Trans R Soc Lond B Biol Sci* 2006, **361:**1899–1909.
55. Cohan FM: **What are bacterial species?** *Annu Rev Microbiol* 2002, **56:**457–487.
56. Philippot L, Andersson SG, Battin TJ, Prosser JI, Schimel JP, Whitman WB, Hallin S: **The ecological coherence of high bacterial taxonomic ranks.** *Nat Rev Microbiol* 2010, **8:**523–529.
57. Bastolla U, Fortuna MA, Pascual-García A, Ferrera A, Luque B, Bascompte J: **The architecture of mutualistic networks minimizes competition and increases biodiversity.** *Nature* 2009, **458:**1018–1020.
58. Koeppel AF, Wu M: **Species matter: the role of competition in the assembly of congeneric bacteria.** *ISME J* 2014, **8:**531–540.
59. Stone L, Roberts A: **The checkerboard score and species distribution.** *Oecologia* 1990, **85:**74–79.
60. Gotelli NJ, Ulrich W: **The empirical bayes approach as a tool to identify non-random species associations.** *Oecologia* 2010, **162:**463–477.
61. Levy R, Borenstein E: **Metabolic modeling of species interaction in the human microbiome elucidates community-level assembly rules.** *Proc Natl Acad Sci USA* 2013, **110:**12804–12809.
62. Mee MT, Collins JJ, Church GM, Wang HH: **Syntrophic exchange in synthetic microbial communities.** *Proc Natl Acad Sci USA* 2014, **111:**E2149–E2156.
63. Klitgord N, Segré D: **Environments that induce synthetic microbial ecosystems.** *PLoS Comp Biol* 2010, **6:**e1001002.
64. Foster KR, Bell T: **Competition, not cooperation, dominates interactions among culturable microbial species.** *Curr Biol* 2012, **22:**1845–1850.
65. Carlson CA, del Giorgio PA, Herndl GJ: **Microbes and the dissipation of energy and respiration: from cells to ecosystems.** *Oceanography* 2007, **20**(2):89100.
66. Lawrence D, Fiegna F, Behrends V, Bundy JG, Phillimore AB, Bell T, Barraclough TG: **Species interactions alter evolutionary responses to a novel environment.** *PLoS Biol* 2012, **10:**e1001330.
67. Cordero OX, Wildschutte H, Kirkup B, Proehl S, Ngo L, Hussain F, Le Roux F, Mincer T, Polz MF: **Ecological populations of bacteria act as socially cohesive units of antibiotic production and resistance.** *Science* 2012, **337:**1228–1231.
68. Velicer GJ, Yu YN: **Evolution of novel cooperative swarming in the bacterium Myxococcus xanthus.** *Nature* 2003, **425:**75–78.
69. Harcombe W: **Novel cooperation experimentally evolved between species.** *Evolution* 2010, **64:**2166–2172.
70. Hosoda K, Suzuki S, Yamauchi Y, Shiroguchi Y, Kashiwagi A, Ono N, Mori K, Yomo T: **Cooperative adaptation to establishment of a synthetic bacterial mutualism.** *PLoS ONE* 2011, **6**(2):e17105.
71. Waite AJ, Shou W: **Adaptation to a new environment allows cooperators to purge cheaters stochastically.** *Proc Natl Acad Sci U S A* 2012, **109:**19079–19086.
72. Ferriere R, Bronstein JL, Rinaldi S, Law R, Gauduchon M: **Cheating and the evolutionary stability of mutualisms.** *Proc R Soc London. Series B: Biol Sci* 2002, **269**(1493):773–780.
73. Morris JJ, Lenski RE, Zinser ER: **The Black Queen Hypothesis: evolution of dependencies through adaptive gene loss.** *MBio* 2012, **3**(2):e00036–12. doi:10.1128/mBio.00036-12.
74. Sachs JL, Hollowell AC: **The origins of cooperative bacterial communities.** *MBio* 2012, **3**(3):e00099–12. doi:10.1128/mBio.00099-12.
75. Kettler GC, Martiny AC, Huang K, Zucker J, Coleman ML, Rodrigue S, Chen F, Lapidus A, Ferriera S, Johnson J, Steglich C, Church GM, Richardson P, Chisholm SW: **Patterns and implications of gene gain and loss in the evolution of Prochlorococcus.** *PLoS Genet* 2007, **3:**e231.
76. Gilpin ME, Diamond JM: **Factors contributing to non-randomness in species co-occurrences on islands.** *Oecologia* 1982, **52:**75–84.

doi:10.1186/s12866-014-0284-5
Cite this article as: Pascual-García *et al.*: Bacteria dialog with Santa Rosalia: Are aggregations of cosmopolitan bacteria mainly explained by habitat filtering or by ecological interactions? *BMC Microbiology* 2014 **14**:284.

3.2. Article [MIC-2]



1) Pascual-García, A., Tamames, J., and Bastolla, U. (2014) Bacteria dialog with Santa Rosalia: Are aggregations of cosmopolitan bacteria mainly explained by habitat filtering or by ecological interactions? *BMC microbiology*, *14*(1), 1-16.

2) Vallés Y., Artacho A., Pascual-García A., Ferrús M.L., Gosalbes M.J., Abellán J.J. and Francino M.P. (2014) Microbial succession in the gut: directional trends of taxonomic and functional change in a birth cohort of Spanish infants. *PLoS Genet.* 10(6): e1004406

OPEN ACCESS Freely available online

PLOS | GENETICS

Microbial Succession in the Gut: Directional Trends of Taxonomic and Functional Change in a Birth Cohort of Spanish Infants

Yvonne Vallès[1], Alejandro Artacho[1], Alberto Pascual-García[2], Maria Loreto Ferrús[1], María José Gosalbes[1,3], Juan José Abellán[1], M. Pilar Francino[1,4]*

1 Unidad Mixta de Investigación en Genómica y Salud, Fundación para el Fomento de la Investigación Sanitaria y Biomédica de la Comunitat Valenciana (FISABIO)-Salud Pública/Institut Cavanilles de Biodiversitat i Biologia Evolutiva (Universitat de València), València, Spain, **2** Centro de Biología Molecular "Severo Ochoa" (CSIC-Universidad Autónoma de Madrid), Madrid, Spain, **3** CIBER en Epidemiología y Salud Pública (CIBERESP), Spain, **4** School of Natural Sciences, University of California Merced, Merced, California, United States of America

Abstract

In spite of its major impact on life-long health, the process of microbial succession in the gut of infants remains poorly understood. Here, we analyze the patterns of taxonomic and functional change in the gut microbiota during the first year of life for a birth cohort of 13 infants. We detect that individual instances of gut colonization vary in the temporal dynamics of microbiota richness, diversity, and composition at both functional and taxonomic levels. Nevertheless, trends discernible in a majority of infants indicate that gut colonization occurs in two distinct phases of succession, separated by the introduction of solid foods to the diet. This change in resource availability causes a sharp decrease in the taxonomic richness of the microbiota due to the loss of rare taxa ($p = 2.06e\text{-}9$), although the number of core genera shared by all infants increases substantially. Moreover, although the gut microbial succession is not strictly deterministic, we detect an overarching directionality of change through time towards the taxonomic and functional composition of the maternal microbiota. Succession is however not complete by the one year mark, as significant differences remain between one-year-olds and their mothers in terms of taxonomic ($p = 0.009$) and functional ($p = 0.004$) microbiota composition, and in taxonomic richness ($p = 2.76e\text{-}37$) and diversity ($p = 0.016$). Our results also indicate that the taxonomic composition of the microbiota shapes its functional capacities. Therefore, the observed inter-individual variability in taxonomic composition during succession is not fully compensated by functional equivalence among bacterial genera and may have important physiological consequences. Finally, network analyses suggest that positive interactions among core genera during community assembly contribute to ensure their permanence within the gut, and highlight an expansion of complexity in the interactions network as the core of taxa shared by all infants grows following the introduction of solid foods.

Citation: Vallès Y, Artacho A, Pascual-García A, Ferrús ML, Gosalbes MJ, et al. (2014) Microbial Succession in the Gut: Directional Trends of Taxonomic and Functional Change in a Birth Cohort of Spanish Infants. PLoS Genet 10(6): e1004406. doi:10.1371/journal.pgen.1004406

Editor: David S. Guttman, University of Toronto, Canada

Received September 27, 2013; **Accepted** April 14, 2014; **Published** June 5, 2014

Funding: This work has been supported by the Spanish MICINN (project SAF2009-13032-C02-02 and project CSD2009-00006 of the CONSOLIDER program). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: francino_pil@gva.es

Introduction

The gastrointestinal tract (GIT) is a complex ecosystem where many factors, biotic and abiotic, play essential roles in reaching and maintaining a homeostatic equilibrium. The gut is endowed with the most diverse and dense microbiota of the human body, which plays fundamental roles in gut maturation, angiogenesis, immune system modulation, digestion, and protection from pathogens [1,2]. Given such important roles for health, the inter-individual variability of the human gut microbiota in adulthood and at any stage of development still defies our expectations. This variability is shocking in light of the ecological assumption that community composition and dynamics respond to and are structured mostly by the environment, "everything is everywhere but the environment selects" [3,4]. The GIT environment, although subject to inter-individual variation in diet and physiological parameters such as motility and transit time, presents a number of physical, chemical and mechanical properties that are mostly similar across individuals, including temperature, pH and surface tension values bound within limited ranges [5]. Consequently, we would expect a substantial degree of inter-individual convergence of GIT bacterial communities as a response to common selective pressures. Therefore, many studies have concentrated their efforts in the detection of a taxonomic core that would be shared by all individuals [6–10]. In that this search has been difficult, this view has recently evolved towards defining a few types of compositional profiles for the GIT microbiota. For example, Arumugam *et al.* [11] have stipulated that there are three universally distributed clusters of well-balanced host-symbiont states named enterotypes, driven mainly by bacterial composition, and that every individual's microbiota pertains to one of these enterotypes. However, the existence of such well-defined clusters of microbiota composition has been contested because their detection is highly dependent on the

Author Summary

Although knowledge of the complex community of microbes that inhabits the human gut is constantly increasing, the successional process through which it develops during infancy remains poorly understood. Particularly, although gut microbiota composition is known to vary through time among infants, the effect of this variability on the functional capacities of the community has not been previously explored. We simultaneously analyze the taxonomic and functional development of the gut microbiota in a birth cohort of healthy infants during the first year of life, showing that individual instances of gut colonization vary in their temporal dynamics and that clear parallelisms exist between functional and taxonomic change. Therefore, taxonomic composition shapes the functional capacities of the microbiota, and, consequently, successional variability may affect host physiology, metabolism and immunity. Nevertheless, we detect some overarching trends in microbiota development, such as the existence of two distinct phases of succession, separated by the introduction of solid foods, and a strong directionality of change towards the taxonomic and functional composition of the maternal microbiota. Understanding the commonalities and differences among individual patterns of gut colonization in healthy infants will enable a better definition of the deviations in this process that result in microbiota imbalances and disease.

methodology employed. Instead, Koren *et al.* [12] propose that gut microbiota composition across individuals is better represented by a series of gradients of taxon abundances that result in a bimodal distribution, where the ends of the spectrum harbor markedly different relative abundances of taxa.

Moreover, theoretical and experimental community ecology indicate that different communities can assemble under identical selective pressures. With basis on the neutral theory of community ecology [13] and on the metacommunity concept [14], each human gut can be considered to harbor a local microbial community, the composition of which will be driven essentially by the stochastic processes associated with resampling from the metacommunity of all gut microbiotas to which it is linked by organismal dispersal. If there are limitations to the dispersal capacities of different species, such processes could result in the assemblage of substantially different communities in spite of the physical, chemical and mechanical characteristics shared by all guts. In addition, within the context of neutral theory, the functional equivalence hypothesis proposes that multiple species may possess similar functional attributes, and it has been shown that species-rich communities are particularly prone to the evolution of functionally equivalent species [15]. This functional equivalence hypothesis is appealing in regards to the inter-individual variation in composition of the GIT microbiota. Under functional equivalence, the taxonomically different assemblages in different individual guts could present similar overall functional profiles, so that the inter-individual variation would have no impact on the host. Metagenomic and metatranscriptomic functional studies indicate that this scenario is plausible, since, in contrast to taxonomic variability, there seems to be conserved functional profiles among the microbiotas of different individuals [8,16–18]. Furthermore, this adds to the growing consensus that ecological community structure and function are better described by functional diversity (*i.e.* diversity of species traits [19]) rather than by taxonomic diversity [20], and that it is the alteration of functional diversity that will perturb the functioning of the ecosystem. The latter is further strengthened for microbial communities, as quantitative gene content analysis reveals specific fingerprints characterizing particular environments in spite of the substantial number of shared essential functions among bacteria [21].

In the case of human-associated microbiotas on which the host relies for specific functions, the alteration of functional diversity within the community can affect health status. Following this, several metagenomic studies have emphasized assessment of the functional diversity present in the GIT microbiota of healthy individuals, so as to be able to detect potential deviations in individuals affected by different diseases [2,8,9,22–25]. Most of these efforts have concentrated on adult individuals, while the assessment of functional capabilities in the GIT microbiota of infants has remained underexplored. However, infancy is the critical period for gut microbiota assembly, during which a constant dialogue with immune and metabolic development is established. Consequently, epidemiological and experimental lines of evidence indicate that the microbe-host interactions set in place during infancy represent a main determinant of life-long health or disease [26–28]. Despite its importance, the process of gut microbiota development in infants is still poorly understood, and has been mostly surveyed at the level of taxonomic succession by means of culture or of molecular analyses based on the 16S rRNA gene. These studies have shown that the differential exposure of the infant to vaginal, fecal and skin bacteria from the mother depending on the mode of birth (*i.e.*, vaginal *vs.* C-section), as well as the type of feeding during the first months of life (*i. e.*, breastmilk *vs.* formula), are main factors influencing the richness, diversity and composition of the gut microbial community [29–32]; that the earlier stages of infant gut microbiota development are characterized by high levels of inter-individual variability and a very uneven distribution of taxa; and that, as infant development progresses, microbial assemblages converge towards an adult-like composition with a more even taxa distribution [6,33,34]. On the other hand, to date, functional diversity in infants has mostly been explored in cross-sectional studies [23,25], and in a few longitudinal studies that have been limited to one [16,35] or a handful of infants [18]. It is important to keep in mind that cross-sectional studies do not follow individuals through time, but rather reflect single snapshots of the microbiota of different individuals of varying ages, and, therefore, cannot inform on the extent of inter-individual variation in microbiota dynamics. Thus far, the functional capabilities of the microbiota in infants have been shown to broadly mirror those of the mother from very early on, in spite of large taxonomic differences, although functions such as vitamin biosynthesis and xenobiotic degradation increase with time. However, much remains to be learnt about the process of functional development of the microbiota during colonization of an infant's GIT.

Taking into account all of the above, the present study explores the patterns of taxonomic and functional change along time during GIT microbiota development in a birth cohort of 13 infants. With this aim, we have collected fecal samples from healthy infants throughout the first year of life, and have obtained metagenomic sequence to characterize the phylogenetic composition and genetic repertoire of the microbiota present in each sample. In addition, in order to assess the progression of the infant's microbiota towards an adult-like state, we have also collected and sequenced the microbiota present in the mother before and one-year after childbirth. Because we obtain both taxonomic and functional data, we can evaluate the functional development of the GIT microbiota and its interactions with taxonomic community assembly, in the context of the dietary and physiological changes that characterize the first year of life.

Furthermore, because our analyses involve the prospective follow up of 13 infants, they allow us to evaluate several previously unexplored aspects of the GIT microbial succession process. Specifically, the availability of longitudinal data for several individuals sheds light on basic questions such as 1) whether taxonomic composition and functional development follow similar trends across individuals, 2) whether succession follows a strictly deterministic course, whereby early microbial assemblages set the stage for the next ones to come, 3) whether taxonomic variation among individuals during succession has an impact on the functional capabilities of the microbiota, and 4) whether community assembly is shaped by relationships of co-occurrence among taxa and how these evolve throughout succession. Overall, our data enable the characterization of microbial succession in the infant gut at unprecedented levels and, in particular, allow us to investigate whether the functional equivalence hypothesis can explain the inter-individual variability observed for this process.

Results/Discussion

Cohort, samples and sequencing

Given that our goal was to investigate the inherent variation in the process of microbial succession in the gut, rather than the specific alterations caused by factors such as type of delivery or infant feeding, we recruited to the study women having healthy pregnancies and stating their intention to exclusively breastfeed their infants during at least three months. We initially recruited 21 women, all residents of the city of Valencia, who were contacted during midwife visits. Due to various factors, we were able to obtain series of 4–5 infant fecal samples during the first year for only 13 of the enrolled women. At the moment of delivery, these women were between 29 and 42 years of age and had not taken antibiotics in at least three months before the onset of labor. Seven women received antibiotic during delivery and an eighth woman did so during the first week after. All 13 infants were born at term (>37 weeks of gestation), ten of them by vaginal delivery and three by C-section. Nine infants were exclusively breastfed during at least three months, three received a few formula feedings during the first days of life, and one was partially breastfed during the first month and formula-fed thereafter (Table 1). In addition to fecal samples, throughout the 12-months sampling period we obtained information regarding the infants' diet, general health and intake of antibiotics and other drugs (Table 1, Table S1), by means of specifically designed questionnaires that were given to the infants' parents. This information allowed us to establish that all infants remained healthy throughout most of the sampling period and that solid foods were introduced into their diets between the 3- and 7-months samplings, following patterns typical of Spanish Mediterranean infant diets [36].

Infant samples were collected at one week (I1), one month (I2), three months (I3, before introduction of solid foods), seven months (I4, after introduction of solid foods) and one year after birth (I5), and maternal samples were collected within one week prior to delivery (MA) and one year after (MB). We obtained 13 samples at each infant and maternal timepoint except for I2, for which only 9 samples were available, for an overall total of 87 samples that were processed for metagenomic pyrosequencing. After quality filtering, we obtained a total of 5,500,784 reads with a mean of 64,119 reads per sample and an average length of 348 bp (range 263–446 bp). For many reads, more than one Open Reading Frame (ORF) was recovered with a total of 9,968,776 ORFs and an average of 114,584 ORFs per sample. Annotation allowed for taxonomic assignment of 9,014,059 ORFs (103,610 per sample) and functional assignment of 675,141 ORFs (7760 per sample). Sequencing and annotation details as well as abundance tables for taxa and functions on a per sample basis are provided in the Supporting Information (Table S2, Table S3, Table S4). All sequences have been deposited in the IMG/M database [37] under the project name "Gut Microbiota of Spanish Mother-Infant Pairs".

The maternal microbiota changes between the perinatal period and one year after childbirth

Several changes are detected between the mother's gut microbiota days before childbirth and that present one year later. MA samples show a higher taxonomic richness ($p = 0.002$), due to a higher representation of rare taxa (abundance under 1% in all samples), but their functional diversity is lower ($p = 0.009$), indicating that they are functionally more redundant than MB samples (Figure 1C–1F). In addition, in clustering analyses based on similarity of microbiota composition, arbitrary clustering patterns are obtained where the MA and MB samples of the same woman do not group together, neither for taxonomic nor for functional composition (Figure S1). MA samples also present a larger range of inter-individual variability at both the taxonomic and functional composition levels (Figure 2A, 2B). These changes suggest a decrease in the host's capacity to regulate microbiota composition and function during late pregnancy, perhaps related to the low-grade inflammation of GIT mucosal surfaces and to the other immune, physiologic, hormonal and metabolic changes that occur during this period. Moreover, our results are in agreement with the recent demonstration that the maternal gut microbiota is dramatically altered between the first and third trimesters of pregnancy [38].

The composition of the maternal GIT microbiota during the perinatal period could be of great importance to the microbial colonization of the infant. Although the *in utero* environment has been considered sterile under normal conditions [30], culture-dependent and 16S rRNA gene pyrosequencing analyses have detected microorganisms in human meconium, amniotic fluid and umbilical cord, even when no rupture of membranes has occurred and in elective Cesareans [39–44]. The suite of changes that occur during late pregnancy [45,46] may facilitate the transport of maternal bacteria to the fetal GIT. In mice, translocation of live intestinal bacteria to mesenteric lymph nodes increases in late pregnancy [47,48], and dendritic cells have been shown to mediate increased bacterial translocation from the gut to blood and adipose tissue in obesity and diabetes [49], conditions similar to late pregnancy in terms of metabolic changes and the presence of a low-grade inflammatory state. Following translocation, intestinal bacteria could be transported in a controlled manner through lymph and blood, potentially reaching sites from which they could be transferred to the offspring, such as the placenta and the mammary glands. In support of this possibility, 16S rRNA gene pyrosequencing has detected very similar communities of organisms in meconium and in colostrum [41,50]. The taxonomic composition of these communities, which are most often dominated by Lactic Acid Bacteria (LAB), does not correspond to the bacterial abundances in maternal perinatal fecal samples [41], suggesting that the mother is able to regulate which bacteria reach the fetus, and/or that a restricted set of bacteria can survive in the fetal GIT to serve as a first inoculum and initiate the GIT colonization process.

Early colonizers and role of maternal transmission in the initial development of the GIT microbiota

The taxonomic composition detected in infants at the first timepoint analyzed, I1, is shown in Figure S2A. At this timepoint,

Table 1. Information regarding mothers and infants obtained from questionnaires answered by the infants' parents.

Sample	Age	Sex	Delivery	Antibiotics Mother[a]	Antibiotics Infant	Diet
MIP01-MA	29	-	-	No	-	-
MIP01-I1	1 Week	Male	Vaginal	-	-	Breast milk
MIP01-I2	1 Month	-	-	-	-	Breast milk
MIP01-I3	3 Months	-	-	-	-	Breast milk
MIP01-I4	7 Months	-	-	-	-	Solid foods
MIP01-I5	1 Year	-	-	-	-	Solid foods
MIP02-MA	36	-	-	No	-	-
MIP02-I1	1 Week	Female	Vaginal	-	-	Mixed
MIP02-I2	1 Month	-	-	-	-	Breast milk
MIP02-I3	3 Months	-	-	-	-	Breast milk
MIP02-I4	7 Months	-	-	-	-	Solid foods
MIP02-I5	1 Year	-	-	-	-	Solid foods
MIP03-MA	30	-	-	No	-	-
MIP03-I1	1 Week	Female	Vaginal	Amoxicillin	Oftalmowell[b]	Breast milk
MIP03-I2	1 Month	-	-	Amoxicillin	-	Breast milk
MIP03-I3	3 Months	-	-	Cefuroxime	-	Breast milk
MIP03-I4	7 Months	-	-	-	-	Solid foods
MIP03-I5	1 Year	-	-	Amoxicillin	Cefuroxime	Solid foods
MIP06-MA	42	-	-	Amoxicillin	-	-
MIP06-I1	1 Week	Female	C-section	-	-	Breast milk
MIP06-I2	1 Month	-	-	-	-	Breast milk
MIP06-I3	3 Months	-	-	-	-	Breast milk
MIP06-I4	7 Months	-	-	-	-	Solid foods
MIP06-I5	1 Year	-	-	-	Amoxicillin	Solid foods
MIP07-MA	31	-	-	Amoxicillin	-	-
MIP07-I1	1 Week	Male	C-section	-	-	Breast milk
MIP07-I3	3 Months	-	-	-	-	Breast milk
MIP07-I4	7 Months	-	-	-	-	Solid foods
MIP07-I5	1 Year	-	-	-	-	Solid foods
MIP08-MA	30	-	-	No	-	-
MIP08-I1	1 Week	Female	Vaginal	-	-	Breast milk
MIP08-I2	1 Month	-	-	-	-	Breast milk
MIP08-I3	3 Months	-	-	-	-	Breast milk
MIP08-I4	7 Months	-	-	-	-	Solid foods
MIP08-I5	1 Year	-	-	-	-	Solid foods
MIP09-MA	30	-	-	No	-	-
MIP09-I1	1 Week	Male	Vaginal	Amoxicillin	-	Mixed
MIP09-I2	1 Month	-	-	Amoxicillin	-	Mixed
MIP09-I3	3 Months	-	-	-	-	Formula
MIP09-I4	7 Months	-	-	-	-	Solid foods
MIP09-I5	1 Year	-	-	-	-	Solid foods
MIP12-MA	31	-		Amoxicillin	-	-
MIP12-I1	1 Week	Female	C-section	-	-	Mixed
MIP12-I2	1 Month	-	-	-	-	Breast milk
MIP12-I3	3 Months	-	-	Cefixime	-	Breast milk
MIP12-I4	7 Months	-	-	-	-	Solid foods
MIP12-I5	1 Year	-	-	-	-	Solid foods
MIP13-MA	31	-	-	Benzylpenicillin	-	-
MIP13-I1	1 Week	Male	Vaginal	Amoxicillin	-	Mixed
MIP13-I3	3 Months	-	-	-	-	Breast milk

Table 1. Cont.

Sample	Age	Sex	Delivery	Antibiotics Mother[a]	Antibiotics Infant	Diet
***MIP13*-I4**	7 Months	-	-	-	-	Solid foods
***MIP13*-I5**	1 Year	-	-	-	-	Solid foods
***MIP16*-MA**	39	-	-	Amoxicillin	-	-
***MIP16*-I1**	1 Week	Male	Vaginal	-	-	Breast milk
***MIP16*-I2**	1 Month	-	-	-	-	Breast milk
***MIP16*-I3**	3 Months	-	-	-	-	Breast milk
***MIP16*-I4**	7 Months	-	-	-	-	Solid foods
***MIP16*-I5**	1 Year	-	-	-	-	Solid foods
***MIP17*-MA**	39	-	-	No	-	-
***MIP17*-I1**	1 Week	Male	Vaginal	-	-	Breast milk
***MIP17*-I3**	3 Months	-	-	-	-	Breast milk
***MIP17*-I4**	7 Months	-	-	-	-	Solid foods
***MIP17*-I5**	1 Year	-	-	-	-	Solid foods
***MIP19*-MA**	33	-	-	No	-	-
***MIP19*-I1**	1 Week	Female	Vaginal	-	-	Breast milk
***MIP19*-I3**	3 Months	-	-	-	-	Breast milk
***MIP19*-I4**	7 Months	-	-	-	-	Solid foods
***MIP19*-I5**	1 Year	-	-	-	-	Solid foods
***MIP21*-MA**	35	-	-	Amoxicillin	-	-
***MIP21*-I1**	1 Week	Male	Vaginal	-	-	Breast milk
***MIP21*-I2**	1 Month	-	-	-	-	Breast milk
***MIP21*-I3**	3 Months	-	-	-	-	Breast milk
***MIP21*-I4**	7 Months	-	-	-	-	Solid foods
***MIP21*-I5**	1 Year	-	-	-	-	Solid foods

MIP: Mother Infant Pair.

[a]For MA samples we report whether antibiotics were given during childbirth and the specific antibiotic given. In the case of C-sections, we report administration of amoxicillin, which is the standard practice in Spanish hospitals. None of the mothers had taken antibiotics before childbirth for at least three months.

[b]Oftalmowell is a combination of gramicidin, neomycin and polymyxin B.

doi:10.1371/journal.pgen.1004406.t001

the GIT microbiota of different infants is quite divergent, since in each one of them a single genus dominates extensively. *Bacteroides* dominance is the most prevalent, being detected in 5 of the neonates, followed by *Clostridium* (3 neonates), *Veillonella* (2 neonates), *Bifidobacterium* (2 neonates), and *Escherichia* (1 neonate). Among the 9 infants who were born vaginally and were breastfed exclusively (MIPs —Mother-Infant Pairs— 1, 3, 8, 16, 17, 19 and 21) or received a little amount of formula early on (MIPs 2 and 13), all five dominance patterns can be found, although *Bacteroides* is the most common. *Bifidobacterium* dominates in one exclusively breastfed infant (MIP17) and in the infant who was only partially breastfed (MIP9), both of whom were vaginally born. On the other hand, the three infants born by C-section had I1 microbiotas dominated by a Firmicutes genus, *i. e.*, *Clostridium* (MIPs 6 and 12) or *Veillonella* (MIP7). This is in agreement with previous studies indicating that C-section delays the establishment of *Bacteroides*, *Bifidobacterium* and *E. coli* [31,51]. The Canonical Correspondence Analysis (CCA) in Figure S3 shows that C-section does influence the taxonomic composition of the infant microbiota at I1, although it only explains 16% of the total variability. Antibiotic use during delivery and supplementation of the infant's diet with formula (Table 1) play a more limited role, as they explain 11% and 7% of the total variability at this timepoint, respectively (Table S5).

The five genera that dominate the I1 microbiota in different neonates may have had an important head start for GIT colonization, as all of them have been identified in meconium, although they were not the most common taxa revealed by 16S rRNA pyrosequencing in term infants [41]. Moreover, we have previously shown that the meconia passed by two of the infants in this cohort (MIPs 2 and 21) contain 16S rRNA gene sequences, including sequences from *Bacteroides* and *Clostridium*, that are also recovered at 100% identity from the corresponding maternal samples and infant samples from different timepoints [41]. This suggests that these bacteria can be acquired *in utero* and then maintained in the infant for long periods of time. In addition, here we detect that one-week-old infants share a substantial, but highly variable among individuals, percentage of GIT microbiota genera with their respective mothers prior to giving birth (between 26% and 88%, average 71%). These taxa could have been acquired *in utero*, during delivery or through breast milk.

The early colonizers of a given environment can have crucial consequences for the further development of the community. Theoretical models of succession differ on whether they consider that those organisms able to establish themselves in a long-term manner in a given environment will be able to colonize it from the start, or, rather, that early succession will be dominated exclusively by "opportunists" or "pioneers" adapted to the transient conditions common to all recently opened spaces. Pioneers are expected to have cosmopolitan distributions, broad dispersal and rapid growth capabilities in order to arrive first and quickly occupy

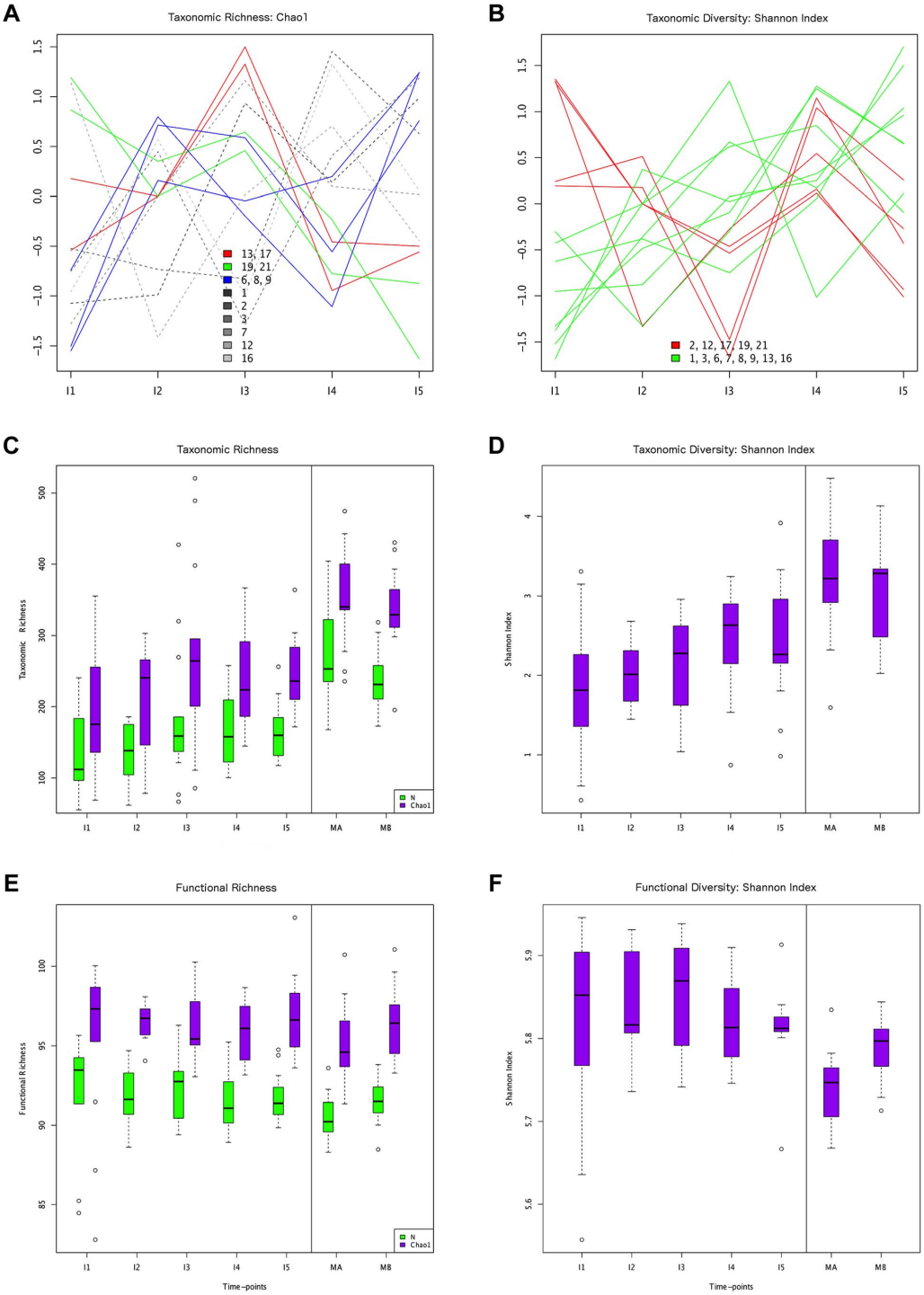

A
Taxonomic Richness: Chao1
13, 17
19, 21
6, 8, 9
1
2
3
7
12
16
I1
I2
I3
I4
I5
B
Taxonomic Diversity: Shannon Index
2, 12, 17, 19, 21
1, 3, 6, 7, 8, 9, 13, 16
C
Taxonomic Richness
Taxonomic Richness
N
Chao1
MA
MB
D
Taxonomic Diversity: Shannon Index
Shannon Index
E
Functional Richness
Functional Richness
Time-points
F
Functional Diversity: Shannon Index
Shannon Index

Figure 1. Different behaviors of taxonomic and functional richness and diversity through infant gut microbiota development. Hierarchical clustering of temporal profiles for (A) taxon richness (Chao1 estimator) and (B) taxon diversity (Shannon index), showing the extent of variation among the 13 infants. Values are centered at the mean of all samples and scaled by the standard deviation. Colored profile clusters have > 95% support based on multiscale bootstrap resampling. The boxplots in (C) and (D) summarize the general behavior of taxon richness and diversity for all infants. Taxon richness (C) shows an increase in median values with time interrupted by the introduction of solid foods (I4), when a decrease in richness is observed. Taxon diversity (B) shows an increase in median values from I1 to I4 followed by a decrease between I4 and I5. Functional richness (E) and diversity (F) show no specific pattern but rather fluctuate with time.
doi:10.1371/journal.pgen.1004406.g001

an empty space [52,53]. Most of the genera that we find dominating at I1 hardly correspond to this definition. Except for *Clostridium* and *Escherichia*, the remaining genera (*Bacteroides*, *Veillonella* and *Bifidobacterium*) are intermediate or slowly growing species with known optimal generation times ranging from one to three hours [54,55]. Moreover, their metabolism is strictly anaerobic, their environmental distribution is not cosmopolitan but host-associated [56], and they can be found at high abundances in later stages of succession. These observations suggest that these organisms are not opportunists taking advantage of a newly available habitat, but rather GIT-specialists, highly competitive in this particular environment. Therefore, the GIT microbial succession does not seem to follow a "facilitation" model, in which pioneers colonize an open space and create the necessary conditions for more specialized late-coming organisms [52]. Although it is possible that a facilitation phase may have taken place at a very rapid pace during the first days after birth, it is still noteworthy that, with the exception of one infant whose microbiota consisted almost exclusively of *Escherichia* and other enterobacteria, all infants at I1 had a microbiota that was already dominated by a strict anaerobe, contrary to the common assumption that early colonizers must be facultative anaerobes [57]. Rather, it suggests that anaerobic conditions are quickly established, and that the strict anaerobes have strong competitive

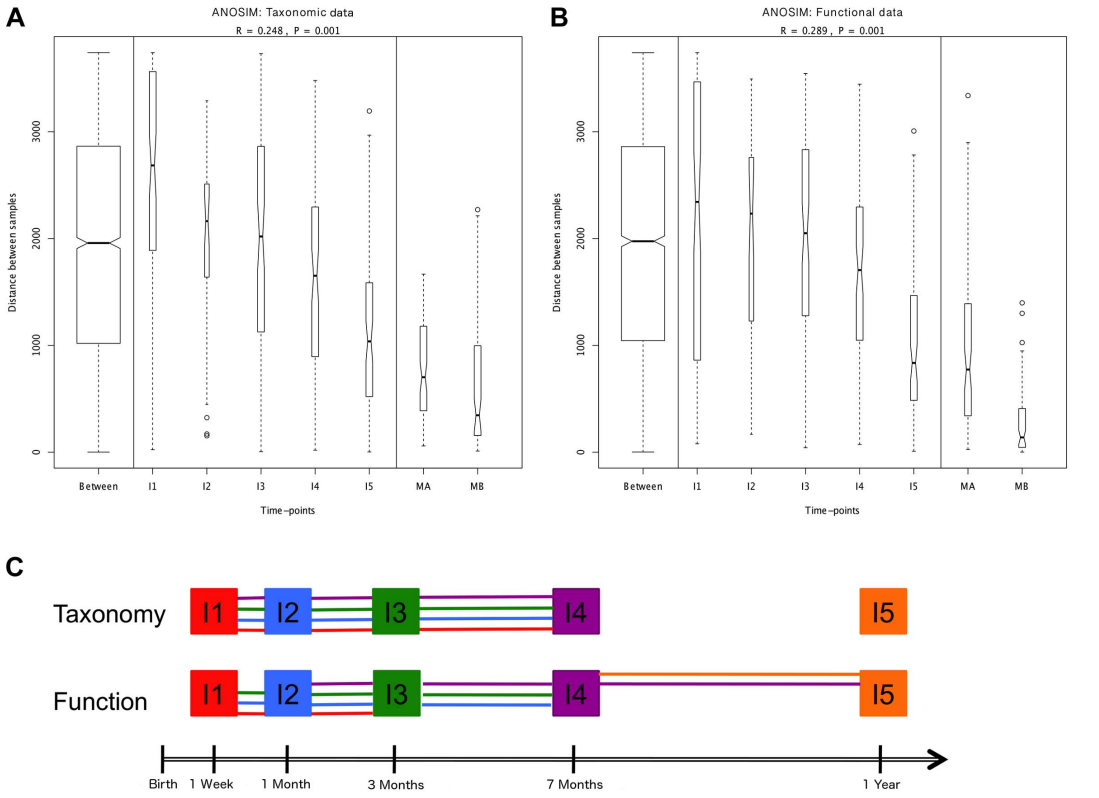


Figure 2. ANOSIM comparison of timepoints. Overall analyses for taxonomic (A) and functional (B) Bray-Curtis distances among all samples. The length of the bows indicates the level of heterogeneity and the width the number of compared samples. Statistically significant differences among timepoints are detected for both taxonomic and functional data. Note the decrease in heterogeneity with time in infants and the larger heterogeneity in MA compared to MB samples. (C) Representation of pairwise ANOSIM analyses between timepoints. Each timepoint is represented by a color and is linked by lines of this color to all timepoints from which it is not significantly different. For functional composition, significant differences appear between timepoints that are more separated in time, indicating directionality along infant development, but no such pattern is detected at the taxonomic level.
doi:10.1371/journal.pgen.1004406.g002

advantages that allow them to rapidly dominate over any facultative anaerobes that could have been present during the very first days after birth, such as the vaginal *Lactobacillus* acquired through the birth canal [29].

If the I1-dominating genera were present in the GIT before the moment of birth, even if at very low abundances, a rapid expansion may have occurred as soon as conditions became favorable. The start of breastfeeding should select for organisms able to grow in its main constituents, such as lactose and Human Milk Oligosaccharides (HMOs). These oligosaccharides are the main growth factors for *Bifidobacterium* [58], but recent work has shown that they can also sustain the efficient growth of *Bacteroides* [59]. Remarkably, although *Bacteroides* has often been reported to be uncommon during the neonatal period, we detected a microbiota dominated by this genus in five of the 13 one-week-old infants. In fact, it should not be surprising that *Bacteroides* might quickly establish, given that it is the only genus besides *Bifidobacterium* known to efficiently grow on HMOs and that it is also one of the most efficient utilizers of the mucin molecules that line the intestinal epithelium [60]. Many species of *Bifidobacterium*, *Escherichia* and *Clostridium* can also utilize mucin, in addition to lactose [61]. *Veillonella*, on the other hand, can't metabolize carbohydrates and requires short-chain fatty acids (SCFA), such as lactate or pyruvate, for growth [62]. Its dominance in two of the one-week-old infants suggests that a short food chain had already been established whereby *Veillonella* could have access to SCFA produced by other GIT genera, for instance by lactose fermentation. In this regard, it can be noted that, in the infants having a high abundance of *Veillonella*, genera that can ferment lactose to SCFA, such as *Clostridium* or *Streptococcus*, were indeed also abundant.

Dynamics of taxonomic and functional richness and diversity during the first year of life

In order to characterize the dynamics of richness and diversity in the infant microbiota from the first week to the one-year mark, we computed the Chao1 estimator [63] and the Shannon index [64], for both taxa and functions (Table S6). Chao1 estimates richness, *i. e.*, the number of taxa or functions present in a community, whereas the Shannon index of diversity takes into account both richness and evenness, *i. e.*, how similar the abundances of the different taxa or functions are. The dynamics of taxon richness along time are presented in Figure 1 for individual infants (A) and across all individuals (C). Chao1 values increase overall between I1 and I5 ($p = 6.18e\text{-}18$), an increase that is present in most of the infants. However, the increase is not linear (linear regression $p = 0.205$, Figure S4A), nor continuous. In most infants, richness is under two thirds of the maternal value (MB) at I1, and then increases from I1 to I2. Although change across all infants is not significant for this first interval ($p = 0.139$), the tendency to increase is reflected in median values (Figure 1C). In the I2–I3 interval, even though richness increases or decreases in similar numbers of infants, overall it is higher at I3 than at I1 ($p = 2.33e\text{-}38$) and I2 ($p = 3.70e\text{-}24$), partly due to the presence of three outliers having very high I3 values (MIPs 7, 13 and 17). Then, from I3 to I4, the interval in which solid foods were introduced, most infants present a decrease in richness, which is significant across individuals ($p = 2.06e\text{-}09$). This decreasing trend may or may not reverse from I4 to I5, so that the change in this interval does not reach significance ($p = 0.107$) and richness values at I5 remain significantly lower than those that had been attained by I3 ($p = 1.13e\text{-}05$), before the introduction of solid foods. Richness at I5 is also significantly lower than that of the mothers ($p = 2.76e\text{-}37$ *vs.* MB), although by this final timepoint most infants have already surpassed two thirds of the MB richness value. Hierarchical clustering analysis of the temporal profiles of richness change for individual infants retrieves three significant clusters, one including infants 6, 8 and 9, another including infants 13 and 17, and a last cluster including infants 19 and 21 (Figure 1A). These clusters do not associate with delivery type, antibiotic use or formula supplementation.

The taxon diversity changes undergone by the different infants during the year are as variable as those seen for taxon richness, but some trends can also be discerned (Figure 1B, 1D). These trends mirror the behavior of richness in some time intervals, but not in others. As seen for richness, taxon diversity increases significantly between I1 and I5, and, in this case, regression analysis indicates that the increase can be considered linear when this entire period is considered (p-value = 0.024; Figure S4B). The pattern of change is similar to that of richness throughout the first three months; however, trends that are opposite to those observed for richness are present after the three months mark, as, in most cases, the Shannon index increases in I3–I4 and decreases in I4–I5. In terms of median values, there are increases between all consecutive timepoints except I4–I5, when the median diversity decreases to a value similar to that attained by I3 ($p = 0.053$; Figure 1D). By I5, taxon diversity is still significantly lower than that of MB ($p = 0.016$) but most infants have reached a Shannon index value that surpasses two thirds of that of their mother, a situation that is again comparable to that of taxon richness. Hierarchical clustering groups the diversity temporal profiles of the infants into two significant clusters, one including infants 2, 12, 17, 19 and 21, and the other, in which the trend towards a linear increase in diversity with time is more pronounced, including infants 1, 3, 6, 7, 8, 9,13 and 16 (Figure 1B). Clustering patterns differ then for taxonomic richness and diversity, with only infants 6, 8 and 9, on one hand, and infants 19 and 21, on the other, clustering together for both parameters.

The opposite trends in taxon richness and diversity from I3 to I5 suggest that changes in richness correspond to the appearance and disappearance of rare taxa, which, if substantial, would respectively result in lower and higher degrees of evenness in the distribution of taxa abundances in the community, captured in the Shannon diversity index. Indeed, rank abundance curves confirm that richness changes are driven mainly by the removal of rare genera in I3–I4, followed by the addition of different rare genera in I4–I5 (data not shown).

Regarding functional richness and diversity in the infant samples (Figure 1E, 1F), their behavior is in sharp contrast to that observed at the taxonomic level, as these functional parameters fluctuate through time across a relatively narrow range of values, with no clear trends to increase or decrease along development (Table S6, Figure S4C, S4D). In fact, even at the earliest sample collection times, the median values of functional richness and diversity in infants are already similar or higher than those obtained for the mothers, particularly in the perinatal samples (MA), which present the lowest values. This indicates that the infant microbiota attains a level of functional complexity similar to that of the mothers from very early on, possibly due in part to the general presence of essential bacterial functions and of those specifically needed for survival in the gut environment.

Succession in infants does not follow a strictly deterministic course

Whether ecological successions are deterministic processes is still a matter of contention. In microbial communities, this question has rarely been explored. Our prospective cohort analysis enables us to address this issue in several complementary ways. We have

already described the lack of a common successional pattern across individuals in terms of the magnitude and direction of taxon richness and diversity changes between timepoints. Clustering analysis of individual samples using the Bray-Curtis distance [65] provides other means of assessing the degree of determinism in the infants' successional paths. Firstly, we analyzed the clustering patterns within each MIP to determine whether they all share the same topology. For both taxonomic and functional composition, MIPs have seemingly idiosyncratic clustering patterns (Figure S5). The most marked tendency is the grouping of I5 with maternal samples, observed in 31% and 62% of the MIPs at the taxonomic and functional levels, respectively, independently of the mode of birth. The lack of a common clustering pattern across the different MIPs reinforces the notion that the infants' successional paths follow non-deterministic dynamics, although a trend of convergence towards the maternal functional composition by the end of the year is suggested.

Global comparisons of all infant and maternal samples at the taxonomic and functional levels also point in this direction. Such comparisons reveal no clear pattern of sample clustering, neither by individual nor by timepoint (Figure 3A–B). The fact that samples from the same timepoint do not cluster together indicates that the microbiota present at each timepoint can not be defined as a well-differentiated, discrete and predictable community, such as the seral communities postulated in some models of vegetational succession [66]. Nevertheless, some degree of unevenness can be observed in the distribution of samples across clusters, pointing towards an effect of age on microbiota composition. The taxonomic heatmap in Figure 3A shows a large cluster (a) that contains 30 infant samples but only one maternal sample, as well as two clusters (b and c) that contain nearly all of the maternal samples and some I4–I5 infant samples. A similar effect can be seen in the functional heatmap (Figure 3B), where a single cluster contains the 26 maternal samples, most of the I5 samples and only a few of the samples from other infant timepoints. In other words, as in the MIP-based analyses, I5 shows here a clear tendency to cluster with the maternal samples, for both taxonomy and function.

Finally, comparison of the heatmaps corresponding to each timepoint in the series (Figure S2) enables us to evaluate whether early microbial assemblages determine the nature of the next ones to come. Rather, it can be seen that the patterns of association among samples from different individuals change through time. For instance, infants 2 and 16 have very similar taxa composition profiles at I1, while they differ widely at all subsequent timepoints. Conversely, infants 17 and 19 are the most similar one-year-olds, whereas at earlier timepoints they had microbiotas dominated by *Bifidobacterium* and *Bacteroides*, respectively. The varying patterns of association among samples through time indicate that early similarity among infants does not predict similar developmental paths or one-year mark outcomes.

Is there directionality in taxonomic and functional change along development?

We next set out to investigate whether, in spite of the lack of determinism in successional paths, an overall pattern of directional

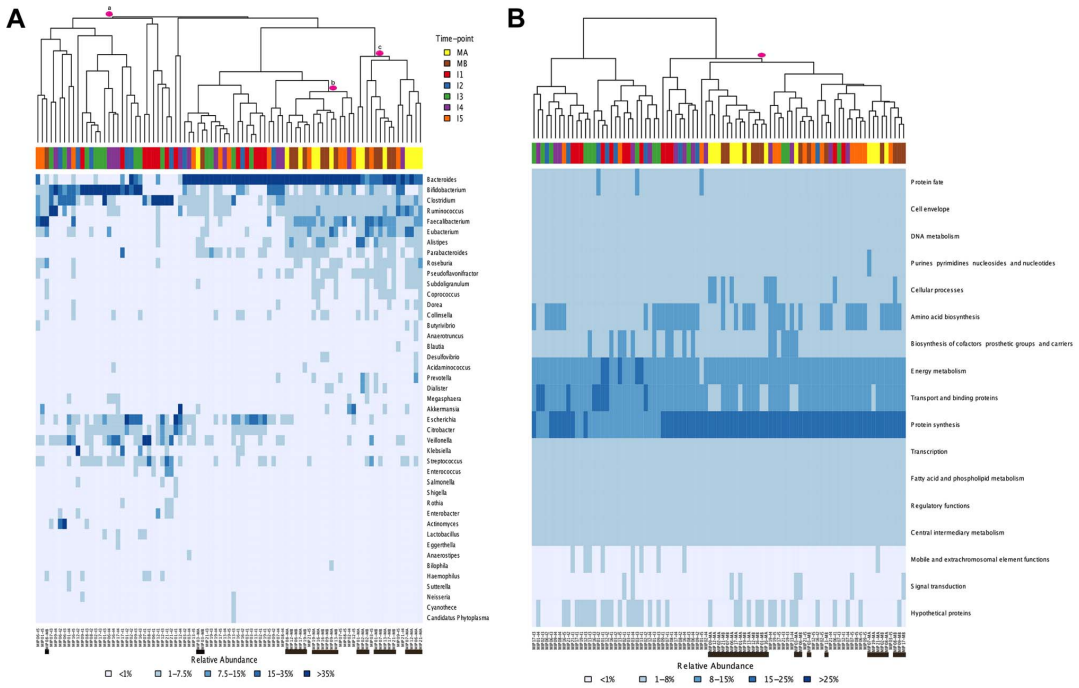


Figure 3. Heatmaps and clustering of individual gut microbiota samples for taxonomic (A) and functional composition (B). Clustering was based on Bray-Curtis distances. (A) Only the genera above 1% abundance in at least one sample are depicted. (B) Functional composition was established based on TIGRFAM main functional roles. Each sample is identified at the bottom of the heatmaps by a code that specifies the MIP to which it belongs and the corresponding timepoint. Maternal samples are additionally highlighted by means of black bars. Colors on top of each heatmap represent the timepoints to which samples belong. Pink circles identify specific clusters referred to in the text.
doi:10.1371/journal.pgen.1004406.g003

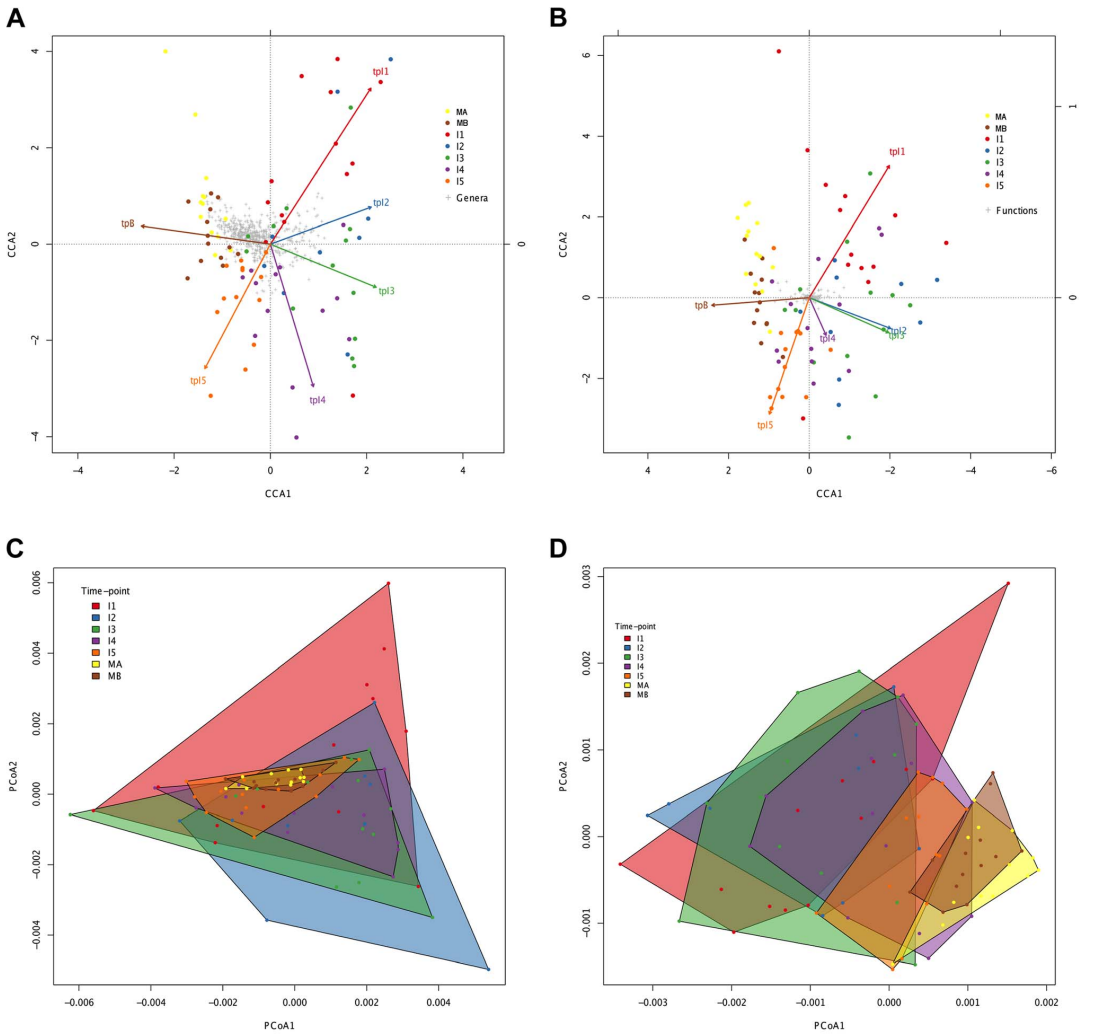


Figure 4. Directionality in taxonomic and functional change through time. Canonical Correspondence Analysis (CCA) of taxonomic (A) and functional (B) data, showing that the main axis (CCA1) separates infant timepoints I1, I2, I3 and I4 from I5, MA and MB. The percent variation explained by the main axis is 60.22% in A and 81.57% in B, while CCA2 explains 14.20% variation in A and 6.99% in B. The direction of the timepoint arrows indicates the main axis of deviation from the reference maternal timepoint (MA). Taxonomic (C) and functional (D) Principal Coordinates Analyses (PCoA) depicting convex hulls enclosing all samples pertaining to a determined timepoint. The percent variation explained by the main axis is 46.60% in C and 30.28% in D, while PCoA2 explains 23.00% variation in C and 16.04% in D. Heterogeneity within timepoints is represented by arrow length (CCA) or convex hull area (PCoA). All analyses identify a progressive change from timepoint to timepoint with clear directionality towards the composition of the mothers.
doi:10.1371/journal.pgen.1004406.g004

change through time towards an adult-like microbiota can be discerned, as suggested by global and MIP-based clustering analyses. To this aim, we employed several multivariate analyses based on the Bray-Curtis distances among samples. We first examined whether there are significant overall differences among the entire set of analyzed timepoints. Comparison of distances between and within timepoints revealed that significant differences exist at both the taxonomic and functional levels (ANOSIM: taxonomic $R = 0.30$, $p = 0.001$ & functional $R = 0.27$, $p = 0.001$). The plots in Figure 2A–B display the amount of variation among samples within and between timepoints and allow us to appreciate the wider divergence between samples in earlier timepoints and the progressive increase in homogeneity as the gut microbiota develops, as previously noted [6]. This can be considered a first clearly directional trend in the data, observable at both the taxonomic and functional levels. We then performed a series of

pairwise ANOSIM analyses in order to detect where the main differences among timepoints lie. All infant samples, including those from the one-year timepoint, are distinguishable from maternal samples both by taxonomy and by function (Table S7). The results obtained for comparisons between infant samples are illustrated in Figure 2C, where each timepoint is represented by a color and is linked by lines of this color to all timepoints from which it does not differ significantly. At the taxonomic level, no pairwise comparison for timepoints I1 to I4 identifies significant differences, whereas each of these timepoints is distinguishable from the I5 timepoint. In other words, this analysis reveals no progressive increase in taxonomic composition distance along time. In contrast, at the functional level, although none of the infant timepoints is significantly different from its immediate neighbor, differences become significative for timepoints that are separated by one or two intermediate timepoints. That is, in this case, larger functional differences appear between timepoints that are more separated in time, indicating a clear directionality along infant development. Nevertheless, this type of analysis does not show a progression in microbiota composition towards the adult state, as all infant timepoints remain distinguishable from those of the mothers, for both taxonomy and function.

In order to further visualize how the compositional variation among samples is distributed, we performed multivariate statistical techniques that provide the coordinates of the samples in a reduced space representing the main variation components. In contrast to the ANOSIM analyses, Canonical Correspondence Analyses (CCA; Figure 4A, 4B) identify a progressive change from timepoint to timepoint with clear directionality towards the adult state for both taxonomy and function. The taxonomic and functional CCAs recover the same pattern, with a slight difference in terms of the proximity between timepoints I2 and I3, which are closer for the functional data set. Although discrete clusters of samples by timepoint are not present, the CCA plots show an orderly displacement from I1 to I5, clearly observed in the changing direction of the timepoint arrows, which indicates the main axis of deviation from the reference maternal timepoint (MA). Moreover, in both cases the first axis of the CCA graph separates the majority of infant samples (I1, I2, I3 and I4) from the one-year-old and maternal samples (I5, MB and MA), indicating that progressive change throughout the first year has resulted in a microbiota that is more similar to that of the mothers. We also analyzed the taxonomic and functional datasets with Principal Coordinates Analyses (PCoA) performed on matrices of Gower distances [67], followed by the drawing of convex hulls enclosing all samples pertaining to a particular timepoint [68]. It can be seen in Figure 4C and 4D that, for both taxonomy and function, there is a general decrease of the area of the convex hulls with age, indicating again a decrease in heterogeneity among coetaneous samples, as well as a time-ordered displacement of the infants' convex hulls towards those of the mothers. We calculated the taxonomic and functional dissimilarities between two timepoints by estimating the non-overlapping areas of their convex hulls (Table S8). As expected, in both cases dissimilarity is lowest between maternal samples and between infant timepoints that are close in time, and is at its peak when I1 convex hulls are compared to the maternal ones. So, both CCA and PCoA coincide in showing a clear time-ordered displacement of taxonomic and functional composition whereby each successive infant timepoint becomes more similar to the mothers.

However, the convex hulls in Figure 4C and 4D point out an interesting difference between taxonomic and functional compositional change. In the case of taxonomic composition, the maternal convex hulls are enclosed within the space occupied by the infant timepoints, which seem to close in around the maternal hulls as time progresses. In contrast, in the case of function, the maternal hulls occupy the rightmost part of the graph and the infant samples progressively shift in that direction, so that some degree of overlap with the maternal hulls is only observed from the I3 timepoint onwards. This suggests that the GIT microbiota undergoes a more pronounced directional shift during succession at the functional than at the taxonomic level.

Parallelisms between taxonomy and function counter the functional equivalence hypothesis

In spite of some differences, we have just shown that the changes in taxonomic and functional microbiota composition with time are similar both in terms of the directionality of change toward the maternal profile and of the progressive reduction of heterogeneity among individual samples. This argues for an effect of the taxonomic composition of the microbiota on its functional gene repertoire. In order to further investigate the relationship between taxonomy and function, we analyzed the functional similarities among GIT microbiota genera. For this, we determined and compared the functional profiles of all genera that reached 1% abundance in at least one sample. Because not enough information was available in a sample per sample basis for each genus, functional profiles were established after pooling all samples for a given timepoint. Functional profiles were defined as vectors containing the relative abundances of each TIGRFAM subrole within a particular genus and timepoint. We then constructed a dendrogram clustering genera by functional profile similarity as measured by the Bray-Curtis distance (Figure 5). The resulting dendrogram mainly follows phylogenetic relationships, suggesting that each phylogenetic group has a characteristic set of functional profiles. At the genus level, the functional profiles computed for the different timepoints generally form an exclusive group, suggesting that either the same species of the genus are present along development or that all members of the genus share similar sets of genes. Moreover, clustering by phylogenetic affiliation also occurs at higher taxonomic ranks, as functional groups comprising only members of specific families and orders are recovered. Six major functional groups are obtained: Group 1, enclosing all Enterobacteriales; Group 2, enclosing all Bacteroidales and Verrucomicrobiales; Group 3, comprising all Selenomonadales, plus the Clostridiales genera *Pseudoflavonifractor* and *Subdoligranulum* and the δ-proteobacteria *Desulfovibrio*; Group 4, enclosing all Pasteurellales; Group 5, comprising most of the Clostridiales, and Group 6, enclosing the Clostridiales genera *Anaerostipes* and *Faecalibacterium*, the Lactobacillales and all Actinobacteria. Interestingly, only members of the phyla Firmicutes (Clostridiales, Lactobacillales and Selenomonadales) and Proteobacteria (Enterobacteriales, Pasteurellales and the genus *Desulfovibrio*) are present in multiple major functional groups. In particular, the order Clostridiales is the most functionally diverse, as it is the only order split into several of the major groups, even though a large majority of genera are found in functional group 5.

Although the general topology of the dendrogram in Figure 5 implies that the functional profile of taxa is strongly related to phylogenetic affiliation, some particular groupings indicate that functional convergence may occur among distantly related taxa. Most remarkable is the clustering in functional group 6 of the Bifidobacteriales and other less abundant Actinobacteria with the Firmicutes order Lactobacillales, which comprises the Lactic Acid Bacteria (LAB). Bifidobacteria are known to share many metabolic properties with the LAB, notably the production of lactic acid as a main endpoint of carbohydrate fermentation. In addition, group 6

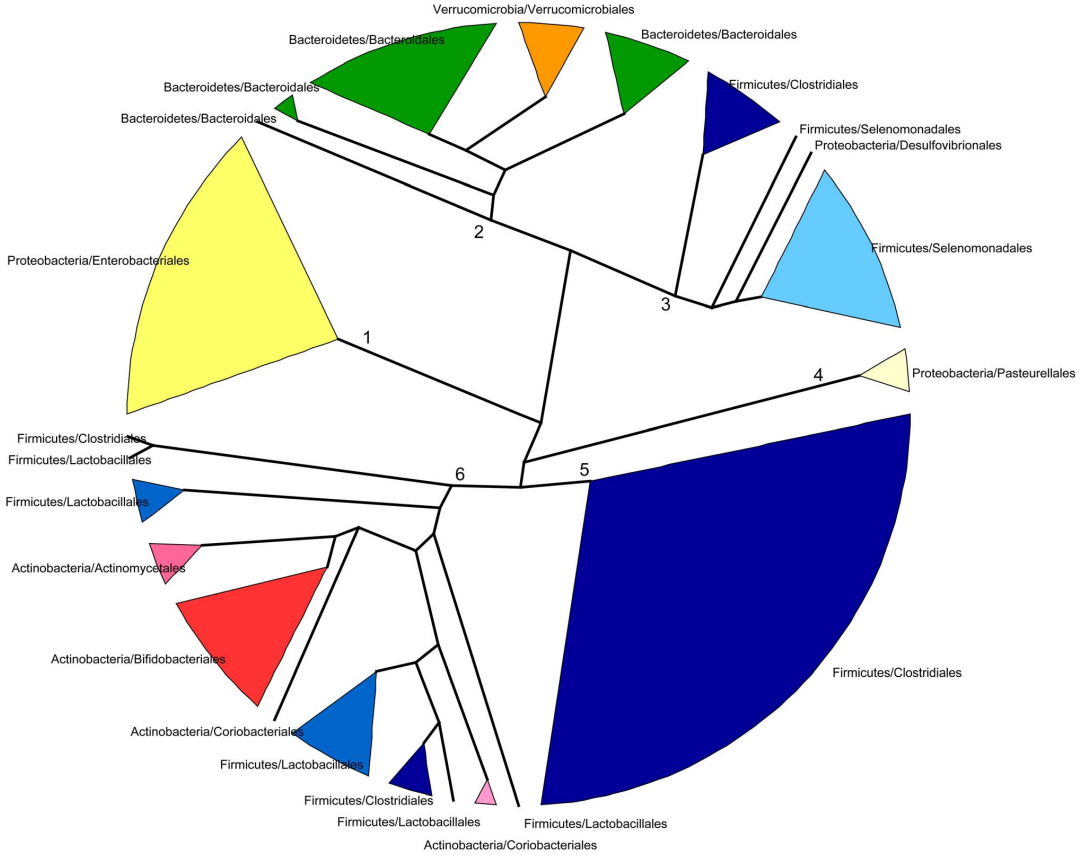


Figure 5. Dendrogram showing six main groups of gut microbiota genera based on functional profile clustering. Functional profiles were defined as the relative abundances of TIGRFAM subroles in a given genus. Only genera present in any sample at >1% abundance and having genes representing at least 50% of the 108 subroles detected in our complete data set were included. Clustering was based on the complete linkage method applied to a matrix of pairwise Bray-Curtis distances between the functional profiles of genera. Branches in the resulting dendrogram were collapsed when genera on the tips pertained to the same order. Orders of the same phylum have different shades of the same color.
doi:10.1371/journal.pgen.1004406.g005

also contains two Clostridiales genera, the acetate-requiring butyrate-producers *Faecalibacterium* and *Anaerostipes*. *Faecalibacterium* can also produce lactate, whereas *Anaerostipes* rather consumes it to produce butyrate. Another interesting grouping is that of the Verrucomicrobiales and the Bacteroidales, mainly represented by *Akkermansia* and *Bacteroides*, two genera that share important metabolic functions in the gut, as both are acetate and propionate producers and highly adept at mucin degradation [60].

Nevertheless, the observed groupings among phylogenetically distant taxa do not indicate functional equivalences that could account for the inter-individual variation in patterns of taxon dominance. This is most evident for timepoint I1, in which the taxonomic discrepancy among samples is maximal and the microbiotas of each individual are mostly dominated by a single genus (Figure S2A). Under the functional equivalence hypothesis, we would expect that the most abundant taxa present in the different samples would have similar functional profiles, independently of their phylogenetic lineage affiliation, and would cluster together into specific functional groups. Rather, the five genera that dominate the microbiota in different I1 infants (*Bacteroides*, *Clostridium*, *Veillonella*, *Bifidobacterium* or *Escherichia*) are found in deeply separated groups of the functional profile tree. This suggests that their functional capabilities are vastly different, and therefore that functional similarity and the functional equivalence hypothesis can't explain their presence as dominating taxa in the microbiotas of different infants.

Dynamics of specific taxa and functions along development

Figures 3A and S2A show that, overall, the infants' samples can have high abundances of bacteria such as *Escherichia*, *Citrobacter*, *Bifidobacterium*, *Veillonella* and *Streptococcus*, in addition to *Clostridium* and *Bacteroides*, which are also common in adults. Venn diagrams allowed us to visualize details of the dynamics of taxa acquired or lost at each particular timepoint and of those that were maintained throughout the whole process of development. We identified a small core of ten genera that are present at all timepoints, in all infants and adults, although at very different abundances,

comprised of *Bacillus*, *Bacteroides*, *Clostridium*, *Enterococcus*, *Escherichia*, *Eubacterium*, *Lactobacillus*, *Prevotella*, *Streptococcus* and *Vibrio*. Of note, this global core of 10 genera includes members of four of the functional groups defined above (Group 1: *Escherichia*; Group 2: *Bacteroides* and *Prevotella*; Group 5: *Clostridium* and *Eubacterium*; and Group 6: *Enterococcus*, *Lactobacillus* and *Streptococcus*). *Bacillus* (order Bacillales) and *Vibrio* (Vibrionales) are not represented in the functional profiles dendrogram because their low abundances precluded the computation of reliable functional profiles.

We also identified separately the core genera of every timepoint (Table S9), and the Venn diagram in Figure 6A shows the intersections of the different infant "timecores". New genera appear at every timecore, some of which remain in all subsequent timecores and are also present in those of the mothers. This is the case of *Bifidobacterium and Ruminococcus*, which join the core at timepoint I2, and of *Pseudoflavonifractor*, which joins at I3. At I4 there is an input of 12 new core taxa that will remain in the I5 timecore, including *Anaerostipes*, *Blautia*, *Coprococcus*, *Dorea*, *Fusobacterium* and *Roseburia*, and 16 new core genera make their appearance at I5, including *Acidaminococcus*, *Alistipes*, *Butyrivibrio*, *Parabacteroides* and *Subdoligranulum*. All of the core genera that are introduced in I4 and I5 are also present in the MB, and, with few exceptions, in the MA maternal timecores. In contrast, several genera of enteric bacteria appearing in the I2 timecore only remain through I3, or are maintained until I5 but are not present in the maternal timecores. Furthermore, all infant timecores except I1 include genera not present in any other infant timecore (in pale yellow in Figure 6A), pointing towards a continuous acquisition and loss of taxa throughout succession. Finally, *Desulfovibrio* and *Dialister*, as well as 17 rare genera, are present in both the MA and MB cores but not in those of any of the infant timepoints.

We also analyzed taxon dynamics by means of abundance plots of specific genera through time (data not shown) and with a Self-Organizing Map approach (SOM) that classified genera into groups with distinct abundance profiles along development. Figure S6A shows the three clusters of distinct temporal profiles (decreasing, increasing or peaking at I3) with >80% support in a bootstrapped SOM procedure. Only 18 genera, including *Klebsiella* and 10 other Proteobacteria, significantly grouped in the decreasing profile cluster, although the individual profiles of numerous other genera, such as *Bifidobacterium*, *Citrobacter*, *Clostridium*, *Enterococcus*, *Escherichia* and *Streptococcus*, also followed decreasing trendlines. A cluster including 11 genera whose abundances significantly peaked at I3 was also recovered. These genera were all rare, even at I3. Finally, the largest cluster grouped 31 genera that significantly increased after the I3 timepoint, mainly belonging to the Firmicutes.

At the functional level, Figure 3B shows that, for the TIGRFAM main functional roles, all samples have rather similar profiles, reflecting the fact that substantial functional requirements are likely shared among the different bacterial communities. Nevertheless, chi-square tests identify highly significant differences in the distribution of all main functional roles across timepoints ($p \leq 0.001$), except for "central intermediary metabolism" ($p = 0.02$) and "unclassified proteins" ($p = 0.4$). "Protein synthesis", "transport and binding proteins" and "energy metabolism" predominate across all samples, with "protein synthesis" being the most abundant role in most cases and one or the other of the latter two roles being the most abundant in a small fraction of the

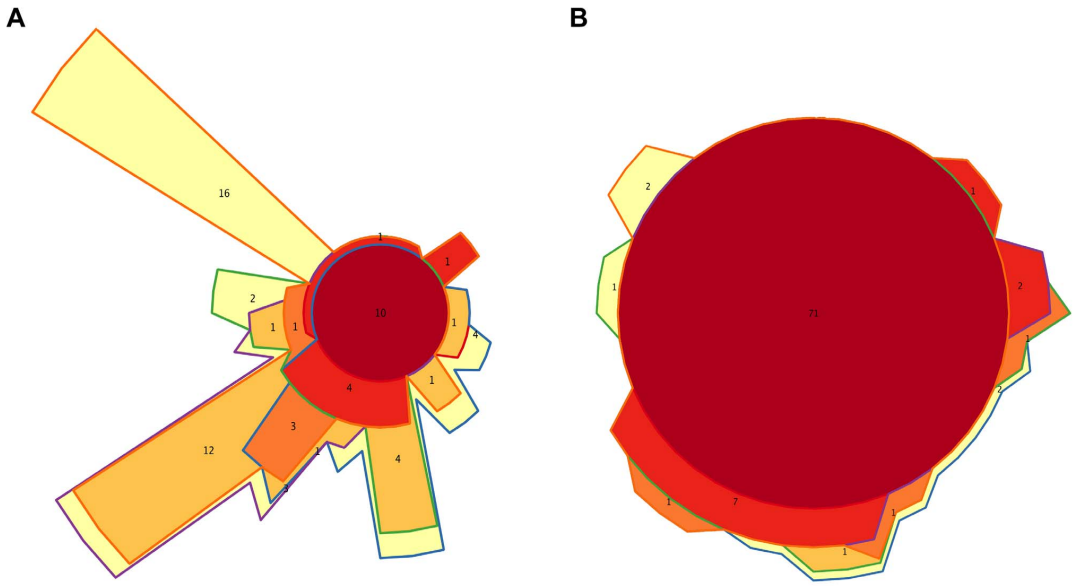


Figure 6. Timecore Venn diagrams. Changes in the core sets of genera (A) or functions (B) present at each infant timepoint. In both cases, areas representing the different timecores are enclosed by lines of the corresponding colors. The red central circles represent the genera or functions present in all five infant timecores; areas filled in dark orange, medium orange, light orange and yellow represent features present in four, three, two or one infant timecores. The number of features included in each section of the diagram is shown and areas are approximately proportional to these numbers.

doi:10.1371/journal.pgen.1004406.g006

infants' samples. Analyses at the TIGRFAM subrole level enable a better differentiation of the functional capacities present in the microbiota at different timepoints. Of the 116 subroles established in the TIGRFAMs database, 108 are detected in at least one of the samples, and 69 represent core functions detected in all. Two additional functions, "nitrogen metabolism" and "one-carbon metabolism", are only absent in some maternal samples, elevating the number of core functions present in all infants to 71. In contrast, there are no functions that are absent from all infant timecores but present in all MA or MB samples. The Venn diagram in Figure 6B displays the intersections of the different infant timecores (Table S10), showing that very few functions beyond those of the common core are present in individual infant timecores or combinations thereof. The timecore of I1 is the most reduced, but is lacking only seven functional subroles that appear in the I2 timecore and remain thereafter. These subroles are those involved in the biosynthesis of polyamines, biotin and pyridoxine, in the transport and binding of nucleosides, purines and pyrimidines, in the tricarboxylic acid cycle of aerobic metabolism, and in cellular chemotaxis and motility, as well as one of the subroles related to mobile and extrachromosomal element functions. In addition, only 12 more functional subroles are present in one or a few of the infant timecores, including "cell envelope surface structures", which is present in timecores I1 to I3, and "nitrogen fixation" and "DNA restriction/modification", which only appear in the I5 and maternal timecores.

The SOM approach also identifies a few temporal trends in the abundance dynamics of TIGRFAM subroles, although with a bootstrap support lower than that obtained for the clustering of taxonomic profiles (Figure S6B). In particular, several subroles follow a sustained decrease from I1 to I5. These include several aerobiosis-related functions, such as the biosynthesis of lipoate and heme, essential cofactors of aerobic metabolism, and the Entner-Doudoroff pathway, an alternative to glycolysis used mostly by *Enterococcus*, *Escherichia* and other Proteobacteria during aerobic conditions. The decrease in this pathway is then concordant with the taxonomic trends described above. Other decreasing subroles are related to cell envelope surface structures and to pathogenesis, although toxin production and resistance functions fluctuate throughout the year without an increasing or decreasing trend.

Potential patterns of association during community assembly based on presence/absence of taxa in diverse environments

To explore how positive and negative associations among taxa may have contributed to shape the gut's ecological succession, we investigated how the main genera detected in the infant and maternal gut microbiota relate within a network based on a wider environmental framework. We employed a previously constructed network based on presence/absence of taxa across a large variety of environments [56], the significance of which has been assessed by means of an appropriate null model (see Materials and Methods; Pascual-García A, Tamames J, Bastolla U, personal communication). For each infant and maternal timepoint, we extracted from this parent network the relationships of the timecore taxa. The subnetwork in Figure 7 represents the ensemble of these relationships for MB and all of the infant timepoints, color-coded according to whether or not they are present at MB and, for those that are, according to the first timepoint in which they appeared (see Figure Legend).

The overall topology of the subnetwork clearly delineates a central cluster populated by numerous links representing significant aggregations, surrounded by a much sparser peripheral "shell". Remarkably, the central cluster exclusively contains taxa and relations that appeared from I1 to I5 and that are also present in the MB timecore (links colored in red, blue, purple or orange), while the outer shell is mainly formed by taxa and relations restricted to the MB timecore (links colored in brown). Network theory indicates that the existence of a central and densely connected set of nodes in a network facilitates system robustness and evolvability, helping adaptation to large fluctuations of the environment and to noise of intrinsic processes [69]. Regarding the temporal assembly of this central cluster, examination of the time of appearance of the different aggregations reveals that few of them existed at I1 (in red), although *Bacteroides*, *Clostridium* and *Enterococcus* formed a transitive aggregation already at this point. Transitive aggregations, where three or more taxa are linked to one another, are highly unlikely to occur by chance and their existence suggests that the involved taxa may sustain mutualistic relationships. In addition to this main triangle, a single other aggregation appears within the central cluster at I1, linking *Prevotella* to *Bacteroides*.

Following with the assembly of the central cluster, several new aggregations are formed at I2 (in blue) enabled by the appearance of *Ruminococcus*, *Faecalibacterium* and *Collinsella*, which are linked into a triangle. In addition, *Ruminococcus* and *Faecalibacterium* form another triangle with *Eubacterium* – which was already present at the I1 timecore without being linked to other genera. These two new triangles are linked to the *Bacteroides-Clostridium-Enterococcus* triangle through a single aggregation between *Faecalibacterium* and *Bacteroides*. Remarkably, in contrast to I2, no new aggregation is formed within the central cluster, or in the surrounding shell, at the I3 timepoint. Although this difference could be influenced by the fact that only 9 samples were available for the I2 timepoint, which could artefactually inflate the I2 timecore, the same result is obtained in a subnetwork based on timecores for the 9 infants who were sampled at all timepoints. This suggests that a stable stage of community assembly had been reached in the infants' gut by one month of age, at least with respect to the core taxa of the microbiota, which was not altered during the remaining months of exclusive milk feeding.

At I4, after the introduction of solid foods, a large number of novel aggregations (in purple) are again enabled by the appearance in the timecore of several Firmicutes genera. In particular, *Dorea* establishes a large number of links at this point, including numerous triangles and several larger cliques (subgraphs in which all nodes are connected to each other) that link different Firmicutes genera, as well as a triangle formed by *Dorea*, *Faecalibacterium* and *Bacteroides*. At I5, another Clostridiales genus, *Anaerotruncus*, and two Bacteroidales, *Parabacteroides* and *Alistipes*, join the central cluster forming numerous aggregations. *Anaerotruncus* links with the Bacteroidales genera *Alistipes* and *Prevotella*, and with nearly all of the Firmicutes genera that appeared at I4, forming numerous triangles and one clique. On their part, *Parabacteroides* and *Alistipes* are also involved in several links and transitive aggregations, including a clique with *Acetivibrio* and *Bacteroides*.

It is worth noting the abundance of transitive relations that are enabled in the central cluster at I4 and I5, consolidating its structure and indicating that the introduction of solid foods to the infants' diet likely promoted an increase in the complexity of community assembly. Moreover, as already mentioned, the genera restricted to the MB timecore do not join the central cluster of the subnetwork, and rather form a surrounding "shell" that is connected with this cluster through a moderate number of aggregations. This suggests that, although community assembly was still not complete by the one-year mark, the main nucleus of the gut community was already established at this point. Interestingly, network theory indicates that core/periphery struc-

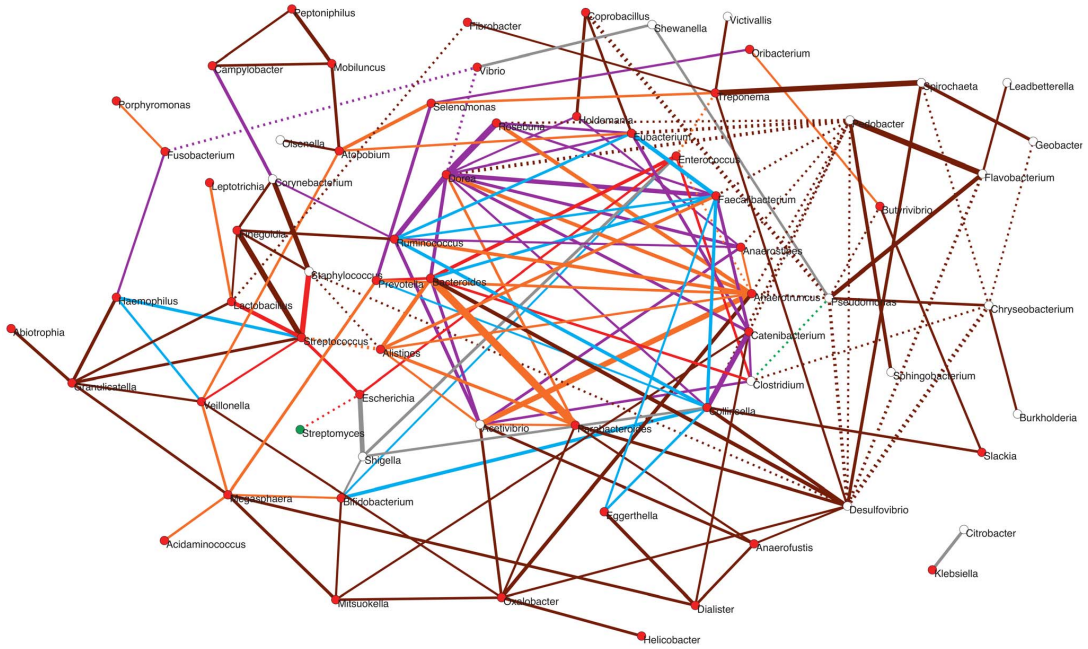


Figure 7. Potential taxon interactions during assembly of the gut microbiota. The represented subnetwork links all genera present in the different infant timecores and in the MB timecore, showing relationships inferred in a parent network based on presence/absence of taxa in multiple environments. We show with continuous lines those relations that have been identified as significant aggregations in the parent network, and with dotted lines the significant segregations. Relations are color-coded according to whether or not they are present in the maternal MB timecore and, for the relationships that are present at MB, according to the first timepoint in which they appeared. Relations that are not observed in the MB timecore are shown in grey; relations present only in MB are colored brown; relations appearing at I5 are colored orange; relations appearing at I4 are purple; relations appearing at I3 are green; relations appearing at I2 are blue; and relations appearing at I1 are red. Nodes are additionally colored according to their dominant environment in the original classification of Tamames *et al.* [56]. A dominant environment was assigned for a given genus when more than half of the samples where it was detected belonged to that environment. Red: host environments; Green: terrestrial environments; White: no particular preference for any environment (*i.e.,* cosmopolitan taxa). The thickness of the network edges represents the significance of the association (z-score).
doi:10.1371/journal.pgen.1004406.g007

tures may form at times of environmental stress, leading to the development of more condensed network structures and the segregation of a network core [69]. In the infant gut, the introduction of solid foods between I3 and I4 represented a major disturbance that must have permanently altered the conditions of the gut environment, and the resulting stress in the subsequent months may have promoted the consolidation of the central cluster.

Besides those that form the central cluster of the subnetwork, several genera appear during infant development that only connect to this cluster through a limited number of direct links or longer paths. In fact, most of the genera present at the I1 timecore are located outside of the central cluster. Remarkably, these include a small (red) star-shaped subgraph with the global-core genus *Streptococcus* at its center, which aggregates *Escherichia* and *Lactobacillus*, also members of the global core, and *Staphylococcus* and *Veillonella*, which appear at I1 but are not maintained in all infant timecores. This subgraph is only connected to the central cluster through a (red) link between *Escherichia* and *Enterococcus*. Several other peripheral genera directly join this subgraph at other timepoints. These additional genera enable a few more aggregative links with the central cluster, as well as several paths linking the subgraph to other peripheral genera, mostly appearing at I5 or MB. On the other hand, the subgraph genera are also involved in several segregative relations appearing at different timepoints, with *Streptomyces*, *Desulfovibrio*, *Fibrobacter* and the central cluster genus *Alistipes*.

Other peripheral genera appearing early on during infant development include *Bifidobacterium*, which is only connected to the central cluster through links to *Collinsella* and *Enterococcus* (in blue), and a series of Proteobacteria. Among these, we find the global core genus *Vibrio*, which never connects to the central cluster, and directly segregates from it through the genus *Dorea*. Several other peripheral genera appear at the I4 and I5 timecores, connecting to the central cluster through direct links (in purple or orange) or aggregative paths. Interestingly, these include the spirochaete *Treponema*, which is considered atypical in urban populations and had until now mostly been detected in rural populations of Africa and South America and in ancient mesoamerican remains [25,70]. In addition, two enteric Proteobacteria, *Citrobacter* and *Klebsiella*, are present as a separate component of the subnetwork, linked to each other but involved in no other relationship. *Citrobacter* and *Klebsiella* are restricted to the I2–I5 and I2–I3 timecores (Table S9), respectively, although they can also reach high abundances in individual infants at other timepoints (Figure S2). This scenario suggests that these taxa, along with *Shigella* and *Shewanella* that only

momentarily join the subnetwork, are not permanent residents of the microbiota, but perhaps opportunists that take advantage of transient conditions in the infant gut.

Regarding the genera that join the periphery of the subnetwork as part of the MB timecore, *Desulfovibrio* and *Oxalobacter* are the ones showing the larger number of aggregations to the central cluster. The aggregations detected for *Desulfovibrio* are of note, since they reflect experimentally established cross-feeding relationships of this H_2-consuming, sulfate-reducing bacterium with the H_2-producer *Collinsella* and the sulfatase-encoding Bacteroidales genera [71]. On the other hand, there is on the right side of the subnetwork a series of MB genera that clearly segregate from the central cluster, mainly through direct segregative links sustained by *Pseudomonas* and *Pedobacter*. Interestingly, most of these genera have been classified as cosmopolitan rather than mainly host-associated [56], and none of them ever reaches abundances of 1% in any sample. In addition, approximately half of these genera are absent from the MA timecore (Table S9), suggesting that they can be easily displaced when there are alterations of the gut environment. This scenario suggests that these late-appearing taxa might be facultative members of the gut microbiota or allochthonous species that frequently make their way to the gut without establishing as main components of the community.

Putting it together: Overall patterns of microbiota development delineate a successional process redirected by the introduction of solid foods

The various patterns of microbiota development described in the preceding sections suggest that during the time-course analyzed we are likely observing two major, distinct colonization phases, separated by the introduction of solid foods to the infants' diets. The first colonization phase would encompass the period during which infants were fed only milk, *i. e.*, timepoints I1 to I3. During this period, the richness, diversity and complexity of interactions among taxa tend to increase in I1–I2, indicating that the relatively simple bacterial communities present by one week can tolerate the arrival and establishment of new species, to which the infants would undoubtedly be exposed during their first weeks of life. The variable behavior of taxon richness and diversity observed across infants during the I2–I3 period suggests that, by three months of age, different infants were at different stages of community development, with some still incorporating new species while others were starting to loose species, most likely due to interspecific competition. In the infants that underwent decreases in richness during this period, changes in Shannon values were not concomitant (Table S6), supporting the notion that interspecific competition purged the community of rare taxa, presumably not well adapted to thrive during this milk-feeding period. Accordingly, *Bifidobacterium* or *Bacteroides*, the only genera capable of thriving on both lactose and HMOs [59], dominated the I3 microbiota in nearly all infants, with the exception of those born by C-section (Figure S2), which supports the notion that C-section delays the establishment of these genera [51]. Moreover, among vaginally delivered infants, *Bacteroides* or *Bifidobacterium* dominated when mothers did or did not receive antibiotics during delivery, respectively. The I3 CCA (Figure S3) confirms that delivery type and use of peripartum antibiotics explain 22% and 12% of the taxonomic composition variation of the infant microbiota at this timepoint (Table S5).

Classical models of succession posit that, after a period of competition leading to species loss, community stability will eventually increase in late successional stages, after which major community shifts will not occur unless a significant disturbance affects the ecosystem [52]. In our data, no such stabilization is observed, as the variable I2–I3 period is followed by a strong decrease in richness in I3–I4, followed by a trend towards richness recovery in I4–I5 (Table S6, Figure 1A, 1C). As stated earlier, these richness changes are mainly due to the loss and gain of rare genera, and are accompanied by opposite trends at the level of diversity (Table S6, Figure 1B, 1D). In addition, the number of core microbiota genera shared by all individuals increases importantly at I4 and I5 (Table S9, Figure 6A), with substantial repercussions on the configuration of relationships among taxa (Figure 7). Most likely, the introduction of solid foods between I3 and I4 contributed importantly to prevent the stabilization of the community, as this chronic disturbance altered the resources present in the gut environment. With solid foods, the variety of nutrients that become available to the infant gut microbiota clearly expands, potentially providing a larger number of niches for different organisms and contributing to the increase in diversity observed at I4. In particular, carbohydrates will now be available in a larger variety of forms, including numerous complex molecules found in cereals, fruits, vegetables and tubers (Table S1), providing a selective challenge for the milk adapted resident community.

Our observation that solid food introduction is followed by a purge in rare taxa is consistent with the idea that fewer species will persist in the face of intense disturbances [72]. In the 7-months infant, the genera that thrived in the milk-adapted microbiota - *i. e.*, *Bifidobacterium* and *Bacteroides* - continue to dominate, with the latter genus being now the most abundant in a majority of individuals. The rise of *Bacteroides* following the introduction of solid foods has been observed in previous studies [34] and is likely due to its large versatility for complex carbohydrate degradation. Nevertheless, some genera that had not been previously detected at high abundances (or only in very few individuals) expand now in the gut microbiota, in agreement with the notion that disturbance should facilitate invasion of the community by new species [73]. This is the case of *Ruminococcus*, which is now found among the most frequent genera in nearly all infants. *Ruminococcus* thrives on oligosaccharides such as raffinose and sucrose that constitute the most abundant soluble saccharides in plant tissues and is capable of partially degrading insoluble plant fibers such as lignin and cellulose [74], which likely explains its competitive advantage after the introduction of cereals, fruits and vegetables into the diet. Another genus that reaches high abundances for the first time in some 7-months-olds is *Akkermansia*, one of the main mucin-degraders in the gut microbiota [60,75]. Mucin production is dependent on the availability of dietary amino acids and should increase with the higher protein content of solid foods, enabling the growth of mucin-specialized bacteria. On the other hand, the disturbance created by solid foods does not seem to enable invasion of the gut community by opportunistic species, as fast growers such as *Escherichia* rather decrease in abundance from I3 to I4. Moreover, the "pathogenesis" functional subrole also decreases markedly after the I3 timepoint, indicating that opportunistic pathogens are not taking advantage of the disturbance.

In the last time interval analyzed, I4–I5, taxon richness tends to increase again mainly due to the acquisition of new rare taxa. This indicates that succession has now entered a second period of net species recruitment, although most incoming taxa have not been able to reach substantial frequencies, suggesting that the pre-established populations retain a competitive advantage. Nevertheless, substantial shifts occur during this period in relative taxon abundances. Several of the most abundant genera at I4 - *i. e.*, *Bifidobacterium*, *Veillonella*, *Escherichia* - decrease substantially in I5. At the same time, the main butyrate producers of the gut

microbiota, *i. e.*, *Faecalibacterium*, *Eubacterium* and *Roseburia*, rise in abundance at this timepoint, with *Faecalibacterium* becoming the second most abundant genus overall (after *Bacteroides*). In addition, other SCFA producers, such as *Blautia* and *Butyrivibrio*, reach frequencies above 1% for the first time in I5. Between I4 and I5, the diet of Spanish Mediterranean infants changes substantially, as it becomes progressively similar to that of adults [36]. During this period the general consumption of animal protein increases importantly, as meats, fish, eggs and dairy products become more prevalent (Table S1). At the same time, the contribution of cereals continues to increase, probably enabling the rise of genera adept at fermenting starches and fiber, such as *Bacteroides* and the butyrate producers *Faecalibacterium*, *Eubacterium* and *Roseburia*.

As a result of these changes, the ranking of taxon abundances observed in the one-year-old infants becomes remarkably similar to that of the mothers, with *Bacteroides*, *Faecalibacterium*, *Clostridium* and *Ruminococcus* present among the five top genera in I5, MA and MB (Figure S2). However, differences exist in the relative abundances of *Bifidobacterium* and *Eubacterium* between mothers and one-year-olds, with the first genus remaining more common in I5 while the latter has not yet reached the high levels at which it is found in MA and MB. Moreover, the richness (Figure 1C), diversity (Figure 1D) and complexity of interactions among taxa (Figure 7) at the one-year mark are still far from those observed in the maternal samples. Similarly, pairwise ANOSIM analyses (Table S7) and ordination techniques (Figure 4) detect differences in taxonomic and functional composition between I5 and the maternal samples, further corroborating that succession was incomplete at the one-year mark. In agreement, recent cross-sectional studies have suggested that an adult-like gut community may not be reached before three years of age [25].

In conclusion, our analyses of GIT microbiota development during the first year of life reveal an incomplete successional process, strongly marked by the introduction of solid foods to the infants' diets. Therefore, important questions regarding microbial succession in the infant GIT still remain for further analysis. A longer sampling period would be necessary to reveal the final progression of the gut microbiota towards an adult-like stage, and a tighter sampling around the time of introduction of solid foods would be required to clarify the transition that accompanies this event. On the other hand, in order to gain an in depth understanding of the ecological and evolutionary processes at play in this environment, we will need to focus on the genetic structure and demographic dynamics of microbial populations as they settle within the gut.

Materials and Methods

Ethics statement

This study was approved by the Ethics Committee of the Center for Public Health Research (CSISP), Valencia, Spain. All women participating in the study read and signed forms of informed consent specifically approved for this project by the Ethics Committee.

Sample collection, pyrosequencing and initial processing of sequencing reads

Fecal samples were collected by the mothers and stored in home freezers until brought to the laboratory, where they were stored at −80°C until processing. Samples were homogenized in a 50% RNA later/phosphate saline buffer solution and centrifuged for two minutes at 2000 rpm. Only the supernatant resulting from the latter spin was used for further processing. DNA was extracted using the Epicenter Master Pure Complete DNA & RNA Purification kit following manufacturer's specifications, except for an additional digestion step at the beginning of the extraction protocol with lysozyme for 30 minutes at 37°. Samples were then prepared for 454 pyrosequencing by adding a barcode and pooling them in groups of 20 samples per run, which provided between 35000 and 70000 reads per sample. Only reads that passed quality controls (average base score quality per read >20) were further analyzed after elimination of read replicates by means of CD-HIT-454 [76]. We addressed downstream analysis at read level rather than at contig level based on the prior assessing of the complexity of our communities, as simulation studies have determined that chimeras are particularly prevalent among contigs lower than 10 kbp in size [77,78]. High-complexity microbial communities lacking dominant populations rarely produce contigs larger than 10 kbp, prompting the recommendation that such data sets should not be assembled at all.

Gene calling, taxonomic assignment and functional annotation

We used a combination of evidence-based and *ab initio* gene calling. In the first step, coding regions were identified based on homology searches at read level via BLASTX [79] against the NCBI-nr protein database considering an e-value cutoff of 0.001. Subsequently, we used GLIMMER3 [80] to identify any coding regions that were missed in the previous step by means of a fine-tuned IMM (Interpolated Markov Model). We used the '-X' GLIMMER3 option, allowing fragmented ORF (Open Reading Frame) identification, and default settings for other options. In order to build the IMM we chose eight complete bacterial genomes from NCBI spanning the main gut microbiome phyla (Firmicutes, Bacteroidetes, Actinobacteria and Proteobacteria) and then extracted the reported ORFs to train the model.

Taxonomic classification was only performed on coding regions found by BLASTX, by using Blast2.lca (https://github.com/emepyc/Blast2lca#readme). This methodology is based on a Last Common Ancestor (LCA) algorithm, which retrieves the most specific taxon associated with the complete set of sequences that hit a certain query, instead of only considering the taxon associated with the closest BLASTX hit, thereby reducing false matches. Eukaryota-related coding regions were filtered out from the analysis based on superkingdom LCA annotation, or on BLASTN searches (0.001 e-value cutoff) against the NCBI-nt eukaryotic subset in the case of those regions identified by the *ab initio* approach. Finally, to functionally annotate the identified coding regions we used HMMER2 [81] against the TIGRFAMs (9.0 release) database of prokaryotic models [82], considering an e-value cutoff of 0.1. HMMER is a protein profile aligner based on hidden Markov models, with high sensitivity for classifying remote homologs [83].

Microbiota richness and diversity

We assessed the taxonomic and functional richness and diversity of the microbiota by means of several estimators. In order to eliminate possible artifacts introduced by read count differences between samples, we first used QIIME [84] for resampling an equal number of reads per sample. The richness estimators N and Chao1 [63] and the Shannon diversity index [64] were then calculated using the library 'vegan' from the R package [85]. The Chao1 estimator was chosen because it has been shown to be one of the most reliable non-parametric estimators of species richness in species-rich samples [86]. The Shannon index was preferred for species diversity because of its use of natural logarithms of relative species abundances, which reduces the weight of the more

abundant species and renders it sensitive to the changes in rare species, which are common in infant gut microbiota samples.

Linear regression analyses were executed to determine the statistical significance of the changes in richness and diversity through time, both over all timepoints and in pairwise comparisons for specific time intervals. Because taxon richness is assumed to follow a Poisson distribution, we employed the 'glm' function implemented in the 'stats' R package to fit generalized linear models. On the other hand, values of the Shannon diversity index were parameterized in the standard unit interval (0, 1) and assumed to follow a Beta distribution; therefore we applied the Beta regression model, as implemented in the 'betareg' [87] R package. Hierarchical clustering analysis of the temporal profiles of richness and diversity change for individual infants was performed in the 'pvclust' R package, which assesses clustering uncertainty by means of multiscale bootstrap resampling [88].

Microbiota composition clustering, directionality and dynamics

The R package was employed for comparative analyses of taxonomic and functional microbiota composition. Heatmaps and clustering analyses were based on the Bray-Curtis distance as a measure of dissimilarity [65]. Directionality in taxonomic and functional composition change through time was assessed by means of various multivariate analyses. First, we employed global and pairwise analyses of similarities (ANOSIM) adjusted for multiple testing to detect whether there were significant differences between taxonomic or functional profiles per timepoint. ANOSIM tests whether there is a significant difference between two or more groups of samples by comparing distances between sample groups to those within groups. In addition, we also performed "Permutational Multivariate Analysis of Variance Using Distance Matrices" (PMANOVA or ADONIS), which yielded similar results to the ANOSIM (data not shown). Both analyses used the Bray-Curtis distance to measure dissimilarity in taxonomic or functional microbiota composition between samples. To explore further the pattern of similarities among timepoints we performed Canonical Correspondence Analysis (CCA) and Principal Coordinates Analysis (PCoA) using Gower distances [67], for both taxonomic and functional data sets. Once the PCoA analyses were executed, we drew convex hulls enclosing all samples pertaining to a particular timepoint and calculated the area of overlap of the polygons representing each timepoint.

The dynamics of individual genera and functions through time were also examined within R. The behavior of different genera was analyzed by means of regression analyses using the Poisson model and the 'GeneFamilies.regression' function from the 'ShotgunFunctionalizeR' library, and also through the drawing of Venn diagrams containing the taxa per individual, MIP or timepoint, using the 'venn' function in the 'gplots' library. Venn diagrams were also constructed to identify taxonomic and functional "timecores" containing the taxa or functions shared across all individuals at a given timepoint using the 'compute.-Venn' function in the 'Venerable' library, and to identify those features restricted to single timecores or combinations thereof. Self-Organizing Maps (SOM) [89] were constructed for both taxonomic and functional data sets, using the function 'som' from the 'som' library. These maps are artificial neural networks that use a neighborhood function to separate a complex, high-dimensional input space into a reduced number of discrete groups with unique behaviors through time. In order to get reliable SOM-based clusters we used the bootstrap method. Firstly, we built 200 different sets of resampled temporal profiles for each feature (genus or function) by resampling entire profiles of randomly selected individuals. Then, we carried out a SOM-based clustering over this 200-fold-sized data set. To build clusters at different support levels, we retrieved only those features whose profiles were classified into the same cluster in at least 60% or 80% of the resampling sets.

Constructing a dendrogram of genus-level functional profiles

Functional profiles were determined for those genera present in any sample at >1% abundance in addition to having genes representing at least 50% of the 108 TIGRFAM functional subroles detected in our complete dataset. Because not enough information was recovered in a sample per sample basis for each genus, the functional profile was established by pooling all the samples of a timepoint. Functional profiles were defined as vectors containing the relative abundances of each one of the 108 TIGRFAM subroles in a particular genus and timepoint. Bray-Curtis distances between functional profiles were computed using the 'bcdist' function from the R 'ecodist' library, and dendrograms based on these distances were drawn using the 'hclust' function from the R 'stats' library with the complete-linkage method.

Extracting gut microbiota taxa co-occurrence networks from a parent network based on diverse environments

We analyzed the relations of the main gut microbiota genera detected in our study within a parent network previously constructed based on the presence/abscence of taxa across a large variety of environments (Pascual-García A, Tamames J, Bastolla U., personal communication). For each infant and maternal timepoint, we considered the group of N taxa observed in all samples of the timepoint (the timecore). For each group we had then $N(N-1)/2$ putative interactions and we determined those that were present in the parent network, which includes all significant associations among 1187 different genera observed in 2322 samples from very different environments. Details about the environments and their classification can be found in [56]. The parent network was obtained from an adaptation of the null model proposed by Navarro-Alberto and Manly [90] where environmental preferences are considered in order to avoid trivial associations. The null model allows for the generation of random realizations of the original data assuming that taxa are not associated. The significance of putative associations can then be assessed by comparing the results obtained from the observed data *versus* those obtained from the random ensemble. As the random realizations do not contain information about real associations, any signal coming from the random ensemble is considered a false positive, serving to establish a restrictive threshold for the estimated false positive rate.

Supporting Information

Figure S1 Heatmaps and clustering of MA and MB maternal samples according to taxonomic composition (A) and main TIGRFAM functional roles (B) based on Bray-Curtis distances. (A) Only the genera above 1% abundance in at least one sample are depicted. Each sample is identified at the bottom of the heatmaps by a code that specifies the MIP (Mother Infant Pair) to which it belongs and the corresponding timepoint.
(PDF)

Figure S2 Heatmaps and clustering of the samples for each timepoint according to taxonomic composition (A) and TIGRFAM main functional roles (B) (details as in Figure S1).
(PDF)

Figure S3 Canonical Correspondence Analyses (CCA) showing the effect of C-section on the taxonomic composition of the microbiota at different timepoints. The proportion of variability explained by C-section delivery is highest at I1 (16%), I2 (22%) and I3 (22%) and decreases at I4 (10%) and I5 (10%), and is always below the proportion of variability explained by the first unconstrained axis.
(PDF)

Figure S4 Linear regressions of richness (Chao1 estimator) and diversity (Shannon index) *vs.* time (A–B taxonomy, C–D function).
(PDF)

Figure S5 Heatmaps and clustering of the samples for each MIP according to taxonomic composition (A) and TIGRFAM main functional roles (B) (details as in Figure S1).
(PDF)

Figure S6 Self-Organizing Maps (SOM) of taxon and function dynamics. SOMs identify patterns of abundance dynamics in the infants throughout development at both taxonomic (A) and functional (B) levels. The number of genera (A) or functions (B) included in each represented cluster is indicated (cluster size). Clusters have 80% and 60% bootstrap support for taxa and functions, respectively. For each cluster, average values on each timepoint along with their corresponding 95% confidence intervals are shown, in a scale centered at the mean of all samples and scaled by the standard deviation.
(PDF)

Table S1 Information on consumption of different foods, obtained from questionnaires answered by the infants' parents.
(DOCX)

Table S2 Details of pyrosequencing reads and annotation per individual sample.
(DOCX)

Table S3 Taxon abundances per sample. Numbers correspond to raw sequence counts.
(TXT)

Table S4 Function abundances per sample. Numbers correspond to raw sequence counts.
(TXT)

Table S5 Variability explained by constrained (CCA1) and unconstrained (CA1 and CA2) axes in Canonical Correspondence Analyses when the constraining variable is delivery type, use of peripartum antibiotic or exclusivity of breastfeeding.
(DOCX)

Table S6 Richness (N and Chao1 estimator) and diversity (Shannon index) for taxonomic and functional data in individual samples.
(DOCX)

Table S7 p-values of ANOSIM pairwise comparisons between timepoints. Statistically significant values ($p<0.05$) are shown in red.
(DOCX)

Table S8 Taxonomic and functional dissimilarities between timepoints estimated as the non-overlapping areas of the convex hulls representing them in the PCoAs of Figure 4C–D. Dissimilarity values above 0.80 are shown in red.
(DOCX)

Table S9 Taxonomic timecores. Timecores are defined as lists of genera present in all individuals at a given timepoint. All genera present in at least one timecore are listed and their presence (1) or absence (0) at each timecore is reported.
(TXT)

Table S10 Functional timecores. Timecores are defined as lists of functions present in all individuals at a given timepoint. All functions present in at least one timecore are listed and their presence (1) or absence (0) at each timecore is reported.
(TXT)

Acknowledgments

We thank Nuria Jiménez and Llúcia Martínez for performing the pyrosequencing reported in this paper at the sequencing facility of the Unidad Mixta de Investigación en Genómica y Salud, FISABIO-Salud Pública. We also thank the midwives who collaborated in the recruitment of study volunteers as well as all the mothers and babies who generously participated in the study.

Author Contributions

Conceived and designed the experiments: YV MPF. Performed the experiments: YV MLF MJG. Analyzed the data: YV AA APG JJA MPF. Contributed reagents/materials/analysis tools: YV AA APG MJG JJA MPF. Wrote the paper: YV AA APG MPF.

References

1. Sekirov I, Russell SL, Antunes LC, Finlay BB (2010) Gut microbiota in health and disease. Physiol Rev 90: 859–904.
2. Collado MC, D'Auria G, Mira A, Francino MP (2013) Human Microbiome and Diseases: A Metagenomic Approach. In: Watson RR and Preedy VR, editors. Bioactive Food as Dietary Interventions for Liver and Gastrointestinal Disease. San Diego: Academic Press. pp. 235–249.
3. Baas-Becking L (1934) Geobiologie of inleiding tot de milieukunde. Van Stockum & Zoon. 263 p.
4. de Wit R, Bouvier T (2006) 'Everything is everywhere, but, the environment selects'; what did Baas Becking and Beijerinck really say? Environ Microbiol 8: 755–758.
5. McConnell EL, Basit AW, Murdan S (2008) Measurements of rat and mouse gastrointestinal pH, fluid and lymphoid tissue, and implications for in-vivo experiments. J Pharm Pharmacol 60: 63–70.
6. Palmer C, Bik EM, DiGiulio DB, Relman DA, Brown PO (2007) Development of the human infant intestinal microbiota. PLoS Biol 5: e177.
7. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, et al. (2007) The human microbiome project. Nature 449: 804–810.
8. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, et al. (2009) A core gut microbiome in obese and lean twins. Nature 457: 480–484.
9. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, et al. (2010) A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464: 59–65.
10. Jalanka-Tuovinen J, Salonen A, Nikkila J, Immonen O, Kekkonen R, et al. (2011) Intestinal microbiota in healthy adults: temporal analysis reveals individual and common core and relation to intestinal symptoms. PLoS One 6: e23035.
11. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, et al. (2011) Enterotypes of the human gut microbiome. Nature 473: 174–180.
12. Koren O, Knights D, Gonzalez A, Waldron L, Segata N, et al. (2013) A guide to enterotypes across the human body: meta-analysis of microbial community structures in human microbiome datasets. PLoS Comput Biol 9: e1002863.
13. Hubbell SP (2001) The unified neutral theory of biodiversity and biogeography. Princeton, New Jersey, USA: Princeton University Press.
14. Gilpin M, editor(1991) Metapopulation Dynamics: Empirical and Theoretical Investigations. London: Academic Press.
15. Hubbell SP (2006) Neutral theory and the evolution of ecological equivalence. Ecology 87: 1387–1398.
16. Vaishampayan PA, Kuehl JV, Froula JL, Morgan JL, Ochman H, et al. (2010) Comparative metagenomics and population dynamics of the gut microbiota in mother and infant. Genome Biol Evol 2: 53–66.
17. Gosalbes MJ, Durban A, Pignatelli M, Abellan JJ, Jimenez-Hernandez N, et al. (2011) Metatranscriptomic approach to analyze the functional human gut microbiota. PLoS One 6: e17447.
18. Valles Y, Gosalbes MJ, de Vries LE, Abellan JJ, Francino MP (2012) Metagenomics and development of the gut microbiota in infants. Clin Microbiol Infect 18 Suppl 4: 21–26.
19. Petchey OL, Gaston KJ (2006) Functional diversity: back to basics and looking forward. Ecol Lett 9: 741–758.
20. McGill BJ, Enquist BJ, Weiher E, Westoby M (2006) Rebuilding community ecology from functional traits. Trends Ecol Evol 21: 178–185.

21. Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, et al. (2005) Comparative metagenomics of microbial communities. Science 308: 554–557.
22. Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, et al. (2006) Metagenomic analysis of the human distal gut microbiome. Science 312: 1355–1359.
23. Kurokawa K, Itoh T, Kuwahara T, Oshima K, Toh H, et al. (2007) Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. DNA Res 14: 169–181.
24. HMP Consortium (2012) Structure, function and diversity of the healthy human microbiome. Nature 486: 207–214.
25. Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, et al. (2012) Human gut microbiome viewed across age and geography. Nature 486: 222–227.
26. Rautava S, Ruuskanen O, Ouwehand A, Salminen S, Isolauri E (2004) The hygiene hypothesis of atopic disease–an extended version. J Pediatr Gastroenterol Nutr 38: 378–388.
27. Noverr MC, Huffnagle GB (2005) The 'microflora hypothesis' of allergic diseases. Clin Exp Allergy 35: 1511–1520.
28. Penders J, Stobberingh EE, van den Brandt PA, Thijs C (2007) The role of the intestinal microbiota in the development of atopic disorders. Allergy 62: 1223–1236.
29. Dominguez-Bello MG, Costello EK, Contreras M, Magris M, Hidalgo G, et al. (2010) Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. Proc Natl Acad Sci U S A 107: 11971–11975.
30. Mackie RI, Sghir A, Gaskins HR (1999) Developmental microbial ecology of the neonatal gastrointestinal tract. Am J Clin Nutr 69: 1035S–1045S.
31. Biasucci G, Benenati B, Morelli L, Bessi E, Boehm G (2008) Cesarean delivery may affect the early biodiversity of intestinal bacteria. J Nutr 138: 1796S–1800S.
32. Azad MB, Konya T, Maughan H, Guttman DS, Field CJ, et al. (2013) Gut microbiota of healthy Canadian infants: profiles by mode of delivery and infant diet at 4 months. CMAJ 185: 385–394.
33. Avershina E, Storro O, Oien T, Johnsen R, Pope P, et al. (2014) Major faecal microbiota shifts in composition and diversity with age in a geographically restricted cohort of mothers and their children. FEMS Microbiol Ecol 87: 280–290.
34. Favier CF, Vaughan EE, De Vos WM, Akkermans AD (2002) Molecular monitoring of succession of bacterial communities in human neonates. Appl Environ Microbiol 68: 219–226.
35. Koenig JE, Spor A, Scalfone N, Fricker AD, Stombaugh J, et al. (2011) Succession of microbial consortia in the developing infant gut microbiome. Proc Natl Acad Sci U S A 108 Suppl 1: 4578–4585.
36. Capdevila F, Vizmanos B, Marti-Henneberg C (1998) Implications of the weaning pattern on macronutrient intake, food volume and energy density in non-breastfed infants during the first year of life. J Am Coll Nutr 17: 256–262.
37. Markowitz VM, Chen IM, Chu K, Szeto E, Palaniappan K, et al. (2012) IMG/M: the integrated metagenome data management and comparative analysis system. Nucleic Acids Res 40: D123–129.
38. Koren O, Goodrich JK, Cullender TC, Spor A, Laitinen K, et al. (2012) Host remodeling of the gut microbiome and metabolic changes during pregnancy. Cell 150: 470–480.
39. Jimenez E, Fernandez L, Marin ML, Martin R, Odriozola JM, et al. (2005) Isolation of commensal bacteria from umbilical cord blood of healthy neonates born by cesarean section. Curr Microbiol 51: 270–274.
40. Jimenez E, Marin ML, Martin R, Odriozola JM, Olivares M, et al. (2008) Is meconium from healthy newborns actually sterile? Res Microbiol 159: 187–193.
41. Gosalbes MJ, Llop S, Valles Y, Moya A, Ballester F, et al. (2013) Meconium microbiota types dominated by lactic acid or enteric bacteria are differentially associated with maternal eczema and respiratory problems in infants. Clin Exp Allergy 43: 198–211.
42. Mshvildadze M, Neu J, Shuster J, Theriaque D, Li N, et al. (2010) Intestinal microbial ecology in premature infants assessed with non-culture-based techniques. J Pediatr 156: 20–25.
43. Madan JC, Salari RC, Saxena D, Davidson L, O'Toole GA, et al. (2012) Gut microbial colonisation in premature neonates predicts neonatal sepsis. Arch Dis Child Fetal Neonatal Ed 97: F456–462.
44. Moles L, Gomez M, Heilig H, Bustos G, Fuentes S, et al. (2013) Bacterial diversity in meconium of preterm neonates and evolution of their fecal microbiota during the first month of life. PLoS One 8: e66986.
45. Mor G, Cardenas I (2010) The immune system in pregnancy: a unique complexity. Am J Reprod Immunol 63: 425–433.
46. Newbern D, Freemark M (2011) Placental hormones and the control of maternal metabolism and fetal growth. Curr Opin Endocrinol Diabetes Obes 18: 409–416.
47. Perez PF, Dore J, Leclerc M, Levenez F, Benyacoub J, et al. (2007) Bacterial imprinting of the neonatal immune system: lessons from maternal cells? Pediatrics 119: e724–732.
48. Donnet-Hughes A, Perez PF, Dore J, Leclerc M, Levenez F, et al. (2010) Potential role of the intestinal microbiota of the mother in neonatal immune education. Proc Nutr Soc 69: 407–415.
49. Amar J, Chabo C, Waget A, Klopp P, Vachoux C, et al. (2011) Intestinal mucosal adherence and translocation of commensal bacteria at the early onset of type 2 diabetes: molecular mechanisms and probiotic treatment. EMBO Mol Med 3: 559–572.
50. Cabrera-Rubio R, Collado M, Laitinen K, Salminen S, Isolauri E, et al. (2012) The human milk microbiome changes over lactation and is shaped by maternal weight and mode delivery. Am J Clin Nutr 96: 544–551.
51. Penders J, Thijs C, Vink C, Stelma FF, Snijders B, et al. (2006) Factors influencing the composition of the intestinal microbiota in early infancy. Pediatrics 118: 511–521.
52. Connell JH, Slatyer RO (1977) Mechanims of Succession in Natural Communities and Their Role in Community Stability and Organization. The American Naturalist 111: 1119–1144.
53. Fierer N, Nemergut D, Knight R, Craine JM (2010) Changes through time: integrating microorganisms into the study of succession. Res Microbiol 161: 635–642.
54. Ng SK, Hamilton IR (1971) Lactate metabolism by *Veillonella parvula*. J Bacteriol 105: 999–1005.
55. Rocha EP (2004) Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization. Genome Res 14: 2279–2286.
56. Tamames J, Abellan JJ, Pignatelli M, Camacho A, Moya A (2010) Environmental distribution of prokaryotic taxa. BMC Microbiol 10: 85.
57. Stark PL, Lee A (1982) The bacterial colonization of the large bowel of pre-term low birth weight neonates. J Hyg (Lond) 89: 59–67.
58. Kunz C, Rudloff S, Baier W, Klein N, Strobel S (2000) Oligosaccharides in human milk: structural, functional, and metabolic aspects. Annu Rev Nutr 20: 699–722.
59. Marcobal A, Barboza M, Froehlich JW, Block DE, German JB, et al. (2010) Consumption of human milk oligosaccharides by gut-related microbes. J Agric Food Chem 58: 5334–5340.
60. Berry D, Stecher B, Schintlmeister A, Reichert J, Brugiroux S, et al. (2013) Host-compound foraging by intestinal microbiota revealed by single-cell stable isotope probing. Proc Natl Acad Sci U S A 110: 4720–4725.
61. McGuckin MA, Linden SK, Sutton P, Florin TH (2011) Mucin dynamics and enteric pathogens. Nat Rev Microbiol 9: 265–278.
62. Rogosa M (1964) The Genus Veillonella. I. General Cultural, Ecological, and Biochemical Considerations. J Bacteriol 87: 162–170.
63. Chao A (1984) Nonparametric estimation of the number of classes in a population. Scandinavian Journal of Statistics 11: 256–270.
64. Shannon CE (1948) A mathematical theory of communication. Bell System Technical Journal 27: 379–656.
65. Bray JR, Curtis J.T. (1957) An ordination of upland forest communities of southern Wisconsin. Ecological Monographs 27: 325–349.
66. Clements FE (1916) Plant succession: an analysis of the development of vegetation. Washington: Carnegie Institution of Washington.
67. Gower JC (1966) Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 53: 325–338.
68. de Berg M, van Kreveld M, Overmars M, Schwarzkopf O (2000) Computational Geometry: Algorithms and Applications. Springer-Verlag.
69. Csermely P, Korcsmaros T, Kiss HJ, London G, Nussinov R (2013) Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. Pharmacol Ther 138: 333–408.
70. De Filippo C, Cavalieri D, Di Paola M, Ramazzotti M, Poullet JB, et al. (2010) Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. Proc Natl Acad Sci U S A 107: 14691–14696.
71. Rey FE, Gonzalez MD, Cheng J, Wu M, Ahern PP, et al. (2013) Metabolic niche of a prominent sulfate-reducing human gut bacterium. Proc Natl Acad Sci U S A 110: 13582–13587.
72. Mackey RL, Currie DJ (2001) The Diversity-Disturbance Relationship: Is It Generally Strong and Peaked? Ecology 62: 3479–3492.
73. Hobbs R, Huenneke LF (1992) Disturbance, Diversity, and Invasion: Implications for Conservation. Conservation Biology 6: 324–337.
74. Cervera-Tison M, Tailford LE, Fuell C, Bruel L, Sulzenbacher G, et al. (2012) Functional analysis of family GH36 alpha-galactosidases from Ruminococcus gnavus E1: insights into the metabolism of a plant oligosaccharide by a human gut symbiont. Appl Environ Microbiol 78: 7720–7732.
75. Derrien M, van Passel MW, van de Bovenkamp JH, Schipper RG, de Vos WM, et al. (2010) Mucin-bacterial interactions in the human oral cavity and digestive tract. Gut Microbes 1: 254–268.
76. Niu B, Fu L, Sun S, Li W (2010) Artificial and natural duplicates in pyrosequencing reads of metagenomic data. BMC Bioinformatics 11: 187.
77. Mavromatis K, Ivanova N, Barry K, Shapiro H, Goltsman E, et al. (2007) Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. Nat Methods 4: 495–500.
78. Pignatelli M, Moya A (2011) Evaluating the fidelity of de novo short read metagenomic assembly using simulated data. PLoS One 6: e19984.
79. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403–410.
80. Delcher AL, Bratke KA, Powers EC, Salzberg SL (2007) Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics 23: 673–679.
81. Finn RD, Clements J, Eddy SR (2011) HMMER web server: interactive sequence similarity searching. Nucleic Acids Res 39: W29–37.
82. Haft DH, Selengut JD, White O (2003) The TIGRFAMs database of protein families. Nucleic Acids Res 31: 371–373.
83. Karplus K, Barrett C, Hughey R (1998) Hidden Markov models for detecting remote protein homologies. Bioinformatics 14: 846–856.

84. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, et al. (2010) QIIME allows analysis of high-throughput community sequencing data. Nat Methods 7: 335–336.
85. R Development Core Team (2010) R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.
86. Basualdo C (2011) Choosing the best non-parametric richness estimator for benthic macroinvertebrates databases. Rev Soc Entomol Argent 70: 27–38.
87. Grün B, Kosmidis I, Zeileis A (2012) Extended Beta Regression in R: Shaken, Stirred, Mixed, and Partitioned. Journal of Statistical Software 48: 1–25.
88. Shimodaira H (2004) Approximately unbiased tests of regions using multistep-multiscale bootstrap resampling. The Annals of Statistics 32: 2343–2775.
89. Kohonen T (1982) Self-Organized Formation of Topologically Correct Feature Maps. Biological Cybernetics 43: 59–69.
90. Navarro-Alberto J, Manly BFJ (2009) Null model analyses of presence-absence matrices need a definition of independence. Popul Ecol 51: 505–512.

Chapter 4

Mutualistic systems

Scientists too often know little about the cultural and historical context of their ideas (...) They rarely acknowledge that their theoretical frames derive from an Anglophone-capitalist model (...) Concepts such as the validity of cost-benefit and competition vs. co-operation terminology, or the superiority of mathematical analysis are uncritically assumed.

Lynn Margulis

Summary

In the results shown for microbial systems, we inferred interactions between microbes. For proteins, we represented the interactions between amino-acids with contacts'matrices, and we developed a procedure to identify global patterns of contacts shared between proteins. In this section we explore mutualistic systems, where we know from experiments the interactions between plants and animals. Our approximation in this case goes an step further modeling the dynamical behaviour of the sytem, thus providing more direct answers to the questions we investigate.

In our work, we consider a dynamical model with within-group competition of Lotka-Volterra type and between-group mutualistic interactions that saturate for large abundance [**?**]. We started reconsidering the assumptions typically made in stability analysis. It is known that, when the interaction matrix has the mathematical property called diagonal stability [Goh (1979)], every feasible equilibrium where all species have positive abundances is dynamically stable. Furthermore, this equilibrium is globally stable, and we note that, in this case, the question about stability translates into the question

of the parameters that guarantee positive abundances, i.e. in its structural stability. In this way, we move from the classical dynamical stability analysis, where the abundances are perturbed for fixed model parameters, to an analysis where the growth rates are perturbed for a given feasible equilibrium.

Analytically, the structural stability analysis is performed after transforming the system. As previously stated, we consider non-linear terms in the mutualistic interactions to avoid divergences in the populations, and we require to transform the system into an effective Lotka-Volterra model. This means that we approximate the non-linear behaviour of the system considering an equivalent linear model, which further allows us to get analytical insight. Indeed, from the fixed point equations of this effective model, it is possible to reduce a system which contains arbitrary interactions into an effective competition system [Bastolla et al. (2005, 2009)]. This transformation facilitates the interpretation of the effects of the different interactions just in terms of competitive interactions –relative to a purely competitive model–, through a parameter called effective interspecific competition. In other words, a higher (lower) value of this parameter with respect to the bare interspecific competition means that the specie feels a higher (lower) competitive load with respect to a system were only competitive interactions are present. Moreover, the relative effect of the specific configuration of mutualistic interactions can be also quantified through this parameter. Furthermore, it is possible to relate the value of the interspecific competition parameter with both dynamical and structural stability [Pascual-García et al. (2015); Ferrera et al. (2015)] and, in turn, with the biodiversity that the system can host [Bastolla et al. (2009); Pascual-García and Bastolla (2015)].

This reasoning allowed us to show that mean field models of mutualism, i.e. systems with fully connected mutualistic interactions, are able to support a higher biodiversity than purely competitive models [Bastolla et al. (2009)]. Moreover, when we relaxed the mean field condition and particular configurations were considered, the nestedness pattern appeared as particularly relevant in the reduction of the effective competition, being other configurations –such as those compartimentalized–, even detrimental.

Our results were vigorously challenged by James *et* al. claiming that other matrix property, the connectance, was more relevant than the nestedness, and further questioning the actual role of mutualistic interactions [James et al. (2012)]. We have shown that the discrepancies were easily rationalized, given that the procedure proposed by James *et* al. generates unfeasible equilibria with high probability [Pascual-García et al. (2014a)], a result that motivated us for further research.

We modeled several mutualistic regimes where the relative strength of the competition to mutualism ratio and the facultative or obligatory mutualistic behaviour of insects heavily determine the structural stability of the system [Pascual-García and Bastolla (2015)]. We also found that the struc-

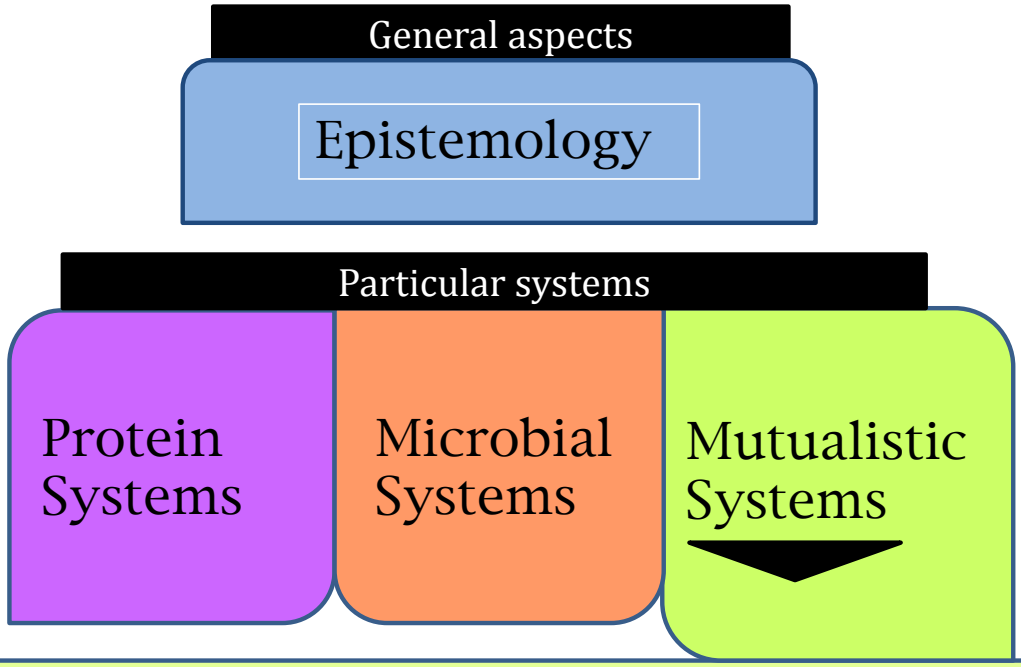
tural stability can be predicted through two parameters. First, the above mentioned effective interspecific competition. And second, the propagation of perturbations in the effective productivities of the system. The effective productivities reflect the effective behaviour of the growth rates when we include mutualistic terms. Since the growth rates are the parameters perturbed in the analysis of structural stability, we measure how perturbations in the growth rates are propagated into the effective productivities.

Using both parameters, we can predict the structural stability and verify our predictions with numerical simulations. Notable results are found for different mutualistic regimes [Pascual-García and Bastolla (2015)], either highlighting the role of connectance or the nestedness. Whereas connectance seems to be always beneficial to reduce the propagation of perturbations –a result strongly reminiscent of the arguments of MacArthur–, the nestedness rather affects the effective interspecific competition, although in some regimes it is insensitive. In this way, the relevance of connectance versus nestedness is translated into physical magnitudes, namely the relevance of the effective interspecific competition or the propagation of perturbations, in the structural stability of ecosystems. Whether any of these topological properties is more relevant for sustaining biodiversity depends in turn on which magnitude –either the effective interspecific competition or the propagation of perturbations–, dominates the regime's behaviour [Pascual-García and Bastolla (2015)]. Interestingly, this behaviour also depends on the specific value of the bare interspecific competition which, if it is high enough, limits the positive effects that can be obtained from mutualistic interactions when it is low.

In summary, working with this system we established a correspondence between the system behaviour and the microscopic observed properties –specified in terms of constraints in the interaction pattern–, such as the nestedness and the connectance. From the point of view of the physical behaviour of the system, we shed some light on how to understand what are, from an evolutionary perspective, the main determinants.

Indeed, our results are consistent [Pascual-García and Bastolla (2015)] with an evolutionary model aiming to explain these microscopic patterns [Suweis et al. (2013)]. We observe that the maximization of biomass of the individual species as an evolutionary target is reached through a minimization of the effective interspecific competition. Furthermore, our modelling approach readily provides theoretical estimations for some constraints that may be experimentally verified. For instance, we found out that, in order to deal with an obligatory mutualistic regime, it is necessary to establish that the biomass of plant species must be more than one million times that of animal species. Therefore, these results confirm that we are actually dealing with a theoretical modeling setting which provides predictions that may be tested through empirical work.

4.1. Article [MUT-1]



1) Bastolla U, Fortuna MA, Pascual-García A, Ferrera A, Luque B, Bascompte J. (2009) The architecture of mutualistic networks minimizes competition and increases biodiversity. *Nature.* 458(7241):1018-20

2) Pascual-García A ., (2010) Explorando el rol de la Competición, el Mutualismo y la Arquitectura en Redes Ecológicas: ¿Qué podemos decir sobre la Biodiversidad?
Published in: *Evolución y Adaptación: 150 años después del origen de las especies.* Editors: Hernán Dopazo and Arcadi Navarro.
ISBN 978-84-92910-06-9

3) Pascual-García A., Ferrera A., and Bastolla, U. (2014) Does mutualism hinder biodiversity? arXiv preprint arXiv:1409.1683

4) Pascual-García A ., Ferrera A., and Bastolla, U. (2015) Effective competition determines the structural stability of model ecosystems. *Under revision.*

5) Ferrera A., Pascual-García A ., and Bastolla, U. (2015) Effective competition determines the global stability of model ecosystems. *Under revision.*

6) Pascual-García A., Bastolla U., (2015) The complexity-stability relation of mutualistic systems reconciles MacArthur and May. *Under revision.*

nature Vol 458|23 April 2009|doi:10.1038/nature07950

LETTERS

The architecture of mutualistic networks minimizes competition and increases biodiversity

Ugo Bastolla[1], Miguel A. Fortuna[2], Alberto Pascual-García[1], Antonio Ferrera[3], Bartolo Luque[3] & Jordi Bascompte[2]

The main theories of biodiversity either neglect species interactions[1,2] or assume that species interact randomly with each other[3,4]. However, recent empirical work has revealed that ecological networks are highly structured[5–7], and the lack of a theory that takes into account the structure of interactions precludes further assessment of the implications of such network patterns for biodiversity. Here we use a combination of analytical and empirical approaches to quantify the influence of network architecture on the number of coexisting species. As a case study we consider mutualistic networks between plants and their animal pollinators or seed dispersers[5,8–11]. These networks have been found to be highly nested[5], with the more specialist species interacting only with proper subsets of the species that interact with the more generalist. We show that nestedness reduces effective interspecific competition and enhances the number of coexisting species. Furthermore, we show that a nested network will naturally emerge if new species are more likely to enter the community where they have minimal competitive load. Nested networks seem to occur in many biological and social contexts[12–14], suggesting that our results are relevant in a wide range of fields.

A long-held tenet in ecology is that the structure of an ecological network can largely affect its dynamics[3,6,7,15,16]. Recent work has unravelled the structure of plant–animal mutualistic networks[5,8–11], but little is known about the implications of these network patterns for the persistence of biodiversity. Previous theory has analysed the dynamics of mutualistic communities without considering their structure[3,17–20]. More recently, ecologists have started numerically to explore the robustness of mutualistic networks[10,21–25], but no study has yet determined how the size of the network depends on its structure. However, understanding the factors determining the number of coexisting species is possibly the most fundamental problem in ecology and conservation biology. Here we analytically quantify whether and to what extent the architecture of mutualistic networks enhances the number of species that can stably coexist in a community (Fig. 1). Also, we explore the emergence of this network architecture through the assembly process. Our analytical approach provides general, insightful results about the equilibrium behaviour instead of simulating the dynamics of our system before such an equilibrium (Supplementary Fig. 1).

We must first derive a baseline biodiversity that will occur in the absence of mutualistic interactions. We therefore begin by considering previous theory that predicts the number of coexisting species when there are only competitive interactions[26,27]. Next we build a generalized model of mutualisms in which species in the same group compete with each other and interact mutualistically with species in the other group (Methods). For direct competition for resources without mutualism, previous work has shown that the largest eigenvalue of the competition matrix limits the maximum biodiversity that the system can attain[26,27]. This predicted maximum number of plant species (similar for animals) can be expressed as

$$\bar{S}^{(\mathrm{P})} = \frac{1-\tilde{\rho}^{(\mathrm{P})}}{\tilde{\rho}^{(\mathrm{P})}} \qquad (1)$$

where $\tilde{\rho}^{(\mathrm{P})}$ is the normalized effective interspecific competition parameter, which can be computed from the main eigenvalue, λ_1, of the normalized competition matrix (Supplementary Methods) as

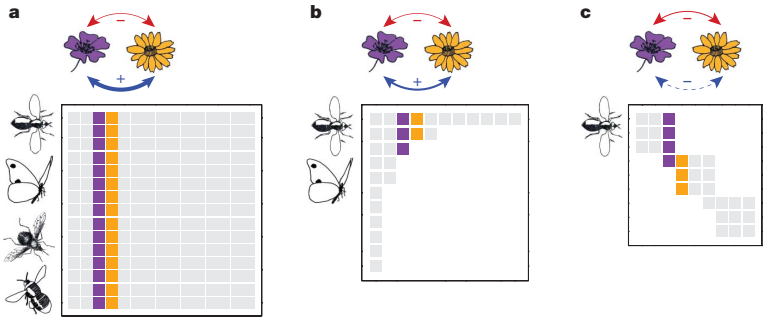


Figure 1 | The structure of mutualistic networks determines the number of coexisting species. Each panel represents a plant–animal network with different structures: **a**, fully connected; **b**, nested; **c**, compartmentalized. Two plants and their respective interactions are highlighted. They compete for resources such as nutrients (red arrow), but also have indirect interactions mediated by their common pollinators (blue arrow), which may change in sign and magnitude (indicated by arrow line style). As the number of shared pollinators is higher, positive effects outweigh negative ones, and the theory predicts a higher number of coexisting species as indicated by the size of the matrices.

[1]Centro de Biología Molecular, Universidad Autónoma de Madrid – CSIC, Madrid 28049, Spain. [2]Integrative Ecology Group, Estación Biológica de Doñana, CSIC, c/ Américo Vespucio s/n, Sevilla 41092, Spain. [3]Departamento de Matemática Aplicada y Estadística, ETSI Aeronáuticos, Universidad Politécnica de Madrid, Plaza Cardenal Cisneros 3, Madrid 28040, Spain.

NATURE|Vol 458|23 April 2009 LETTERS

$$\tilde{\rho}^{(\mathrm{P})} = \frac{\lambda_1 - 1}{S^{(\mathrm{P})} - 1} \quad (2)$$

Here $S^{(\mathrm{P})}$ is the observed number of plant species, which gives the dimensions of the interaction matrices. Qualitatively, the larger is $\tilde{\rho}^{(\mathrm{P})}$, the smaller is the number of species that can stably coexist in a purely competitive system. To obtain explicit analytical formulae, we will henceforth consider direct competition of mean-field type assuming that all species within a set compete with each other with identical intensities (this can be relaxed in numerical simulations; Supplementary Methods). In this case, the quantity computed using equation (2) is equal to the direct competition parameter, $\rho^{(\mathrm{P})}$.

Now that we have set up the baseline limit to the number of coexisting species defined by equation (1), we can incorporate mutualism between plants and animals and quantify the new limit to biodiversity. It is still possible to derive an effective competition matrix that includes the effect of mutualism. The maximum eigenvalue of this matrix limits biodiversity through equations (1) and (2). We first consider the fully connected mutualistic network in which all plants interact with all animals (Fig. 1a). The normalized effective interspecific competition, $\tilde{\rho}^{(\mathrm{P})}_{\mathrm{mut}}$, is related to the direct competition without mutualism as follows, where $a^{(\mathrm{P})}$ is a parameter (Supplementary Information equation (7)) that is proportional to the strength of mutualistic interactions:

$$\tilde{\rho}^{(\mathrm{P})}_{\mathrm{mut}} = \frac{\rho^{(\mathrm{P})} - a^{(\mathrm{P})}}{1 - a^{(\mathrm{P})}} \quad (3)$$

Stable solutions exist for $a^{(\mathrm{P})} < \rho^{(\mathrm{P})}$. We can see from equation (3) that $\tilde{\rho}^{(\mathrm{P})}_{\mathrm{mut}}$ is smaller than $\rho^{(\mathrm{P})}$. This means that mutualism always reduces the effective interspecific competition in a fully connected plant–animal network. The predicted maximum number of plant species in the presence of mutualism, $\bar{S}^{(\mathrm{P})}_{\mathrm{mut}}$, becomes (Supplementary Methods)

$$\bar{S}^{(\mathrm{P})}_{\mathrm{mut}} = \frac{1 - \tilde{\rho}^{(\mathrm{P})}_{\mathrm{mut}}}{\tilde{\rho}^{(\mathrm{P})}_{\mathrm{mut}}} = \frac{\bar{S}^{(\mathrm{P})}}{1 - a^{(\mathrm{P})}/\rho^{(\mathrm{P})}} \quad (4)$$

which is strictly greater than $\bar{S}^{(\mathrm{P})}$, proving that fully connected mutualistic networks increase the number of coexisting species by reducing the effective interspecific competition.

Having quantified the increase in biodiversity due to mutualism in the fully connected case, we proceed by assessing how this mutualistic effect is shaped by the structure of mutualistic networks (Fig. 1b, c). We will repeat the above arguments relaxing the assumption that plant and animal species interact with all species in the other group. Whereas the effective competition matrix in the case of mean-field mutualism contained terms describing an average identical effect of one species on another, now the elements of the effective competition matrix, $C^{(\mathrm{P})}_{ij}$, are different and have to be written explicitly as (Supplementary Methods)

$$C^{(\mathrm{P})}_{ij} = \delta_{ij} + \frac{1}{\bar{S}^{(\mathrm{P})}} + R\left(\frac{1}{S^{(\mathrm{A})} + \bar{S}^{(\mathrm{A})}} n^{(\mathrm{P})}_i n^{(\mathrm{P})}_j - n^{(\mathrm{P})}_{ij}\right) \quad (5)$$

where δ_{ij} is the Kronecker delta function (1 if $i = j$, 0 otherwise), R is the mutualism-to-competition ratio (Supplementary Information equation (23)), $n^{(\mathrm{P})}_i$ is the number of interactions of plant species i and $n^{(\mathrm{P})}_{ij}$ is the number of shared interactions between species i and j. Importantly, the right-hand side of equation (5) decreases with the nestedness of the mutualistic network (as defined in Methods). As a consequence, by inspection nestedness reduces the effective interspecific competition for a given distribution of number of interactions across plant species and fixed parameters. Because the predicted maximum number of plant species (equation (4)) increases with decreasing effective competition, the model predicts that the more nested is the matrix, the higher is the maximum biodiversity.

To explicitly quantify the increase in biodiversity (from the baseline of an exclusively competitive system) due to the nested architecture of mutualistic networks, we computed the derivative of the predicted maximum number of plant species (equation (4)) with respect to the mutualism-to-competition ratio:

$$\left.\frac{1}{\bar{S}^{(\mathrm{P})}_{\mathrm{mut}}}\frac{\partial \bar{S}^{(\mathrm{P})}_{\mathrm{mut}}}{\partial R}\right|_{R=0} = \left(1 + \frac{1}{\bar{S}^{(\mathrm{P})}}\right)\langle n^{(\mathrm{P})}\rangle \left[\bar{S}^{(\mathrm{P})}\left(\hat{\eta}^{(\mathrm{P})} - \frac{\langle n^{(\mathrm{P})}\rangle}{S^{(\mathrm{A})} + \bar{S}^{(\mathrm{A})}}\right) - (1 - \hat{\eta}^{(\mathrm{P})}) + \frac{\langle (n^{(\mathrm{P})})^2\rangle - \langle n^{(\mathrm{P})}\rangle^2}{\langle n^{(\mathrm{P})}\rangle (S^{(\mathrm{A})} + \bar{S}^{(\mathrm{A})})}\frac{S^{(\mathrm{P})} + \bar{S}^{(\mathrm{P})}}{S^{(\mathrm{P})} - 1}\right] \quad (6)$$

Here $\langle n^{(\mathrm{P})}\rangle = \sum_i n^{(\mathrm{P})}_i / S^{(\mathrm{P})}$ and $\langle (n^{(\mathrm{P})})^2\rangle = \sum_i (n^{(\mathrm{P})}_i)^2 / S^{(\mathrm{P})}$ are the mean and mean-square number of mutualistic interactions per plant species, respectively. This derivative increases with the parameter $\hat{\eta}^{(\mathrm{P})} = \sum_{i \neq j} n^{(\mathrm{P})}_{ij} \Big/ \left((S^{(\mathrm{P})} - 1)\sum_k n^{(\mathrm{P})}_k\right)$, which is highly correlated with the measure of nestedness defined in Methods. As seen above, mutualism of the fully connected type always increases the number of coexisting species, setting a maximum limit to biodiversity (fully connected networks have the maximum numbers of absolute and shared mutualistic interactions; Fig. 1a). Structured networks, however, may increase the effective competition and reduce biodiversity if there are not enough shared interactions (that is, for low nestedness; Fig. 1c), or if direct competition is strong so that the predicted maximum numbers of species in the absence of mutualism, $\bar{S}^{(\mathrm{A})}$ and $\bar{S}^{(\mathrm{P})}$, are small. Therefore, the architecture of mutualistic networks highly conditions the sign and magnitude of the effect of mutualism on the number of coexisting species. Nestedness provides the maximum number of species given a certain number of interactions (Fig. 1b). The next question is to unravel how nested mutualistic networks arise in the first place. In Supplementary Methods, we analytically show that a new species entering the community will experience the lowest competitive load, and will therefore be most likely to be incorporated into the community, if it

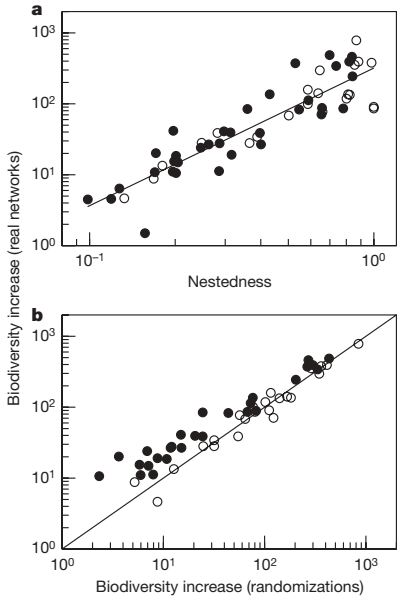


Figure 2 | The nested architecture of real mutualistic networks increases their biodiversity. a, The increase in the predicted maximum biodiversity (sum of plant and animal species) of a mutualistic network as a function of its value of nestedness. Each symbol represents a real network. **b**, Relationship between the increase in the predicted maximum biodiversity for real networks versus randomizations. All significantly nested networks (filled symbols) show a higher increase in biodiversity. The increase in biodiversity is calculated as a numerical approximation to equation (6). The observed numbers of species ($S^{(\mathrm{P})}$ and $S^{(\mathrm{A})}$) are given in Supplementary Table 1. Other parameters are $\bar{S}^{(\mathrm{P})} = \bar{S}^{(\mathrm{A})} = 50$ and $R = 0.005$.

interacts with the most generalist species. This naturally leads to a nested network.

To illustrate the predicted effect of network architecture on biodiversity, we incorporate the structure of each one of 56 real mutualistic networks (Supplementary Table 1) into our analytical expression (equation (5)). In Fig. 2a, we plot the increase in biodiversity in relation to the baseline limit without mutualism (equation (6)) against the level of nestedness. As can be seen, real communities that are more nested show higher increases in biodiversity. It is possible, however, that this increase is mediated by a covariant variable such as the number of species or interactions. To rule this out, we use an alternative way of exploring the role of network structure that keeps constant all variables but nestedness. Figure 2b shows the comparative increase in biodiversity for both real and randomized networks (Methods). In the bulk of communities (45 of 56, $P = 2.0 \times 10^{-6}$, binomial test), the real architecture induces a higher increase in biodiversity than the randomization. More importantly, all networks that are significantly nested (Methods; filled symbols in Fig. 2b) have a greater increase in biodiversity than do their randomizations. Nestedness may be correlated with other properties of network structure such as degree distribution or disassortativity, and the overall contribution to biodiversity increase may therefore be a composite of all these properties that shape the architecture of mutualistic networks.

Our analytical framework can complement previous non-interacting or mean-field approaches to ecology[1,2], by quantifying the importance of network structure for biodiversity. Ideally, this could provide an assessment of the relative contributions of different mechanisms to biodiversity maintenance, a critical task at present in the face of global change. A variety of systems can be described as similar cooperative networks[12–14]. The dynamics of such systems can be captured by appropriate versions of the mutualistic model studied here. Therefore, our analysis can be extended to address questions such as to what extent systemic risk depends on the structure of the financial systems[13], how the optimum number of companies is determined by the architecture of contractor–manufacturer networks[14], and to what degree the structure of social networks favours the evolution of cooperation[28].

METHODS SUMMARY

We used a mutualistic model defined as a system of differential equations. It describes the dynamics of a community of n plant species and m animal species as a function of their intrinsic growth rates, interspecific competition, and mutualistic effects represented as nonlinear, saturating functional responses (Holling type II). We controlled the structure of the plant–animal mutualistic network and were able to analytically solve the model for several network architectures.

We analytically estimated nestedness by averaging the number of shared interactions between two given plants relative to their respective numbers of interactions. In a completely nested matrix, the sets of interactions overlap, therefore maximizing the above quantity. This analytical measure of nestedness allowed us to directly relate nestedness to the effective competition matrix, and to write our analytical solutions as a function of nestedness.

We assessed the significance of nestedness by estimating the probability, p, that a randomization of the network is equally or more nested than the real matrix[5]. Our randomizations assumed that the probability of an interaction was proportional to the generalization level of both the plant and the animal species[5].

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 24 August 2008; accepted 5 March 2009.

1. Alonso, D., Etienne, R. S. & McKane, A. J. The merits of neutral theory. *Trends Ecol. Evol.* **21,** 451–457 (2006).
2. Volkov, I., Banavar, J. R., Hubbell, S. P. & Maritan, A. Patterns of relative species abundance in rainforests and coral reefs. *Nature* **450,** 45–49 (2007).
3. May, R. M. *Stability and Complexity of Model Ecosystems* (Princeton Univ. Press, 1974).
4. Chesson, P. Mechanisms of maintenance of species diversity. *Annu. Rev. Ecol. Syst.* **31,** 343–366 (2000).
5. Bascompte, J., Jordano, P., Melián, C. J. & Olesen, J. M. The nested assembly of plant-animal mutualistic networks. *Proc. Natl Acad. Sci. USA* **100,** 9383–9387 (2003).
6. Montoya, J. M., Pimm, S. L. & Solé, R. V. Ecological networks and their fragility. *Nature* **442,** 259–264 (2006).
7. Pascual, M. & Dunne, J. A. (eds). *Ecological Networks: Linking Structure to Dynamics in Food Webs* (Oxford Univ. Press, 2006).
8. Jordano, P., Bascompte, J., Olesen, J. M. & Invariant properties in coevolutionary networks of plant-animal interactions. *Ecol. Lett.* **6,** 69–81 (2003).
9. Vázquez, D. P., Aizen, M. A. & Asymmetric specialization: a pervasive feature of plant-pollinator interactions. *Ecology* **85,** 1251–1257 (2004).
10. Bascompte, J., Jordano, P. & Olesen, J. M. Asymmetric coevolutionary networks facilitate biodiversity maintenance. *Science* **312,** 431–433 (2006).
11. Olesen, J. M., Bascompte, J., Dupont, Y. L., Jordano, P. & The modularity of pollination networks. *Proc. Natl Acad. Sci. USA* **104,** 19891–19896 (2007).
12. Guimarães, P. R. Jr, Sazima, C., Furtado dos Reis, S. & Sazima, I. The nested structure of marine cleaning symbiosis: is it like flowers and bees? *Biol. Lett.* **3,** 51–54 (2007).
13. May, R. M., Levin, S. A. & Sugihara, G. Ecology for bankers. *Nature* **451,** 893–895 (2008).
14. Saavedra, S., Reed-Tsochas, F. & Uzzi, B. A simple model of bipartite cooperation for ecological and organizational networks. *Nature* **457,** 463–466 (2009).
15. Sugihara, G. *Niche Hierarchy: Structure Assembly and Organization in Natural Communities*. PhD thesis, Princeton Univ. (1982).
16. Sugihara, G. Graph theory, homology and food webs. *Proc. Symp. Appl. Math.* **30,** 83–101 (1984).
17. Wright, D. H. A simple, stable model of mutualism incorporating handling time. *Am. Nat.* **134,** 664–667 (1989).
18. Pachepsky, E., Taylor, T. & Jones, S. Mutualism promotes diversity and stability in a simple artificial ecosystem. *Artif. Life* **8,** 5–24 (2002).
19. Tokita, K. & Yasutomi, A. Emergence of a complex and stable network in a model ecosystem with extinction and mutation. *Theor. Popul. Biol.* **63,** 131–146 (2003).
20. Rikvold, P. A. & Zia, R. K. P. Punctuated equilibria and 1/f noise in a biological coevolution model with individual-base dynamics. *Phys. Rev. E* **68,** 031913 (2003).
21. Memmott, J., Waser, N. M. & Price, M. V. Tolerance of pollinator networks to species extinctions. *Proc. R. Soc. Lond. B* **271,** 2605–2611 (2004).
22. Fortuna, M. A. & Bascompte, J. Habitat loss and the structure of plant-animal mutualistic networks. *Ecol. Lett.* **9,** 281–286 (2006).
23. Burgos, E. *et al.* Why nestedness in mutualistic networks? *J. Theor. Biol.* **249,** 307–313 (2007).
24. Rezende, E. L., Lavabre, J. E., Guimarães, P. R. Jr, Jordano, P. & Bascompte, J. Nonrandom coextinctions in phylogenetically structured mutualistic networks. *Nature* **448,** 925–928 (2007).
25. Okuyama, T. & Holland, J. N. Network structural properties mediate the stability of mutualistic networks. *Ecol. Lett.* **11,** 208–216 (2008).
26. Bastolla, U., Lässig, M., Manrubia, S. C. & Valleriani, A. Biodiversity in model ecosystems, I: coexistence conditions for competing species. *J. Theor. Biol.* **235,** 521–530 (2005).
27. Bastolla, U., Lässig, M., Manrubia, S. C. & Valleriani, A. Biodiversity in model ecosystems, II: species assembly and food web structure. *J. Theor. Biol.* **235,** 531–539 (2005).
28. Lieberman, E., Hauert, C. & Nowak, M. A. Evolutionary dynamics on graphs. *Nature* **433,** 312–316 (2005).
29. Holland, J. N., Okuyama, T. & DeAngelis, D. L. Comment on ''Asymmetric coevolutionary networks facilitate biodiversity maintenance''. *Science* **313,** 1887 (2006).
30. Atmar, W. & Patterson, B. D. The measure of order and disorder in the distribution of species in fragmented habitat. *Oecologia* **96,** 373–382 (1993).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgments We thank P. Jordano and J. Olesen for providing data and insight, A. Ramirez Ortiz for discussions and P. Buston and D. Stouffer for comments on a previous draft. J. Olesen provided the drawings in Fig. 1. Funding was provided by the Spanish Ministry of Science and Technology (through a Ramon y Cajal Contract and a Consolider Ingenio Project to U.B., a PhD Fellowship to M.A.F. and a grant to B.L.) and by the European Heads of Research Councils, the European Science Foundation, and the EC Sixth Framework Programme through a European Young Investigator Award (J.B.). Research at the Centro de Biología Molecular Severo Ochoa is facilitated by an institutional grant from the Ramón Areces Foundation.

Author Contributions U.B., jointly with A.P.-G., A.F. and B.L., performed the analytical development. M.A.F. analysed the real data and, jointly with B.L., performed the simulations. J.B. compiled the real data and, jointly with U.B., designed the study and wrote the first version of the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to J.B. (bascompte@ebd.csic.es).

doi:10.1038/nature07950 nature

METHODS

The mutualistic model. The dynamical equation for the population of plant species i is

$$\frac{dN_i^{(P)}}{dt} = \alpha_i^{(P)} N_i^{(P)} - \sum_{j\in\mathbf{P}} \beta_{ij}^{(P)} N_i^{(P)} N_j^{(P)} + \sum_{k\in\mathbf{A}} \frac{\gamma_{ik}^{(P)} N_i^{(P)} N_k^{(A)}}{1 + h^{(P)} \sum_{l\in\mathbf{A}} \gamma_{il}^{(P)} N_l^{(A)}} \quad (7)$$

where upper indices (P) and (A) denote 'plant' and 'animal', respectively, N_i represents the number of individuals of species i and **P** and **A** indicate the sets of plant and animal species, respectively. The parameter α_i represents the intrinsic growth rate in the absence of mutualism, and β_{ij} represents the direct interspecific competition for resources between species i and j (for example light and nutrients in the case of plants, and breeding sites in the case of animals). The last term describes the mutualistic interaction, through nonlinear functional responses representing a saturation of consumers as the resources increase. The parameter γ_{ik} defines the per capita mutualistic strength of animal k on plant i, and h can be interpreted as a handling time. The equations for animal populations can be written in a symmetric form by interchanging the indices (A) and (P). Equation (7) incorporates all elements recently adduced as necessary ingredients for a realistic model of facultative mutualism[17,29], plus additional ones such as the explicit interspecific competition term. It generalizes previous mutualistic models and allows the reconciliation of previous results on particular cases (Supplementary Methods).

Fixed points of the model. We can analytically obtain the fixed points of model (7) through some algebraic transformations and Taylor expansions (see Supplementary Methods for the full analytical development). There are two different solutions. The first is characterized by small equilibrium biomasses, $N \ll 1/h\gamma$. Because the mutualistic strength, γ, has to remain small for this to be stable, we call this regime weak mutualism. A second type of fixed point, which we refer to as strong mutualism, corresponds to equilibrium biomasses, N, of order $1/h\gamma$. As soon as the weak-mutualism fixed point becomes unstable, the strong-mutualism fixed point becomes stable. Because mutualistic networks are built upon weak dependences[10], the weak-mutualism solution seems the most plausible; it is the one considered in the main text, whereas the strong-mutualism regime is described in Supplementary Methods.

The weak-mutualism fixed-point equations can be written in the form of a linear system, $\sum_j C_{ij}^{(P)} N_j^{(P)} = p_i^{(P)}$, where $p_i^{(P)}$ are the entries of the effective productivity vector (Supplementary Methods). We show in Supplementary Methods that the necessary and sufficient condition for dynamic stability in the weak-mutualism regime is that all equilibrium biomasses are positive and the effective competition matrix is positive definite (that is, all eigenvalues are real and positive).

Measuring nestedness. The level of nestedness of the mutualistic matrix is usually estimated by means of appropriate software[5,12,30]. Here we introduced an explicit definition of nestedness that makes the calculation more straightforward and had the advantage of being related to the form of the effective competition matrix. For plant species, it reads

$$\eta^{(P)} = \frac{\sum_{i<j} n_{ij}^{(P)}}{\sum_{i<j} \min(n_i^{(P)}, n_j^{(P)})}$$

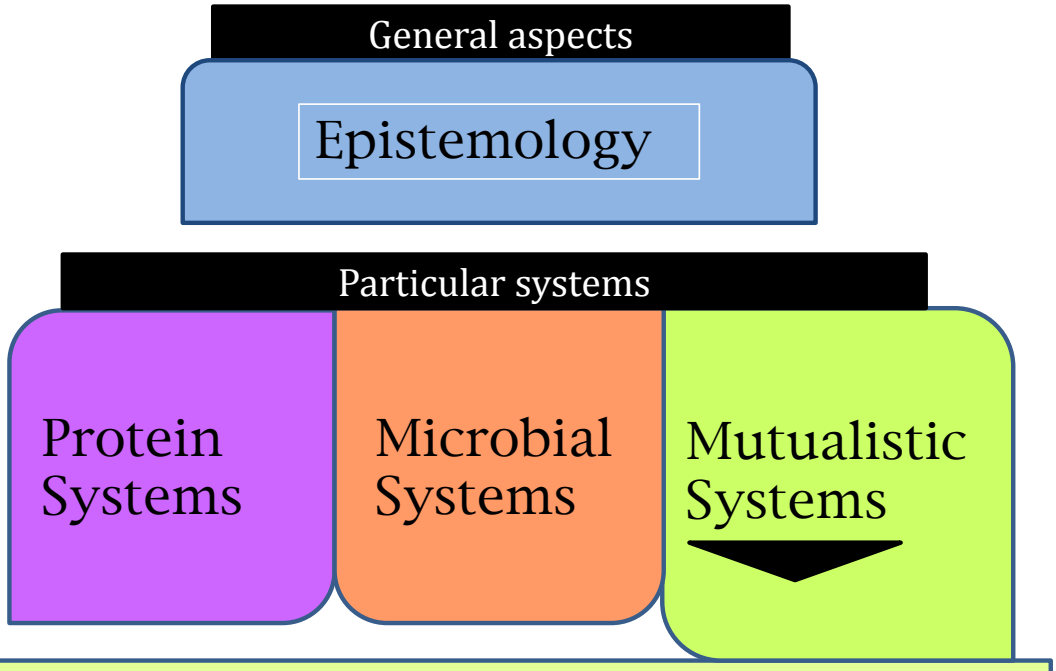
Here $\min(n_i^{(P)}, n_j^{(P)})$ refers to the smaller of the two values $n_i^{(P)}$ and $n_j^{(P)}$. A symmetric definition holds for animal species. This nestedness index ranges from zero to one, and is highly correlated with previous measures of nestedness.

To assess the significance of nestedness in a real community, we used a population of randomizations of the real community. Our null model randomized the interaction matrix probabilistically maintaining the generalization level of both the plant and the animal species. Specifically, the probability of an interaction between plant i and animal j, π_{ij}, is given by the following expression[5], where p_i and q_j are the fractions of occupied cells in row i and column j, respectively:

$$\pi_{ij} = \frac{p_i + q_j}{2}$$

As a statistic indicating significance, we estimated the probability, p, that a randomization was equally or more nested than the real matrix[5].

4.2. Article [MUT-2]



1) Bastolla U, Fortuna MA, Pascual-García A, Ferrera A, Luque B, Bascompte J. (2009) The architecture of mutualistic networks minimizes competition and increases biodiversity. *Nature.* 458(7241):1018-20

2) Pascual-García A ., (2010) Explorando el rol de la Competición, el Mutualismo y la Arquitectura en Redes Ecológicas: ¿Qué podemos decir sobre la Biodiversidad?
Published in: *Evolución y Adaptación: 150 años después del origen de las especies.* Editors: Hernán Dopazo and Arcadi Navarro.
ISBN 978-84-92910-06-9

3) Pascual-García A., Ferrera A., and Bastolla, U. (2014) Does mutualism hinder biodiversity? arXiv preprint arXiv:1409.1683

4) Pascual-García A ., Ferrera A., and Bastolla, U. (2015) Effective competition determines the structural stability of model ecosystems. *Under revision.*

5) Ferrera A., Pascual-García A ., and Bastolla, U. (2015) Effective competition determines the global stability of model ecosystems. *Under revision.*

6) Pascual-García A., Bastolla U., (2015) The complexity-stability relation of mutualistic systems reconciles MacArthur and May. ***Under revision.***

The complexity-stability relation of mutualistic systems reconciles MacArthur and May

Alberto Pascual-García[1] and Ugo Bastolla[1,2]

[1] Centro de Biologia Molecular "Severo Ochoa"
CSIC-UAM Cantoblanco, 28049 Madrid, Spain

[2] E-mail: ubastolla@cbm.csic.es

Abstract

Which properties of ecosystems favour stability against environmental perturbations and help maintening biodiversity is a key question of theoretical ecology that has recently reconsidered mutualistic systems, generating intense controversy. Despite adopting similar models, some works found that mutualism increases species persistence while others found the opposite result, and there is disagreement on which properties have the strongest influence. Here we address this debate under the point of view of structural stability against global perturbations. We show that structural stability can be predicted through two quantities, the effective interspecific competition ρ^{eff} and the propagation of perturbations η'. The mutualistic network architecture and parameters affect these control variables in a complex way that rationalize previous contradictory results. In particular, mutualism decreases the effective competition ρ^{eff}, thereby enhancing structural stability and persistence, only when the direct interspecific competition is weak. In the weak mutualistic regime, mutualistic interactions reduce ρ^{eff} when networks are nested. Strong mutualistic interactions close to saturation and obligatory mutualism in the weak-strong regime influence structural stability mainly by reducing the propagation of environmental perturbations in highly connected networks. This mechanism is reminiscent of MacArthur's proposal that ecosystem complexity enhances stability. We note that predatory interactions influence the propagation of perturbations in the same way as mutualistic interactions, but their influence on ρ^{eff} is the opposite one. In conclusion, the relationship between the architecture of mutualistic networks and their persistence is complex, but analytic theory allows to predict simple trends that shed new light on the debate on the relationship between mutualism and biodiversity.

Which properties of ecosystems enhance their stability against environmental perturbation, favouring the maintenance of biodiversity, is a key question of theoretical ecology.

In this context, the concept of complexity had a preeminent role [1], starting from the inspiring proposal by MacArthur that it favours the stability of ecosystems [2] and its apparent falsification in the model of May [3]. Nowadays, ecologists prefer to talk about ecosystem architecture [4], but the essence of the question remains the same. Much of this recent theoretical work on ecosystem stability addressed mutualistic networks of plants and pollinators and plants and seed dispersers [5–11], which had been overlooked in previous years partly because of the lack of field data and partly because they become rapidly unstable if modelled with simple Lotka-Volterra models [12, 13], a difficulty that was overcome implementing non-linear functional responses [14]. Despite intense work, these theoretical investigations disagree on the effect of mutualism on persistence. Whereas some studies indicated that, in some conditions, mutualism increases the persistence of model ecosystems [6, 7, 9, 11], others reached the opposite conclusion [8], and different studies highlighted either nestedness [6, 7, 11] or connectance [8] as the network property that most influences persistence.

The classical approach to the stability debate pioneered by May [3] assumes that the equilibrium is feasible (i.e. all abundances are positive), randomly extracts the interaction parameters, and tests the resulting dynamical stability. Nevertheless, if the interaction matrix has a mathematical property called diagonal stability, every feasible equilibrium is globally dynamically stable [15], so that the study of dynamical stability can be generalized to quantifying the perturbations of parameters that maintain feasibility, i.e. structural stability. Despite the structural stability of ecosystems against environmental perturbations is arguably a main determinant of the maintenance of biodiversity, its study is less common in theoretical ecology than in other fields of computational biology [16], with some recent exceptions [6, 9, 11, 17].

We found that structural stability can be analytically predicted based on two quantities: the effective interspecific competition parameter [17] and the propagation of perturbations, a quantity strongly related with the ideas put forward by MacArthur [2]. This framework rationalizes the contradictory results of earlier analysis, and it also encompasses the recent proposal by Suweis et al. that the main effect of mutualism consists in enhancing species abundances [10], which are largely determined by the effective competition.

Our numerical experiments place the model ecosystem in a feasible and stable equilibrium and globally perturbate the intrinsic growth rates. The dynamics is then simulated, extinctions are recorded, and the structural stability is measured as the relative perturbation Δ_c above which half of the simulations result in the extinction of at least one species. Following previous work [6, 8, 9], we model two groups of species (plants and animals, denoted by the superscripts P and A) that interact through within-group competition of Lotka-Volterra (LV) type and between-group mutualistic interactions that saturate for large abundance [18] (see Methods and [19]). We report here only the equations for plants, since those for animals can be obtained interchanging the superscripts P and A.

We compare different parameter regimes and network architectures by choosing the un-

perturbed growth rates α_i in an equivalent way for different interaction matrices, such that the equilibrium is feasible for large perturbations. The α_i that maximize this structural stability can be analytically computed when the mutualistic growth rates are linear [17]. This procedure has been recently proposed [11], but it cannot be applied if mutualism approaches saturation. Here and in previous work [9] we choose α_i imposing that the resulting equilibrium abundances are almost equal for species in the same guild, which maximizes the abundance of the rarest species. These ideal growth rates $\overline{\alpha}_i^{(\mathrm{P})}$ are given by

$$\overline{\alpha}_i^{(\mathrm{P})} = \sum_{j\in\mathbf{P}} \beta_{ij}^{(\mathrm{P})} \overline{N}_j^{(\mathrm{P})} - \sum_{k\in\mathbf{A}} \frac{\gamma_{ik}^{(\mathrm{P})} \overline{N}_k^{(\mathrm{A})}}{1 + h_i^{(\mathrm{P})} \sum_{l\in\mathbf{A}} \gamma_{il}^{(\mathrm{P})} \overline{N}_l^{(\mathrm{A})}} \tag{1}$$

With the above ansatz, the $\overline{\alpha}_i^{(\mathrm{P})}$ are negatively correlated with the number of mutualistic interactions, establishing a trade-off between the number of interactions and their metabolic cost. On the other hand, the $\overline{\alpha}_i^{(\mathrm{P})}$ of a system with pure competition and fully connected competition matrix are identically distributed. In the simulations of James *et al.* [8] the growth rates were identically distributed also for sparse mutualistic networks, producing unfeasible equilibria with high probability. This explains why they found that mutualistic interactions modelled in this way hinder biodiversity. Choosing growth rates that are almost ideal both for pure competition and for mutualism provides a fairer comparison [9].

Depending on the sign of the intrinsic growth rates, mutualism can be obligatory, if the α_i are positive for plants and negative for animals, or facultative if they are all positive. A natural way to achieve obligatory mutualism is to choose equilibrium abundances larger for plants than for animals and large maximum mutualistic growth rates $1/h_i^{(\mathrm{A})}$ for animals (see Extended Data).

Once a feasible equilibrium is set, we test its dynamical stability by considering the equivalent Lotka-Volterra system with effective mutualistic interactions $\gamma_{ik}^{(\mathrm{eff,P})}$ (see Box). The equilibrium is locally stable if these effective interactions are small, which happens both for $\gamma_0 < \gamma_0^{(1)}$ (weak mutualism, $z_i \ll 1$) and $\gamma_0 > \gamma_0^{(2)}$ (strong mutualism, $z_i \gg 1$), see Extended data Fig.S2. Plots of the critical mutualistic strengths $\gamma_0^{(1)}$ and $\gamma_0^{(2)}$ for different network architectures are reported in the Extended Data, Fig.S7.

Structural stability can be analytically predicted through the effective competition matrix $C^{(\mathrm{P})}$, Eq.(B2) [17], which represent the direct competition between species in the same guild plus their indirect interactions through species in the other guild, and allows separating the equilibrium equations of the two groups as $\sum_j C_{ij}^{(\mathrm{P})} \overline{N}_j^{(\mathrm{P})} = p_i^{(\mathrm{P})}$, where $p^{(\mathrm{P})}$, Eq.(B6) is the effective productivity vector. While the complete interaction matrix has both positive and negative signs, we expect that the components of $C^{(\mathrm{P})}$ are positive, so that the Perron-Frobenius theorem implies that its main eigenvector has only positive components [22]. As shown elsewhere [17, 23], the optimal distribution of productivities $p^{(\mathrm{P})}$ that provide maximal structural stability must be directed along the main eigenvector $v^{(P),1}$ of $C^{(\mathrm{P})}$. For large systems, the equilibrium is feasible only if $p^{(\mathrm{P})}$ is almost parallel

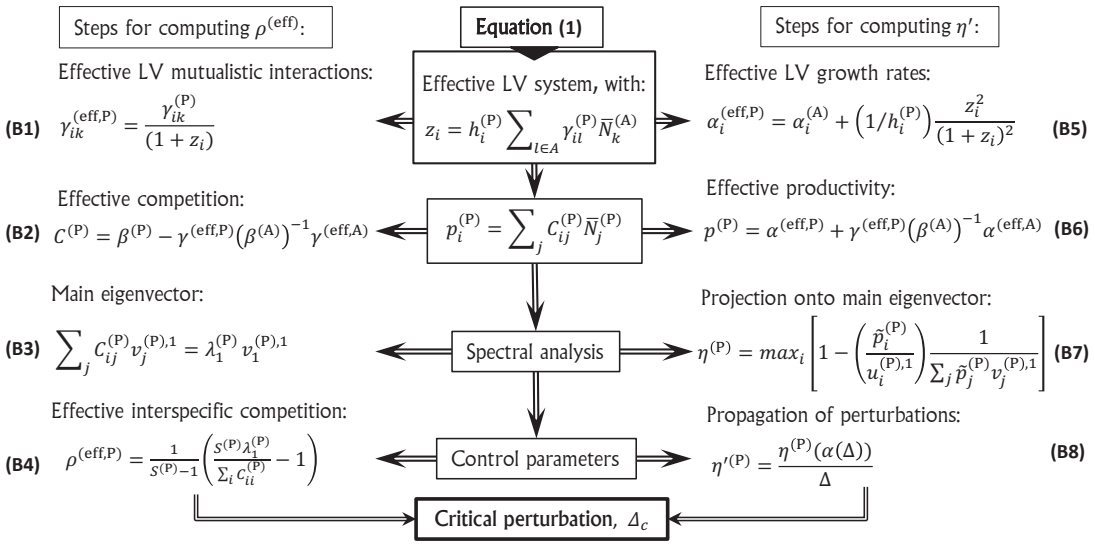


Figure 1: Flux of the computation of the interspecific effective competition parameter ρ^{eff} and the propagation of perturbations $\eta\prime$.

to $v^{(P),1}$, which implies that the equilibrium abundances are also almost parallel to $v^{(P),1}$ and are given by

$$\overline{N}_i^{(\mathrm{P})} \approx \frac{v_i^{(P),1}}{\left(\rho^{(\mathrm{eff,P})}(S^{(\mathrm{P})}-1)+1\right)\left(\sum_i C_{ii}/S\right)}, \tag{2}$$

where $\rho^{(\mathrm{eff,P})}$, Eq.(B4), represents the mean effective competition between different species of the same guild. The smaller $\rho^{(\mathrm{eff,P})}$, the larger species abundances.

As previously shown [17], the smaller is $\rho^{(\mathrm{eff,P})}$ the easier it is to fulfill the condition for feasibility, which is expressed as an inequality on $\eta^{(\mathrm{P})}$, Eq.(B7). In our model, global environmental perturbations of relative size Δ affect the intrinsic growth rates α and through them the productivities p_i and the feasibility factor η. The term $\eta\prime$ Eq.(B8) expresses the rate at which a perturbations of the α_i tends to increase η. Combined with the feasibility condition Eq.(B7), this leads to the following analytic expression of the critical relative perturbation of growth rates $\Delta_c^{(\mathrm{P})}$ above which extinctions occurr:

$$\Delta_c^{(\mathrm{P})} = \frac{1}{\eta\prime^{(\mathrm{P})}} \left[\left(\frac{S^{(\mathrm{eff,P})}}{S^{(\mathrm{P})}+S^{(\mathrm{eff,P})}}\right)\left(1 - f_1^{(\mathrm{P})} \frac{n_c^{(\mathrm{P})}}{\langle N^{(\mathrm{P})}\rangle}\right)\right], \tag{3}$$

where $S^{(\mathrm{eff,P})} = \left(1-\rho^{(\mathrm{eff,P})}\right)/\rho^{(\mathrm{eff,P})}$ sets a biodiversity scale that is larger the smaller $\rho^{(\mathrm{eff,P})}$, $S^{(\mathrm{P})}$ is the number of species, the term $f_1^{(\mathrm{P})}$ is the fraction of plant species that are the only connection of at least one animal species in obligatory mutualism, and

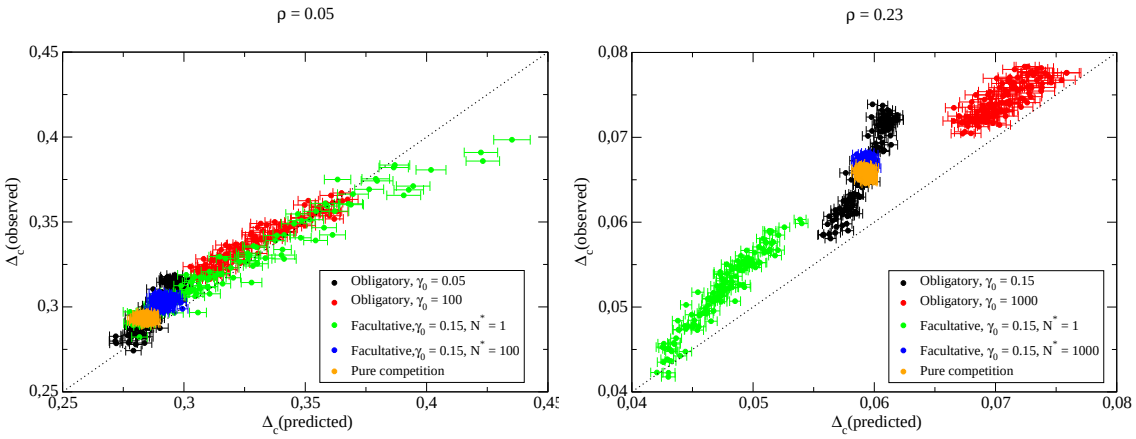


Figure 2: Structural stability Δ_c, defined as the relative perturbation of intrinsic growth rates such that the probability that at least one species gets extinct is 0.5. The figures represent observed versus predicted Δ_c in eight different mutualistic regimes and in pure competition. Each point represent one different mutualistic network. The direct competition parameter is $\rho = 0.05$ (left) and $\rho = 0.23$ (right).

$n_c^{(\mathrm{P})}/\langle N^{(\mathrm{P})}\rangle = \left(1+\gamma_0\left(h^{(\mathrm{A})}\right)^2\sqrt{\hat{N}^{(\mathrm{P})}/\hat{N}^{(\mathrm{A})}}\right)^{-1}$ is the minimum plant abundance that maintains an animal species (zero for facultative mutualism). $\Delta_c^{(\mathrm{A})}$ is analogous, with $n_c^{(\mathrm{A})} = 0$, and the critical perturbation Δ_c is the smaller between $\Delta_c^{(\mathrm{A})}$ and $\Delta_c^{(\mathrm{P})}$. We tested the above equation with simulations in several regimes, presented in Fig.2. The figure also shows purely competitive systems ($\gamma_0 = 0$), evidencing that mutualism enhances structural stability for some regimes and networks but not for others. The good agreement between predicted and measured Δ_c shows that ρ^{eff} and η' are sufficient to predict structural stability.

For purely competitive systems $\eta'^{(\mathrm{P})} \approx 1$, whereas for mutualistic systems it can be estimated as the ratio between the standard deviation of the productivity under pertubations of the growth rates, which is proportional to the square root of the number of mutualistic links, and the average productivity, which is proportional to the number of links (see Extended Data). Therefore, the larger the number of links is, the smaller $\eta'^{(\mathrm{P})}$ and the larger structural stability.

A simple computation, reported in the Extended Data, shows that mutualistic interactions reduce the effective interspecific competition $\rho^{(\mathrm{eff,P})}$, i.e. $\rho^{(\mathrm{eff,P})} < \rho^{(\mathrm{P})}$, only if the direct interspecific competition $\rho^{(\mathrm{P})}$ is weak, otherwise mutualism increases $\rho^{(\mathrm{eff,P})}$. In particular, in the weak mutualistic regime in which all $z_i \ll 1$ (small γ_0, small $\overline{N}$ and

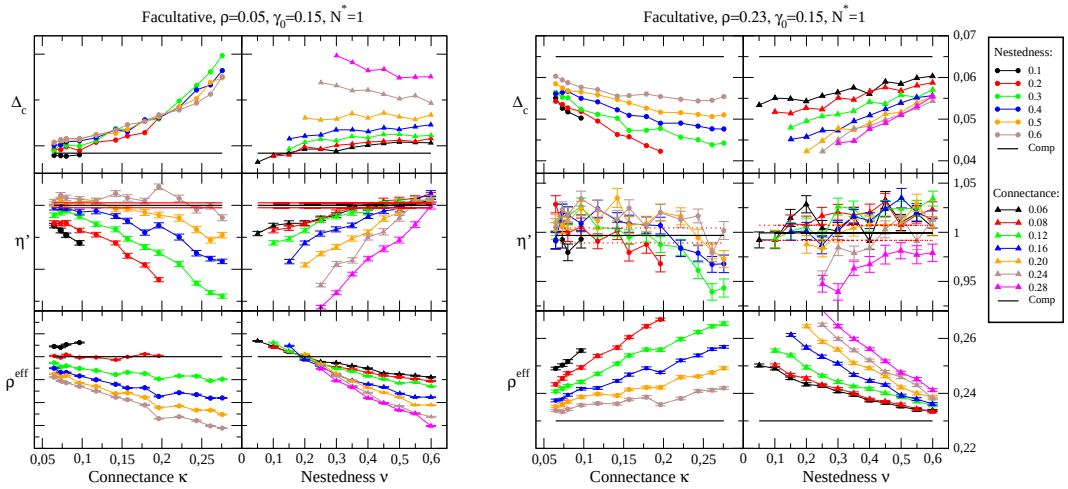


Figure 3: Structural stability Δ_c, effective interspecific competition parameter ρ^{eff} and propagation of perturbations η' versus nestedness and connectance of mutualistic networks for two of the regimes of Fig.2 representing facultative weak mutualism. Note that the effective competition is smaller than the direct competition for small $\rho = 0.05$ while the opposite occurs for large $\rho = 0.23$.

small connectance) it holds

$$\rho^{(\text{eff,P})} = \frac{(\gamma_0)^2}{(1-\rho^{(\text{P})})} \langle \mu_{ii} \rangle \left(\rho^{(\text{P})} - \rho^{(P),\text{c}} \right) , \tag{4}$$

where the critical competition parameter $\rho^{(P),\text{c}}$ only depends on the mutualistic network and on the competition $\rho^{(\text{A})}$ between animals (see Extended Data, where μ_{ii} is also defined). From now on we omit the superscripts to simplify the notation. For fully connected networks it always holds $\rho^{\text{c}} = 1$, therefore mutualism always reduces ρ^{eff} as previously shown [6]. For sparse networks, ρ^{c} and consequently ρ^{eff} decrease with the nestedness (Extended Data section 10.1, Fig.S3 and Fig.S4). At the same time, the propagation of perturbations $\eta\prime$ increases with nestedness, giving the opposite effect on structural stability, and it decreases with connectance, as expected. When the effect of ρ^{eff} prevails, as in regime A of our simulations, structural stability is positively influenced by nestedness (Fig.3A). When the effect of $\eta\prime$ prevails, as in regime B, structural stability increases with connectance (Fig.3B), as previously reported by James at al. [8].

The situation is different for strong mutualism with saturated interactions ($\gamma_0 > \gamma_0^{(2)}$, large $\overline{N}$ or large connectance). In this case, for small connectance and large S/S_0 it holds $\rho^c < 0$, i.e. mutualism increases ρ^{eff}, in particular for less connected networks (Fig.3 and Extended Data Eq.(S35), while $\eta\prime$ decreases with connectance, enhancing structural stability (Extended Data Fig.S5). For obligatory mutualism that is weak for plants and

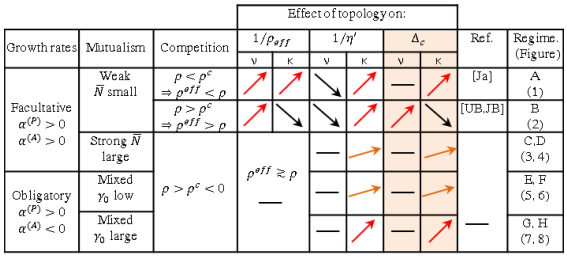


Figure 4: Graphical summary of the studied regimes, described in the first three columns. The arrows indicate the influence of network architecture, connectance κ and nestedness ν, on $1\rho^{\text{eff}}$, $1/\eta'$ (the inverse variables are used because they are positively related with structural stability) and Δ_c. The tenth column cites the publication where qualitatively similar results were reported.

strong for animals ρ^{eff} does not depend on γ_0 and is almost equal to ρ. In this case, network architecture influences structural stability mainly through the decrease of $\eta\prime$ with connectance. A graphical summary of these different regimes is presented in Fig.4.

Effective competition also determines dynamical stability. We conjecture that the linearised interaction matrix is diagonally stable, and therefore the linearised dynamical system is globally stable [15], if $C^{(\text{P})}$ and $C^{(\text{A})}$ are positive definite [21]. This conjecture is justified in the Extended Data, and it is supported by simulations (Extended data Fig.S2). Since $1-\rho^{(\text{eff,P})}$ is proportional to the average of the minor eigenvalues of $C^{(\text{P})}$, the smaller $\rho^{(\text{eff,P})}$ is, the less likely it is that $C^{(\text{P})}$ has a negative eigenvalue and the equilibrium is unstable. This argument predicts a negative correlation between $\rho^{(\text{eff,P})}$ and the critical mutualistic strength $\gamma_0^{(1)}$, supported by Extended data Fig.S6. Thus we expect that $\gamma_0^{(1)}$ is positively correlated with structural stability, see Fig. S8, and it is inversely correlated with nestedness in the weak mutualistic regime, $1/\gamma_0^{(2)}$ is positively correlated with connectance in the strong mutualistic regime (Extended data Fig.S7), and mutualism is dynamically stable for all values of the mutualistic strength γ_0 in the weak–strong regime, as confirmed by our simulations.

Taking all these results together, we see that the relationship between network architecture and structural stability is complex and changes in different regimes, which explains why previous works that studied only one regime obtained qualitatively different results [6–8, 11]. In particular, mutualistic interactions decrease the effective competition for weak direct competition, otherwise they increase it. It would be interesting to investigate this interplay between competition and mutualism in economic systems. When mutualism is far from saturation nestedness is the property that most influences structural stability, whereas at saturation, in particular for weak-strong obligatory mutualism, connectance has the strongest influence. The negative correlation between the propagation of perturbations and the connectance is highly reminiscent of the argument used by

MacArthur to argue that ecosystem complexity favours stability [2]. However for some regimes of parameters the connectance hinders biodiversity and the effective competition is the relevant variable, a result closer to the perspective of May [3]. While the propagation of perturbations behaves equally in predatory ecosystems as in obligatory mutualism, we can analytically see that predatory interactions reduce the effective competition for strong direct competition and increase it for weak direct competition, opposite to mutualistic interactions (see Extended Data).

Interestingly, the feasibility condition induces constraints on ecological parameters, such as the trade-off between the number of mutualistic interactions and the dissipation rate of animals. Moreover, for obligatory mutualism the ratio between plant and animal abundances must be $N^{(\mathrm{P})}/N^{(\mathrm{A})} > 2 \cdot 10^5$, consistent with the empirical estimate $N^{(\mathrm{P})}/N^{(\mathrm{A})} \approx 5 \cdot 10^6$ (see [24] and Extended Data). The framework of effective competition can also integrate the recent proposal that the ecological effect of mutualistic interactions is mainly due to their influence on species abundances [10], since feasibility induces an inverse relation between ρ^{eff} and abundance, so that when ρ^{eff} is reduced structural stability, dynamical stability and species abundance increase at the same time.

In conclusion, the study of the structure and dynamics of complex ecological systems, though still in its infancy, is already giving important insights, allowing computational models to detect different regimes that facilitate the comparison with field data and may enhance their predictive power [25].

Acknowledgements

We acknowledge discussions with Antonio Ferrera and Bartolo Luque. This work was supported by grant BFU2012-20020 of the Spanish Government.

Author contributions

U.B. and A.P.G. designed the study. U.B. performed the analytic computations and A.P.G. performed and analyzed the simulations. All authors wrote the paper.

Author Information

Correspondence should be addressed to U.B. (ubastolla@cbm.csic.es)

References

[1] Ives, A. R. and Carpenter, S. R. (2007). Stability and diversity of ecosystems. Science, 317:58-62.

[2] MacArthur, R. (1955). Fluctuations of animal populations and a measure of community stability. Ecology, 36:533-536.

[3] May, R.M. Will a large complex system be stable? *Nature* **238,** 413-414 (1972).

[4] Bascompte J (2010) Ecology. Structure and dynamics of ecological networks. Science 329:765-6.

[5] Bascompte, J., Jordano, P., Melian, C. J. and Olesen, J. M. (2003) The nested assembly of plant-animal mutualistic networks. Proc. Natl Acad. Sci. USA 100, 93839387.

[6] Bastolla, U., Fortuna, M.A., Pascual-García, A., Ferrera, A., Luque, B. & Bascompte, J. The architecture of mutualistic networks minimizes competition and increases biodiversity, *Nature* **458,** 1018-1020 (2009).

[7] Thébault E, Fontaine C (2010) Stability of ecological communities and the architecture of mutualistic and trophic networks. Science 329:853-6.

[8] James, A., Pitchford, J.W. & Plank, M.J. Disentangling nestedness from models of ecological complexity, *Nature* **487,** 227-230 (2012).

[9] Pascual-García, A., Ferrera, A. and Bastolla, U. Comment to "Disentangling nestedness from models of ecological complexity" http://arxiv.org/abs/1409.1683

[10] Suweis, S., Simini, F., Banavar, J.R. and Maritan, A. (2013) Emergence of structural and dynamical properties of ecological mutualistic networks. Nature 500, 449-452.

[11] Rohr RP, Saavedra S, Bascompte J (2014) Ecological networks. On the structural stability of mutualistic systems. Science 345:1253497.

[12] May, RM (1982) Mutualistic interactions among species. Nature 296, 803-804.

[13] Goh, B.S. (1979) Stability in models of mutualism. Am. Nat. 113, 261-275.

[14] Holland, J.N., DeAngelis, D.L. and Bronstein, J.L. (2002). Population dynamics and mutualism: functional responses of benefits and costs. Am. Nat., 159, 231244.

[15] Goh B.S. Global stability in many-species systems. The American Naturalist (1977) 111, 135-143.

[16] Thom, R. (1994) Structural stability and morphogenesis. Addison-Wesley, Boston, 1994.

[17] Bastolla, U., Lässig M., Manrubia, S.C. & Valleriani A. Biodiversity in model ecosystems, I: coexistence conditions for competing species. *J. Theor. Biol.* **235,** 521-530 (2005).

[18] Okuyama, T. & Holland, J. N. Network structural properties mediate the stability of mutualistic networks. *Ecol. Lett.* **11,** 208-216 (2008)

[19] We use the following notation: $N_i^{(\mathrm{P})}$ denotes the abundance of plant species i, $\alpha_i^{(\mathrm{P})}$ its intrinsic growth rate, $\beta^{(\mathrm{P})}$ and $\beta^{(\mathrm{A})}$ the direct competition matrices of LV type, $\gamma^{(\mathrm{P})}$ and $\gamma^{(\mathrm{A})}$ the mutualistic matrices and $1/h_i$ is the maximum mutualistic growth rate of species i, whose inverse h_i is related with the handling time [20]. The dynamical equations depend on the mutualistic network and on seven meta-parameters: the direct interspecific competition parameters $\rho^{(\mathrm{A})}$ and $\rho^{(\mathrm{P})} \in [0, 1]$, the maximum mutualistic growth rates $1/h^{(\mathrm{A})}$ and $1/h^{(\mathrm{P})}$, the mutualistic strength γ_0, the ratio $\hat{N}^{(\mathrm{A})}/\hat{N}^{(\mathrm{P})}$ of animal and plant abundances, and the equilibrium abundances $\overline{N}$, which determine the intrinsic growth rates α_i. We present results obtained in different regimes of parameters, as described in Table of the Methods section.

[20] Wright, D. H. (1989) A simple, stable model of mutualism incorporating handling time. Am. Nat. 134, 664667.

[21] Ferrera A, Pascual-García A and Bastolla U (2014) Effective competition determines the global stability of model ecosystems. Submitted.

[22] Meyer, C. (2000) Matrix analysis and applied linear algebra. SIAM, ISBN 0-89871-454-0.

[23] Pascual-García A, Ferrera A and Bastolla U (2014) Effective competition determines the structural stability of model ecosystems. Submitted.

[24] Jorgensen, S. E. and Svirezhev, Y. M. (2004). Towards a thermodynamic theory for ecological systems. Elsevier.

[25] Burkle LA, Alarcón R (2011) The future of plant-pollinator diversity: understanding interaction networks across time, space, and global change. Am J Bot. 98:528-38.

Methods

The mutualistic networks studied in this paper have $S^{(\mathrm{A})} = 46$ animal and $S^{(\mathrm{P})} = 47$ plant species, as found in a field study at Cainama, Venezuela (Ramirez 1989). We randomly generate 125 networks with different combinations of connectance and nestedness (see Extended Data).

For each network, we model the multi-species population dynamics through the equations

$$\frac{1}{N_i^{(\mathrm{P})}}\frac{\mathrm{d}N_i^{(\mathrm{P})}}{\mathrm{d}t} = \alpha_i^{(\mathrm{P})} - \sum_{j\in\mathbf{P}} \beta_{ij}^{(\mathrm{P})} N_j^{(\mathrm{P})} + \sum_{k\in\mathbf{A}} \frac{\gamma_{ik}^{(\mathrm{P})} N_k^{(\mathrm{A})}}{1 + h_i^{(\mathrm{P})} \sum_{l\in\mathbf{A}} \gamma_{il}^{(\mathrm{P})} N_l^{(\mathrm{A})}} \,. \tag{5}$$

[6, 8, 9], where $N_i^{(\mathrm{P})}$ denotes the abundance of plant species i, $\alpha_i^{(\mathrm{P})}$ is its intrinsic growth rate in the absence of other species, $\beta_{ij}^{(\mathrm{P})}$ is the direct competition matrix and $\gamma_{ik}^{(\mathrm{P})}$ is the mutualistic matrix. The equations for animals can be obtained interchanging the superscripts P and A, and they will be omitted in the rest of this section. For simplicity, we assume that equivalent interaction parameters are identically distributed. The parameters $\hat{N}^{(\mathrm{P})}$ and $\hat{N}^{(\mathrm{A})}$ set the scale of carrying capacity for plant and animal populations, respectively, and the parameters $\rho^{(\mathrm{P})}, \rho^{(\mathrm{A})} \in [0, 1]$ set the interspecific competition measured in units of intraspecific competition, yielding the direct competition

$$\beta_{ij}^{(\mathrm{P})} = \frac{b_{ij}}{\hat{N}^{(\mathrm{P})}} \left(\rho^{(\mathrm{P})} + \delta_{ij}(1 - \rho^{(\mathrm{P})})\right) \,, \tag{6}$$

where δ_{ij} is Kronecker's delta and b_{ij} are dimensionless numbers uniformly distributed in $[1 - \delta_b, 1 + \delta_b]$. We parameterize mutualistic interactions as

$$\gamma_{ik}^{(\mathrm{P})} = a_{ik} \frac{\gamma_0 c_{ik}^{(\mathrm{P})}}{\sqrt{\hat{N}^{(\mathrm{P})} \hat{N}^{(\mathrm{A})}}} \tag{7}$$

where γ_0 measure the strength of mutualism with respect to competition, a_{ik} is the adjacency matrix of the mutualistic network, the dimensionless parameters $c^{(\mathrm{P})}$ are uniformly distributed between $1-\delta_c$ and $1+\delta_c$ if $a_{ik} = 1$ and are zero if $a_{ik} = 0$. The maximum mutualistic growth rates $1/h_i$ are chosen equal to $1/H^{(\mathrm{P})}$ for plants, and $1/H^{(\mathrm{A})}$ for animals. Finally, we uniformly extract the equilibrium abundances in $[\overline{N}\hat{N}^{(\mathrm{P})}(1-\delta_N), \overline{N}\hat{N}^{(\mathrm{P})}(1+\delta_N)]$. Growth rates $\overline{\alpha}_i^{(\mathrm{A})}$ are determined such that the equilibrium abundances satisfy the fixed point equations i.e. Eq.(1) of the main text. With these assumptions, the dynamical equations depend on the mutualistic network and on seven meta-parameters that determine the interaction matrices ($\rho^{(\mathrm{A})}$, $\rho^{(\mathrm{P})}$, $H^{(\mathrm{A})}$, $H^{(\mathrm{P})}$, γ_0 and $\hat{N}^{(\mathrm{A})}/\hat{N}^{(\mathrm{P})}$) and the equilibrium abundances $\overline{N}$ and through them the α_i, and on the 3 parameters δ_b, δ_c and δ_N that control the broadness of the distributions. We present results for several regimes of meta-parameters, described in the table below.

For each network and each set of metaparameters, we randomly draw 50 realizations of the interaction matrices and the equilibrium abundances, and we determine the critical

Parameter	A	B	C	D	E	F	G	H
$S^{(\mathrm{A})}$, $S^{(\mathrm{P})}$	46, 47							
α	Facultative				Obligatory			
$\rho^{(\mathrm{A})}$, $\rho^{(\mathrm{P})}$	0.05		0.23		0.05		0.23	
γ_0	0.15				0.05	100	0.15	10^3
$\overline{N}$	1	100	1	1000	1			
$H^{(\mathrm{A})}$	0.1				0.23		0.066	
$H^{(\mathrm{P})}$	0.1				0.25			
$\hat{N}^{(\mathrm{P})}/\hat{N}^{(\mathrm{A})}$	1				$7 \cdot 10^7$			
δ_b	0.15							
δ_c	0.15							
δ_N	0.15							

Table 1: Metaparameters regimes presented in the figures.

values of γ_0 at which the system looses dynamical stability (see below). Computations are only performed for γ_0 in the allowed range. Subsequently, we generate 100 random perturbations of all the intrinsic growth rates, $\alpha_i = \overline{\alpha}_i\,(1 + \Delta r_i)$, where r_i is a random number extracted in $[-1, 1]$, we integrate ecological dynamics with the Bulirsch-Stoer algorithm with adaptive step until convergence, considering extinct species whose abundance falls below 10^{-8} of the initial value. For each value of Δ we record the fraction of simulations in which at least one species got extinct and through interpolation we obtain the critical perturbation Δ_c at which this fraction equals 0.5.

To assess dynamical stability, we take the derivatives of Eq.(5) close to the fixed point and we transform the dynamical system into an equivalent Lotka-Volterra system, obtaining the effective mutualistic interaction Eq.(B1) and effective growth rates Eq.(B5). Any species can have either weak mutualism, if its equilibrium mutualistic growth rate is far from saturation ($z_i \ll 1$), or strong mutualism ($z_i \gg 1$). The effective mutualistic interactions increase with γ_0 in the weak regime and decrease in the strong regime, reaching a maximum in between. The equilibrium is locally stable if the eigenvalues of the community matrix have negative real parts, which happens if the mutualistic matrices γ^{eff} of Eq.(B1) are small, i.e. both for $\gamma_0 < \gamma_0^{(1)}$ and $\gamma_0 > \gamma_0^{(2)}$.

The effective competition matrix [17] must be computed from the linearized system as Eq.(B2) Its main eigenvalue allows to compute the effective competition $\rho^{(\mathrm{eff,P})}$, Eq.(B4). It is easy to see that $\rho^{(\mathrm{eff,P})}$ represents the effective competition between different species, averaged with weights given by the main eigenvector of $C^{(\mathrm{P})}$ (see also Extended Data). Since $1 - \rho^{(\mathrm{eff,P})}$ is proportional to the average of the minor eigenvalues, positivity of $C^{(\mathrm{P})}$ requires that $\rho^{(\mathrm{eff,P})} < 1$.

From the equilibrium equation $\sum_j C_{ij}^{(\mathrm{P})} \overline{N}_j^{(\mathrm{P})} = p_i^{(\mathrm{P})}$, where the effective productivity

vector $p^{(\mathrm{P})}$ is given by Eq.(B6), we obtain the following necessary condition for feasibility [17, 23]:

$$\eta^{(\mathrm{P})} \equiv \max_i \left(1 - \frac{p_i^{(\mathrm{P})}}{v_i^{(P),1} p^{(P),1}}\right) \leq \frac{S^{(\mathrm{eff,P})}}{S^{(\mathrm{P})} + S^{(\mathrm{eff,P})}} \left(1 - \frac{n_c^{(\mathrm{P})}}{\langle N^{(\mathrm{P})} \rangle}\right) , \tag{8}$$

where $S^{(\mathrm{P})}$ is the number of species, $v^{(P),1}$ is the main eigenvector of $C^{(\mathrm{P})}$, $v^{(P),1}$ is the main eigenvector of $C^{(\mathrm{P})}$, $p^{(P),1} = \sum_j p_j^{(P)} v_j^{(P),1}$ and n_c is the critical abundance below which extinctions take place.

To predict structural stability, we have to compute how the relative perturbation of intrinsic growth rates α_i affects the quantity η defined above. This computation is complicated by the fact that a change in α_i modifies the equilibrium abundances, and consequently the effective growth rates and mutualistic interactions through Eq.(B0). We can simplify the result by noting that, when mutualistic interactions are far or close to saturation, the change in $\gamma^{(\mathrm{eff,P})}$ and $\alpha^{(\mathrm{eff,P})}$ due to a change in equilibrium abundances is small and it can be neglected, except for obligatory mutualism (see below). Thus, we fix $\gamma^{(\mathrm{eff,P})}$ and the effective competition matrix $C^{(\mathrm{P})}$ and we compute the perturbed $\alpha^{(\mathrm{eff,P})}$ as $\alpha_i^{(\mathrm{eff,P})}(\Delta) = \alpha_i^{(\mathrm{eff,P})}(\Delta = 0) + \Delta\alpha_i$ and from them we obtain the perturbed productivities $p^{(\mathrm{P})}(\alpha(\Delta))$ we project them onto the main eigenvector of the competition matrix (assumed unchanged), and we compute the perturbed η according to Eq.(8). From this we obtain $\eta'^{(\mathrm{P})} = \eta^{(\mathrm{P})}(\Delta)/\Delta$.

Part IV

Discussion

We have inherited from our forefathers the keen longing for unified, all-embracing knowledge. (...) But the spread, both in and width and depth, of the multifarious branches of knowledge by during the last hundred odd years has confronted us with a queer dilemma. We feel clearly that we are only now beginning to acquire reliable material for welding together the sum total of all that is known into a whole; but, on the other hand, it has become next to impossible for a single mind fully to command more than a small specialized portion of it. I can see no other escape from this dilemma (lest our true who aim be lost for ever) than that some of us should venture to embark on a synthesis of facts and theories, albeit with second-hand and incomplete knowledge of some of them and at the risk of making fools of ourselves. So much for my apology.

Erwin Schrödinger

Discussion

A person who never made a mistake
never tried anything new.
Albert Einstein

In this thesis, we focused on emergent patterns in different complex biological systems. Our work is motivated by the observation of novel microscopic evolutionary patterns, which are the consequence of an unknown behaviour. We developed methods to characterize quantitatively these patterns and to address their significance with statistical and mechanistic models.

The characterization of these patterns is made taking into account that an emergent behaviour is associated with a restriction in the regions of the phase space that the system visits. This implies that there are some values of the variables that are not observed, due to the existence of internal or external constraints limiting the dynamics.

However, identifying the mechanisms leading to the observed patterns is a difficult challenge. Some of them may be directly traceable from the environment, whereas others would be rather the consequence of the interaction between the components of the system, or of the evolutionary events taking place in the population. We have focused on the interactions between components because they represent a description of the fluxes of information in the system. These fluxes determine the viable values of the variables monitoring the state of the components, thus limiting the observation of other values.

Nevertheless, an exhaustive exploration of the interactions'configurations is unfeasible for large systems, and our strategy to reduce this search focuses on interaction patterns emerging from the evolutionary process. This choice is justified after observing that changes in the evolutionary scale take place after longer periods of time than in the physical scale. Thus it may be considered that the patterns that become fixed are the consequence of the long term performance of the system.

In this sense, the evolutionary constraints would reflect the dynamics performance under the most representative environmental fluctuations within the evolutionary scale, where the environment should be considered here in its widest acceptation –thus including ecological interactions or any other

relevant ingredient external to the organism–. On the other hand, other observed constraints would be rather a consequence of the intrinsic variability within the physical scale, and they should be viewed as more contingent. Altogether, evolutionary constraints represent primary constraints, and this assertion should hold unless the amplitude of unexpected fluctuations in the physical scale becomes so large that the organism is no longer viable in such environment.

The identification of evolutionary patterns of interactions is achieved through the analysis of microstates within the evolutionary space. This means that we look for systems containing the same components –or equivalent in some sense–, evolving in different environments. We expect that the patterns commonly found are the consequence of their adaptation to some subset of underlying environmental or evolutionary features, which are shared in spite of the particular environment where each system inhabits.

In addition, interaction patterns are interesting because it is difficult to demonstrate that they are the main drivers of the observed behaviour unless the dynamics can be modeled explicitly. For instance, the composition of bacterial taxa in different environments could be the consequence of underlying physico-chemical conditions such as light, temperature or salinity, that readily favour the presence of certain taxa. If this were the case, the observed composition of taxa would be perfectly traceable from the environmental features, and thus it can hardly be considered an emergent property, or its emergence strength should be considered very weak, following the definition provided in the first manuscript [Pascual-García (2015)]. If the composition was instead the consequence of ecological interactions, the traceability between the microscopic description and taxa composition would be very difficult, and the emergence strength would depend on the scope of the constraints present in the system [Pascual-García (2015); Ryan (2007)].

Methodological comparison between the systems

This methodological perspective has been applied to systems where the experimental information available is very different. Therefore, we will explore differences and similarities between the systems we have presented in this thesis and the methodologies followed through a series of concepts that we label and summarize in Tab. 4.1, and briefly explain.

The experimental data we considered (Tab. 4.1 (a)) provide a representation of microstates in the evolutionary space. In the case of proteins each microstate is a model of a protein structure, represented through its native amino-acids contacts [Pascual-García et al. (2009, 2010)]. Next, we dealt with metagenomic samples found in different environments, where bacterial genes were sequenced and classified in the analysis of microbial communities. In this case, each sample should be considered as a microstate in the evolutionary space. Finally, we considered a description of mutualistic communities

	Protein structures	Microbial communities	Mutualistic communities
(a) Experimental microstates representation	Protein structures	Compositional / abundances samples	Interaction matrices
(b) Pattern observed	Common native contacts	Aggregation / segregation	Nestedness / connectance
(c) Procedure to identify patterns	Protein structure alignment	Pairwise aggregation / segregation measures	Network analysis
(d) Null model to address significance of the patterns	Similarity between unrelated proteins	Network built under independence assumptions	Matrix randomization with constraints
(e) Working hypothesis	Folds can be objectively defined	Ecological interactions are prevalent	Role of mutualistic configurations on structural stability
(f) Working procedure	Classification monitoring transitivity violations	Evaluation of indirect evidences	Mechanistic models

Table 4.1: Comparison of the three systems considered at the different research stages.

in terms of their composition, and further in experimental data describing the mutualistic interactions between plants and animals.

Our starting point in the analysis of these systems was the identification of relevant interaction patterns (see Table 4.1 (b) and (c)). These patterns, when observed through the microstates, are interpreted in terms of evolutionary and physical constraints, what allows us to rationalize their effect on the behaviour of the system. Therefore, to address this point we focus on the comparison of microstates. We discussed in the epistemological section that the generation of distances or (dis)similarities is a critical exercise in scientific research, being the basis of dimensionality reduction techniques and the identification of constraints [Pascual-García (2015)].

For protein structures, we searched for similarities through a protein structure alignment algorithm aiming to find conserved interactions irrespective of the sequence of amino-acids of the different proteins, thus just on the basis of $\alpha-$carbons similarly arranged in the space. For microbial samples, we searched for significant aggregation or segregation between taxa, i.e. the systematic presence (absence) of pairs of taxa along the different samples, which is reminiscent of the search of conserved inter-residues contacts in proteins. We addressed this task both with classical measures of ecological ressemblance [Legendre and Legendre (2012)] and with a novel probabilistic measure we developed [Pascual-García et al. (2014b)]. For mutualistic systems, it may be thought that the approximation could be similar and we should also look for common properties comparing the interaction matrices. However, mutualistic interaction matrices depend on the specific species sorting –which is not the case for protein structures, where interactions are built following the order imposed by the backbone–. Thus, we characterized the matrices with measures accounting for the different patterns that may be found in a matrix, namely number of observations, degree, assortativity, etc. We further addressed the significance of these measures in the whole ensemble, and then we compared the matrices looking for any measure that is systematically significant.

The statistical significance of these comparisons was established in all cases with respect to a statistical null model (Tab. 4.1 (d)). The similarity between protein structures was determined analysing a set of unrelated protein structures from which we obtained an extreme value distribution. Comparisons departing significantly from this distribution were considered a signature of structural ressemblance which is compatible with an evolutionary interpretation. Significant aggregation and segregation in microbial data were obtained invoking a null model were independence between species was considered, i.e. it was assumed that the species do not interact. From this model, we generated sets of artificial matrices with the same row and column totals and further reflecting coarse grained environmental features. Then, we recomputed the aggregation and segregation measures from these

artificial matrices. With this procedure, we were able to build again statistical distributions, from which we tested the significance of our observations. For mutualistic systems, the significance of interaction patterns was addressed generating artificial matrices where both the number of observations and the row and column totals were conserved on average. This is how we were able to test the significance of global properties of the networks, such as nestedness.

Once we tested the significance of the observed patterns, we built a null hypothesis encompassing the ensemble of microstates, aiming to link it with the observed emergent property, and which rejection may lead to accept an alternative hypothesis (what we call *working hypothesis* Tab. 4.1 (e)). From the space of protein structures, we considered as null hypothesis that the similarities between proteins are so pervasive that we deal with a continuous space, and therefore it is not possible to find discrete folds. With bacterial samples, the most economic hypothesis is that the observed significant aggregations and segregations are a consequence of habitat filtering. In mutualistic systems, we considered that competition is the main driver of the system's stability, and in turn of its biodiversity. In addition, we considered that global network properties in the interactions are not relevant to explain these system's biodiversity.

We tested these hypotheses designing computational experiments for each system (Tab. 4.1 (f)). For the space of structures, the main hypothesis was challenged imposing a classification scheme based on the formal definition of equivalence class, and by monitoring the error we incur in this process. The relative role of habitat filtering versus ecological interactions in microbial communities was examined through several experiments which results would support the null or the alternative hypothesis. For mutualistic systems, we tested our hypothesis with mechanistic models that directly measure the role of the different configurations on the system's stability. One may wonder again why we did not follow a similar mechanistic approximation in the analysis of protein structures, –namely, evaluating the relevance of the patterns found in the protein folding properties–, given that we also know the interaction patterns. The answer is simple: protein folding dynamics is much more complex than populations dynamics. However there is no doubt that the combination of evolutionary patterns obtained from protein structures with dynamical models is a promising area of research. Indeed, as we said in the introduction of protein systems, the modelization of unknown protein structures considering the structures already solved for their homologs is probably the most succesful modeling technique up to date.

Predictive power of results

The outcome of the different experiments has been explained in the results, and they will not be repeated here. But it is important to say that

the value of the results arises from its predictive power, and not from a self-referential reading of the similarities between the different models.

For protein structures, we predicted that the definition of equivalence classes should be possible if the molecular clock hypothesis approximately holds. This hypothesis was verified with the representative data set we considered. Furthermore, the classification found was cross-validated with existing classifications (in the case of SCOP, manually curated by experts). We also found that the distribution of protein structures follows a power law distribution, confirming that most of the new folds that may be discovered will be singletons. These results provide the basis for the development of computational models that automatically classify protein structures, and to develop methods for predicting new ones.

For microbes, we predicted several aggregations that correspond with known mutualistic interactions. In addition, the analysis was consistent with an independent analysis performed on the assemblage of bacteria in gut for infants during their first year of life. It may be also possible to validate these interactions with experimental and computational data. For instance, it could be used to guide the development of bottom-up approaches for builiding bacterial communities with computational models, considering flux balance analysis [Orth et al. (2010)]. Using these models, a coexisting community of three species with spatially explicit interactions has already been built [Harcombe et al. (2014)], but increasing this number would require to consider a combinatorial number of species configurations that could be reduced following our predictions.

For plant-pollinator communities, we explicitly provided an expression for structural stability that has been numerically validated. In addition, we predicted the ratio between plant and animal biomasses for the feasibility of obligatory mutualism, which is consistent with estimations computed from experimental data. More accurate predictions require the existence of new experimental data to fix other parameters such as the interaction strengths, which are very difficult to obtain for this kind of systems –and we would like to acknowledge here the important work performed by field scientists–. An experimental alternative arises from setups working with bacterial species, where the experiments are increasingly controlled and it is much faster to obtain results. For instance, an approximation has already been proposed on which the Lotka-Volterra models were inferred from time-series analyses of metagenomic data with low, but significant, predictive power [Stein et al. (2013)].

In summary, it is possible to make a positive reading of the predictive power of our results. A positive reading arises when we highlight predictions to emphasize our findings. Nevertheless, along this thesis we have stated the importance of a methodological approximation. This is not because we think that the models are important in themselves, but rather because these

methods are powerful for hypothesis rejection.

And, whereas predictions made with coarse evolutionary data are many times limited, the number of hypothesis rejected may be huge. For instance, when we talked about the importance of evolutionary conserved interactions, we pointed out that they may be considered a scaffold over which the system has organized its dynamics due to the presence of long-term constraints. And thus, with this approach we reduce the number of microstates that we need to consider to understand the dynamical behaviour of the system.

Therefore, a negative reading –i.e. when we highlight the results rejected– is readily valuable, given that it has already been achieved considering experimental data and it rejects a large amount of possibilities. In this sense, we should *also* evaluate these models considering their ability to provide a solid scaffold, following top-down approaches, to facilitate the development of new bottom-up models and experiments.

It may be pertinent to ask ourselves whether a negative reading of results should be a standard for a fair development of scientific knowledge, given that we avoid any bias to reach a positive interpretation of results.

Emergence and evolution

We would like to finish this thesis discussing the interpretation of results from an evolutionary perspective. In the articles, we already discussed the interpretation of the results in terms of evolutionary events. For proteins, we pointed out that the existence of both global and local similarities may reflect the relative dominance of gene duplication versus more dramatic evolutionary events. For bacteria, we conjectured that the high number of aggregations we found could be a signature of prevalent syntrophic relationships. For mutualistic communities, we explored the influence of mutualism and of interactions architectures for the system's stability. We made an evolutionary interpretation relating the stability of the system with assemblage processes, where some configurations would be more beneficial for the establishment of new species.

Nevertheless, if we explore the evolutionary mechanisms through which these patterns were selected, we observe several difficulties to provide a clear answer. For instance, for protein structures we observe that the existence of both global and local similarities can be explained saying that these are traits selected for proper protein function. In this sense, the mechanism is clearly explained because there is supervenience –i.e. a clear bottom-up causal relationship– between correct protein function and the individuals selected in the evolutionary process. However, a pattern such as the nestedness reflects a selected trait in the pool of species, and it is difficult to interpret this fact within an evolutionary context taking the individual as the *unique* object of selection. Given that the effect of the pattern on individuals is a downward effect, it should be considered how any benefit provided by a co-

llective pattern is incorporated in the genetic pools of different populations of species, if the selection process acts on single individuals independently.

A solution to this question may be found broading the concept of object of selection, in a sense that we justify as follows. Let us consider that the fitness of any individual f_i can be decomposed in two components, where the first component reflects the fitness f_{ij}^{int} of the individual as a consequence of its ecological interactions with other species j, and the second its fitness $f_i^{\overline{int}}$ due to any other process, thus $f_i = f_i^{int} + f_{ij}^{\overline{int}}$.

Now consider a particular example of two individuals belonging to two different species, a and b, which interact mutualistically. Further consider that there is any evolutionary event affecting the fitness $f \rightarrow \hat{f}$, for instance of the species a, and its effect is negative for the species. If the evolutionary event affects to some gene directly involved in the correct performance of the interaction with b, we will get $\hat{f}_a^{\overline{int}} = f_a^{\overline{int}}$ and $\hat{f}_{ab}^{int} < f_{ab}^{int}$. In this situation, if this evolutionary event affects the outcome of the interaction for the species b, the decrease in the fitness that the species a experiences will also affect the fitness of the species b. In this way, the regions of the genome of the species b involved in the interaction, will be subject to a change in the selection pressure. In this example, there will be regions of the genomes of a and b with some degree of evolutionary coupling, i.e. they *coevolve*. In addition, if $f_{ab}^{int} >> f_a^{\overline{int}}$ and $f_{ba}^{int} >> f_b^{\overline{int}}$ the importance of this tandem in the evolutionary process would be even larger than that of the individuals, and we should talk about a new object of selection, namely the regions of the genomes of both species that are coupled.

It is important to emphasize that the selection process over these expanded objects of selection can be described in the same terms as for individuals: there is a relative competition between all the objects of selection found in the system at the different scales. Indeed, this reasoning could be scaled up towards larger objects of selection. However, it is expected that the selective effects be weaker for larger objects, given that the strength of long-range interactions should decay with distance.

Therefore, we find a picture where different regions of the genome in a given species are coupled with regions of other species with different strengths, dependent on the relative effect of interaction on the fitness of the individuals. It is important to note that this picture stands on a rather artificial assumption, namely that there is a clear separation in two components of the fitness. Instead, genes are typically involved in several processes, and deciphering their contribution to any of these components would be difficult, if not impossible. But what we would like to emphasize is that, as soon as there is significant coevolution, it is difficult to accept the classical view stating that the individual is the unique object of selection.

For instance, in the discussion of our first work on Microbial Ecology [Pascual-García et al. (2014b)], we observe a large number of aggregations

that could be explained through syntrophic relationships, i.e. a metabolic coupling between species of bacteria that would be beneficial for both species. This kind of interactions may have notable consequences for bacterial speciation, given that the selection process is heavily influenced by interactions, and it should be considered as obligate mutualism [Morris et al. (2013)]. Furthermore, it could be a mechanism to explain the large variations in the length of genomes observed in bacteria, because some necessary functions in a given individual could be performed by mutualistic partners and, eliminating these genes, the species would avoid its energetic cost. For mutualistic communities of plants and pollinators, the patterns analysed involve the whole community; but, given that interactions are weaker than in bacterial communities and generation times longer, we would expect a lower selective effect.

Broadening our notion of object of selection allow us sto rationalize the role that concepts such as stability of communities may have in evolution. More stable communities, would favour the selection of those traits more competent to sustaining the interactions, what should be considered as a mechanism of downward selection. Indeed, the relevance of group selection would prevail if the fitness of the group is larger than the sum of the fitness of the individuals due to the interactions [Mayr (1997)], what reflects the popular notion around emergent properties stating that *the whole is more than the sum of the parts*.

This kind of higher level organizations may help organisms to *buffer* certain environmental conditions, and thus the next question would be which kind of stability is favoured and why in each situation. An interesting theoretical framework summarizing this triplet between fitness, stability and environment was proposed by Demetrius [Demetrius (2013)]. He identifies two types of stability strategies –starting from age structured populations [Demetrius et al. (2004)] and further generalizing to other systems [Demetrius and Manke (2005)]–, based on the abundance and degree of hetereogeneity of resources.

In our case, we also dealt with types of stability in our work which optimization cannot be simultaneously maximized –such as stability against unfolding or misfolding in protein structures [Bastolla (2014); Nido et al. (2015)], or dynamical *versus* structural stability in ecosystems [Pascual-García and Bastolla (2015); Bastolla et al. (2005, 2009)]–, and that may be dependent on environmental features. For instance, we conjecture that dynamical stability would be favoured when the environment is relatively stable, thus facilitating the increase in abundances of few species. On the other hand, when large fluctuations are prevalent in the environment, the system rather needs to be stable against changes in those parameters more directly affected by the environment –such as growth rates and interactions–, thus favouring structural stability. Similarly, in a plain environment, stability against unfolding would

be selected, given that the protein has a lower probability of being influenced by perturbations that may lead to a misfolded configuration. Stability against misfolding would be favoured instead in crowded environments –as the cell–, through negative selection [Bastolla (2014)].

In this way, it seems plausible to think that the consequences derived from the appearance of an emergent group behaviour increase the fitness of the group of components involved in the interaction processes –given some environmental conditions– and, if this increase is large enough, it may represent a powerful mechanism for accelerating evolutionary processes.

This perspective would allow us to think in more specific emergent macroscopic properties linking the extended objects of selection with the biological function performed. The macroscopic emergent property will be any function associated to the whole group, emerging from certain microscopic behaviour.

As we said, the net effect of these new functions over the individuals would be a reduction in any negative effects that environmental fluctuations may have on individual fitness. In this way, individuals would be subjected to a lower selective pressure, what may allow the populations involved to increase genetic diversity. For instance, for mutualistic systems, we have shown that an increase in structural stability was related to a decrease in the effective competition that each specie feels. And, in turn, a reduction in the effective competition can be related to a higher capacity of the system to host more coexisting species, i.e. with an increase in its biodiversity [Bastolla et al. (2009); Pascual-García and Bastolla (2015)]. Certainly, understanding in which conditions an increase of diversity is favoured is an urgent task, and the perspective obtained from the analysis of complex biological systems within the evolutionary process seems to provide new insights to this question.

Therefore, the discovery and analysis of emergent patterns in complex biological systems, although still in its infancy, already provides interesting results. In this thesis, we have addressed the epistemological questions surrounding this kind of research, and patterns which belong to very different systems. We have seen that it is possible to answer interesting questions about the emergence and evolution of biological systems working around the identification of microscopic evolutionary patterns. And, also importantly, we showed that these questions open speculative hypotheses for future work.

Considering some general concepts is necessary to provide the means to build general theories in biological systems, which sometimes come from very different areas of knowledge, obiously surpassing the capability of any scientist to know in detail each of these areas [Schrödinger (1992)]. This represents a challenge that should not be considered as dangerous for the development of science, as long as we acknowledge where are the limits of the conclusions derived from the starting assumptions. However, we observe that

this kind of approximation is sometimes seen with suspicion by specialists in the different areas. This –many times founded– suspicion is useful, and it is a responsability for those scientists who aim to link new areas to demonstrate the validity of their approximations.

This should not be a limitation to enrol ourselves in this adventure aiming to establish long-range linkages in human knowledge. In the end, these linkages are precisely interactions between mankind's components, that will affect the cultural microscopic dynamics and may lead to new collective behaviours. It is a matter of time that we observe the consequences of these interactions and, although the current situation in the world leaves little room for optimism, we believe that these consequences will be positive for the fitness of our planet –whatever that means–.

Part V

Conclusions

It looks strange
and it looks strange
and it looks very strange;
and then suddenly
it doesn't look strange at all
and you can't understand what made
it look strange in the first place.

Gertrude Stein

Conclusions

Protein systems

1. Transitivity violation, a measure arising from the mathematical definition of equivalence classes, allow to determine a similarity threshold above which protein structures can be objectively classified into folds. In the region of high similarity, protein structure space is discrete and the proteins are related through global structural similarity. In the region of low similarity, there exist still significant local structural similarity, but the space is continuous. We speculate that the main evolutionary event which drove the existence of folds is gene duplication and subsequent structure divergence, whereas local similarities reflect more dramatic events, such as large insertions or deletions.

2. We proposed a new measure of protein structure divergence based on the number of shared native contact that we called contact divergence. This measure has a deep analogy with sequence divergence, thus allowing to conclude that structure is two to four times more conserved than sequence. We further evaluated the role of protein function, finding that function conservation strongly constrains the structural divergence, whereas it is possible to observe function change with global structural conservation.

Microbial systems

3. We inferred a large number of significant aggregations and segregations between bacterial taxa from data obtained from samples subject to next generation sequencing experiments classified into different environments. Many of these aggregations are difficult to explain through habitat filtering since the number of different environments in which aggregating taxa coexist is large, and cosmopolitan taxa have a strong propensity to aggregate. Furthermore, aggregation present a marked community structure. We retrieved known cooperative interactions among these aggregations, which, together with the higher pro-

pensity to aggregate for closely related species, allow us to conjecture an important role of mutualistic relationships in bacterial speciation.

4. The analysis of the ecological succession in microbial taxa sampled from infant guts during their early development reveals a systematic convergence –in terms of function and composition of taxa– of the communities towards those found in their mothers. Although the process must be divided in two trends related with the different diets (milk *versus* solid-food intake), there is a detectable core of bacteria growing in time. Comparing this core with the results found in our previous work, we identify these species as obligatory bacteria in the gut, being the excluded taxa rare facultative bacteria.

Mutualistic systems

5. We analysed the structural stability of mutualistic communities, which is a key determinant for sustaining biodiversity. We showed that the structural stability can be predicted quantifying two quantities: the propagation of perturbations and the effective interspecific competition. This prediction is in good agreement with numerical simulations. In this way, we identified the regimes where mutualism increases the structural stability, finding that reduced interspecific direct competition is a necessary condition for mutualism having a stabilizing effect. Furthermore, we rationalized the effects that specific architectures of the mutualistic interactions have on the structural stability, solving previous discrepancies found in the literature.

Epistemology

6. We analyzed concepts arising from the analysis and modelization of complex biological systems from a novel topological perspective, showing that difficulties in the determination of the system's boundaries and classification schemes have their origin in the extensional vagueness contained in the conceptual setting of the models. Furthermore, we suggested how to quantify the emergence strength of a pattern by assessing the traceability of the corresponding behaviour and we related it to the scope of the constraints existing in the system.

Conclusiones

Proteínas

1. Las violaciones de transitividad, una medida basada en la definición matemática de clases de equivalencia, permite determinar un umbral de similaridad sobre el cual las estructuras de proteínas pueden ser clasificadas objetivamente en *folds*. En la región de alta similaridad, el espacio de estructuras de proteínas es discreto y las proteínas están relacionadas a través de similaridades globales. En la región de baja similaridad, aún existen relaciones significativas que reflejan similaridades locales, pero el espacio es continuo. Especulamos que el principal evento evolutivo que explica la existencia de estos conjuntos es la duplicación génica, y la subsiguiente divergencia estructural, mientras que las similaridades locales reflejan eventos más dramáticos, tales como grandes inserciones y deleciones.

2. Hemos propuesto una medida de divergencia estructural entre proteínas basada en el número de contactos nativos compartidos, que llamamos divergencia de contactos. Esta medida tiene una profunda analogía con la divergencia en secuencia, y permite concluir que la estructura es de dos a cuatro veces más conservada que la secuencia. Además evaluamos el rol de la función de las proteínas, y encontramos que la conservación en función constriñe fuertemente la divergencia estructural, mientras que es posible encontrar cambio de función con conservación global de la estructura.

Sistemas microbianos

3. Hemos inferido un gran número de agregaciones y segregaciones significantivas entre taxones bacterianos, para datos obtenidos mediante experimentos de secuenciación masiva clasificados en ambientes distintos. Muchas de estas agregaciones son difícilmente explicables considerando el filtro ambiental, dado que el número de ambientes distintos en los cuales taxones que agregan coexisten es elevado. Además, las

agregaciones presentan una estructura clara de comunidad. Recuperamos interacciones cooperativas conocidas entre estas agregaciones, las cuales, junto con la alta propensidad a agregar para especies estrechamente emparentadas, nos permite conjeturar sobre un rol importante de las relaciones mutualistas en la especiación bacteriana.

4. El análisis de la sucesión ecológica en taxones microbianos muestreados en intestinos de bebés durante su desarrollo temprano revela una convergencia sistemática –en términos de función y composición de los taxones– de las comunidades hacia aquéllas obtenidas de sus madres. Aunque el proceso puede ser dividido en dos tramos relacionados con las diferentes dietas (consumo de leche *versus* alimentos sólidos), hay un núcleo detectable de bacterías que crece a lo largo del tiempo. Comparando este núcleo con los resultados que encontramos en nuestro trabajo anterior, identificamos estas especies como bacterias obligatorias en el intestino, siendo las especies excluídas bacterias facultativas raras.

Sistemas mutualistas

5. Hemos analizado la estabilidad estructural de comunidades mutualistas, la cual es clave en el sostenimiento de la biodiversidad. Hemos mostrado que la estabilidad estructural se puede predecir cuantificando dos magnitudes: la propagación de perturbaciones y la competición efectiva interespecífica. Obtenemos un buen acuerdo entre esta predicción y las simulaciones numéricas. De este modo, hemos identificado los regímenes en donde el mutualismo incrementa la estabilidad estructural, y encontramos que una condición necesaria es que la competición directa interespecífica sea moderada. Es más, hemos racionalizado los efectos que tienen las arquitecturas específicas de las interacciones mutualistas en la estabilidad estructural, resolviendo discrepancias previas encontradas en la literatura.

Epistemología

6. Hemos analizado distintos conceptos que surgen en el análisis y modelización de sistemas biológicos complejos desde una novedosa perspectiva topológica, mostrando que dificultades en la determinación de los límites de los sistemas y en los esquemas de clasificación, tienen su origen en la vaguedad extensional del aparato conceptual de los modelos. Es más, sugerimos cómo cuantificar la fuerza emergente de un

patrón a través de la trazabilidad del comportamiento correspondiente, que relacionamos con el alcance de las restricciones existentes en el sistema.

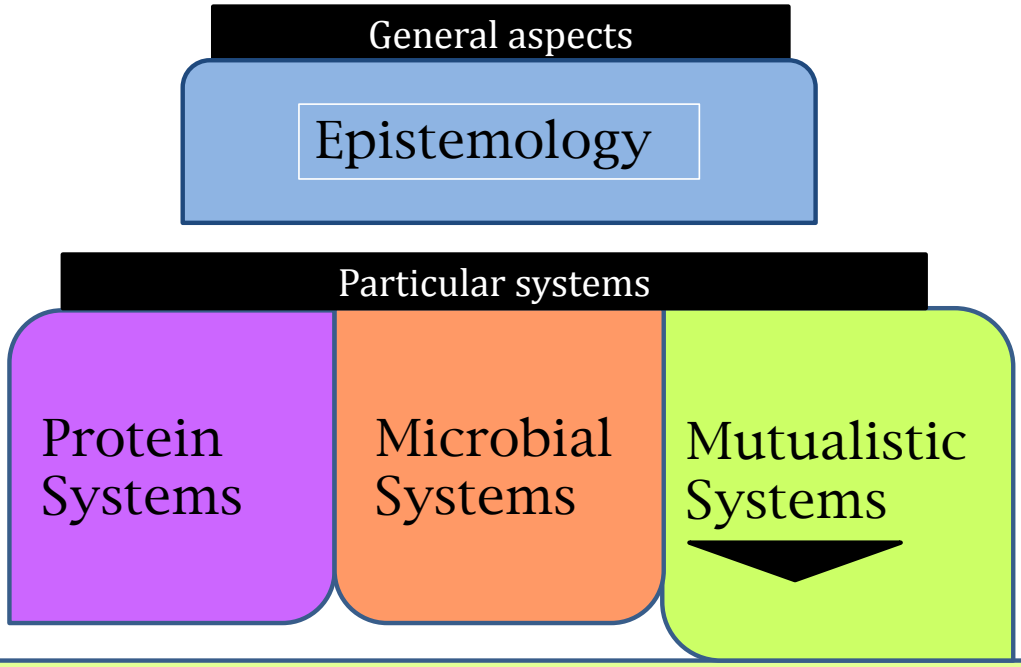
Part VI

Appendix

Appendix A

Supplementary Materials Mutualistic Systems

A.1. Supplementary Material Article [MUT-1]



1) Bastolla U, Fortuna MA, Pascual-García A, Ferrera A, Luque B, Bascompte J. (2009) The architecture of mutualistic networks minimizes competition and increases biodiversity. *Nature.* 458(7241):1018-20

2) Pascual-García A ., (2010) Explorando el rol de la Competición, el Mutualismo y la Arquitectura en Redes Ecológicas: ¿Qué podemos decir sobre la Biodiversidad?
Published in: *Evolución y Adaptación: 150 años después del origen de las especies.* Editors: Hernán Dopazo and Arcadi Navarro.
ISBN 978-84-92910-06-9

3) Pascual-García A., Ferrera A., and Bastolla, U. (2014) Does mutualism hinder biodiversity? arXiv preprint arXiv:1409.1683

4) Pascual-García A ., Ferrera A., and Bastolla, U. (2015) Effective competition determines the structural stability of model ecosystems. *Under revision.*

5) Ferrera A., Pascual-García A ., and Bastolla, U. (2015) Effective competition determines the global stability of model ecosystems. *Under revision.*

6) Pascual-García A., Bastolla U., (2015) The complexity-stability relation of mutualistic systems reconciles MacArthur and May. *Under revision.*

doi: 10.1038/nature07950 nature

SUPPLEMENTARY INFORMATION

1 Community model

We consider here model communities of plants and pollinators or seed dispersers where species in the same group are in competition between each other and interact mutualistically with species in the other group. We represent through $N_i^{(\mathrm{P})}$ and $N_k^{(\mathrm{A})}$ the species abundance density of the i-th species of plant and the k-th species of animal respectively, and with $S^{(\mathrm{P})}$ and $S^{(\mathrm{A})}$ the observed number of such species. The intrinsic growth rates in the absence of competition and mutualism are represented as $\alpha_i^{(\mathrm{P})}$ for plants and $\alpha_k^{(\mathrm{A})}$ for animals. In the latter case, they may be either positive or negative, representing the difference between the growth rate in the absence of any plant and the death rate. This choice has no relevant effect on the qualitative results.

For the sake of mathematical simplicity, we represent direct competitive interactions between species i and j through a linear functional response, $-\beta_{ij}^{(\mathrm{P})} N_j^{(\mathrm{P})}$ in the case of plants and $-\beta_{kl}^{(\mathrm{A})} N_l^{(\mathrm{A})}$ for animals. The competition matrices $\beta_{ij}^{(\mathrm{P,A})}$ are assumed to be symmetric and positive, with all positive or zero elements, both for plants and for animals.

The mutualistic interactions between plants and animals are modeled through non-linear functional responses of Holling Type II[1], $f(N) = (\gamma N) / (1 + h\gamma N)$. The denominator of the Holling term slows down the functional response when the densities are large, $N \approx 1/h\gamma$, limiting the maximum growth rate as $1/h$ and preventing it from diverging in the large N limit. The parameter h can be interpreted as a handling time. The mutualistic interactions network is described by a matrix $\gamma_{ij}^{(\mathrm{P})}$ whose non-negative elements represent the increase of the growth rate of the plant species i per unit of animal biomass j, in the limit of very small animal biomass. Similarly, $\gamma_{ji}^{(\mathrm{A})}$ has all non-negative elements that represent the increase of the growth rate of the animal species j per unit of plant biomass i.

doi: 10.1038/nature07950 SUPPLEMENTARY INFORMATION

The resulting dynamical equations for the plant populations are

$$\frac{1}{N_i^{(\mathrm{P})}}\frac{\mathrm{d}N_i^{(\mathrm{P})}}{\mathrm{d}t} = \alpha_i^{(\mathrm{P})} - \sum_{j\in\mathbf{P}} \beta_{ij}^{(\mathrm{P})} N_j^{(\mathrm{P})} + \sum_{k\in\mathbf{A}} \frac{\gamma_{ik}^{(\mathrm{P})} N_k^{(\mathrm{A})}}{1 + h^{(\mathrm{P})} \sum_{l\in\mathbf{A}} \gamma_{il}^{(\mathrm{P})} N_l^{(\mathrm{A})}} \,. \tag{1}$$

The equations for animal populations can be written in a symmetric form interchanging the indices A and P. When not otherwise stated, we will in the following only write down equations for plants.

In the next section we analyze the fixed points and dynamical stability of the model, leaving for section 3 the interesting problem of how the effective competition limits the structural stability of the model, and correspondingly its maximum biodiversity. Section 4 is a summary of the main results of the stability analysis. This is followed in Section 5 by an analysis of an assembling network, showing that a new species entering the community is favored by interacting with the most generalist species. Section 6 explores the robustness of our analytical results when other interaction types are included (6.1) and when departing from the mean field assumptions using numerical simulations (6.2). We conclude this online material with three appendices, one with the proof of the dynamical stability condition, the second with the numerical calculation of the predicted maximum biodiversity, and the last one presenting the mutualistic networks analyzed in this paper.

2 Fixed points and dynamical stability

We will consider here the fixed points of the dynamical system, defined by the equations $\mathrm{d}N_i^{(\mathrm{A,P})}/\mathrm{d}t = 0$, and analyze their stability. In order to get analytic expressions, we will exploit the fact that the handling time h is small compared with the typical intrinsic time of growth $1/\alpha$. We find two different types of solution. The first one is characterized by small equilibrium biomasses, $N \ll 1/h\gamma$. In this limit, we can expand the functional

doi: 10.1038/nature07950 SUPPLEMENTARY INFORMATION

response in a Taylor series, whose dominant term yields a linear system of fixed point equations. We call this regime *weak mutualism*. A second type of fixed points correspond to equilibrium biomasses N of order $1/h\gamma$. In this case the linear system is not a valid approximation, but it is now possible to get analytic insight by neglecting the terms $h\alpha$ with respect to $h\gamma N$. We call this regime *strong mutualism*.

Furthermore, in order to simplify the analytic expressions, we will consider mainly direct competition matrices $\beta_{ij}^{(\mathrm{P})}$ of mean field type, with $\beta_{ij}^{(\mathrm{P})} = \beta_0^{(\mathrm{P})}\left(\rho^{(\mathrm{P})} + (1-\rho^{(\mathrm{P})})\delta_{ij}\right)$ (see ref. 2), where δ_{ij} is Kronecker's delta (one if $i = j$ and zero otherwise). The dimensionless parameters $\rho^{(\mathrm{P})} < 1$ measure the extent of interspecific competition between different species of the same group.

2.1 Pure competition

For a purely competitive system, i.e., $\gamma_{ik} \equiv 0$, the fixed point densities $\{N_i\}$ satisfy the system of equations

$$\sum_{ij} \beta_{ij}^{(\mathrm{P})} N_j^{(\mathrm{P})} = \alpha_i^{(\mathrm{P})} . \tag{2}$$

The analytical expressions are symmetrical for the case of the animals. The necessary and sufficient conditions for dynamic stability are that (i) all equilibrium biomasses must be positive; and (ii) the direct competition matrix β must be positive definite.

2.2 Weak mutualism: mean field

We now integrate mutualistic interactions into the competitive community. If the equilibrium densities are small, $N \ll 1/h\gamma$, which is a valid approximation within the weak mutualism regime, the fixed point equations for the plant communities at the dominant order in h can be written in the form of a linear system,

doi: 10.1038/nature07950 SUPPLEMENTARY INFORMATION

$$\sum_j C_{ij}^{(\mathrm{P})} N_j^{(\mathrm{P})} = p_i^{(\mathrm{P})} \,. \tag{3}$$

These equations are mathematically equivalent to the fixed points of a purely competitive system, Eq.(2). We call the vector p_i *effective productivity* and the matrix C_{ij} *effective competition*. We will prove in Appendix A that, in analogy with the purely competitive system, the equilibrium fixed point is stable if and only if the effective competition matrix is positive (i.e., all of its eigenvalues are positive) and all the equilibrium densities are positive.

At zero order in h, the effective productivity and the effective competition are given by the expressions

$$p_i^{(\mathrm{P})} = \alpha_i^{(\mathrm{P})} + \sum_{k,l} \gamma_{ik}^{(\mathrm{P})} \left(\beta^{(\mathrm{A})}\right)_{kl}^{(-1)} \alpha_l^{(\mathrm{A})}, \tag{4}$$

$$C_{ij}^{(\mathrm{P})} = \beta_{ij}^{(\mathrm{P})} - \sum_{k,l} \gamma_{ik}^{(\mathrm{P})} \left(\beta^{(\mathrm{A})}\right)_{kl}^{(-1)} \gamma_{lj}^{(\mathrm{A})}. \tag{5}$$

First order corrections in h are straightforward to compute, and do not change the qualitative picture. They will be omitted in the following.

We first consider mean field mutualist interactions, with all species of plants and animals interacting between each other with equal per capita mutualistic effect, $\gamma_{ik}^{(\mathrm{A,P})} = \gamma_0^{(\mathrm{A,P})}$. We will relax this assumption later on. With this assumption, the effective competition matrix turns out to be of mean field type,

$$C_{ij}^{(\mathrm{P})} = \beta_0^{(\mathrm{P})}(1-a^{(\mathrm{P})}) \left[\delta_{ij}\left(1-\rho_{\mathrm{mut}}^{(\mathrm{P})}\right) + \rho_{\mathrm{mut}}^{(\mathrm{P})}\right] , \tag{6}$$

$$a^{(\mathrm{P})} = \frac{\gamma_0^{(\mathrm{P})}\gamma_0^{(\mathrm{A})}}{\beta_0^{(\mathrm{A})}\beta_0^{(\mathrm{P})}} \frac{S^{(\mathrm{A})}}{(S^{(\mathrm{A})}\rho^{(\mathrm{A})} + (1-\rho^{(\mathrm{A})}))} \,. \tag{7}$$

doi: 10.1038/nature07950 SUPPLEMENTARY INFORMATION

The effective interspecific competition is given by

$$\rho_{\rm mut}^{(\rm P)} = \frac{\rho^{(\rm P)} - a^{(\rm P)}}{1 - a^{(\rm P)}} < \rho^{(\rm P)}. \tag{8}$$

We see from the above expression that, for the mean field system and for values of the parameter $a^{(\rm P)} \in [0, \rho^{(\rm P)}]$, the effective interspecies competition $\rho_{\rm mut}^{(\rm P)}$ is smaller than the bare competition $\rho^{(\rm P)}$, i.e., mutualistic interactions of mean field type reduce the effective interspecies competition. Eq. (8) is valid for $a^{(\rm P)} < \rho^{(\rm P)} + (1 - \rho^{(\rm P)})/S^{(\rm P)}$. At this point, the main eigenvalue λ_1 of the effective competition matrix becomes negative and the community enters into the strong mutualism regime.

Stability of the weak mutualism fixed point requires that the effective competition matrix is positive. The eigenvalues are $\lambda_1 = (1-a)\left(S\rho_{\rm mut}^{(\rm P)} + (1 - \rho_{\rm mut}^{(\rm P)})\right) = S(\rho^{(\rm P)} - a) + (1 - \rho^{(\rm P)})$ and $\lambda_k = (1-a)(1 - \rho_{\rm mut}^{(\rm P)}) = (1 - \rho^{(\rm P)})$ $(k > 1)$. Positivity of the competition matrix requires that $S(\rho^{(\rm P)} - a) + (1 - \rho^{(\rm P)}) > 0$, which in turn yields the condition

$$\gamma_0^{(\rm P)}\gamma_0^{(\rm A)} < \beta_0^{(\rm P)}\beta_0^{(\rm A)} \left(\rho^{(\rm A)} + \frac{1 - \rho^{(\rm A)}}{S^{(\rm A)}}\right)\left(\rho^{(\rm P)} + \frac{1 - \rho^{(\rm P)}}{S^{(\rm P)}}\right), \tag{9}$$

which generalizes the result presented in ref. 3 to the case where the interspecific competition is not zero. Notice that, if $\rho^{(\rm P)}$ and $\rho^{(\rm A)}$ are not zero, the maximum value of mutualistic interactions in the weak mutualism regime does not vanish for large ecosystems (large $S^{(\rm P)}$ and $S^{(\rm A)}$), but it is limited as $\gamma_0^{(\rm P)}\gamma_0^{(\rm A)} < \beta_0^{(\rm P)}\beta_0^{(\rm A)}\rho^{(\rm A)}\rho^{(\rm P)}$.

2.3 Strong mutualism: mean field

For mutualistic interactions stronger than Eq.(9) the weak mutualism fixed point is not stable, and we have to consider the strong regime in which the equilibrium biomasses are of order $1/h$. In order to get analytic results, we neglect higher order terms in h, such as

doi: 10.1038/nature07950 SUPPLEMENTARY INFORMATION

$h\alpha$. We consider mean field systems in which all pairs of species interact with the same strength. In this case, positivity of the equilibrium biomasses requires that

$$\gamma_0^{(\mathrm{P})}\gamma_0^{(\mathrm{A})} > \beta_0^{(\mathrm{P})}\beta_0^{(\mathrm{A})}\left(\rho^{(\mathrm{A})}+\frac{1-\rho^{(\mathrm{A})}}{S^{(\mathrm{A})}}\right)\left(\rho^{(\mathrm{P})}+\frac{1-\rho^{(\mathrm{P})}}{S^{(\mathrm{P})}}\right). \tag{10}$$

Therefore, we see from Eq.(9) that, as soon as the weak mutualism fixed point ceases to be stable, the strong mutualism fixed point becomes stable. For the mean field case, the strong mutualism fixed point allows coexistence of an arbitrary number of species, independent of the values of the intrinsic growth rates $\alpha_i^{(\mathrm{P,A})}$.

2.4 Strong and weak mutualism can not coexist

In the general case, stability of the strong mutualism fixed point with positive densities requires that the effective competition matrices $C^{(\mathrm{A})}$ and $C^{(\mathrm{P})}$ are not positive definite, i.e., at least one of their eigenvalues is negative or zero. The proof goes like this. In the strong mutualist regime we can neglect the terms $h\alpha_i$ and the fixed point equations are

$$N^{(\mathrm{P})} = \sum_j \left(\beta^{(\mathrm{P})}\right)^{-1}_{ij} \sum_k \frac{\gamma_{ik}^{(\mathrm{P})} N_k^{(\mathrm{A})}}{1+h^{(\mathrm{P})}\sum_{l\in\mathbf{A}}\gamma_{il}^{(\mathrm{P})}N_l^{(\mathrm{A})}}\,. \tag{11}$$

Since $\gamma N/(1+h\gamma N) \leq \gamma N$, it follows that, in the strong mutualism regime,

$$N_i^{(\mathrm{P})} \leq \sum_j \left(\left(\beta^{(\mathrm{P})}\right)^{-1}\gamma^{(\mathrm{P})}\left(\beta^{(\mathrm{A})}\right)^{-1}\gamma^{(\mathrm{A})}\right)_{ij} N_j^{(\mathrm{P})} \equiv \sum_j M_{ij}^{(\mathrm{P})} N_j^{(\mathrm{P})}\,. \tag{12}$$

We have defined here the mutualistic matrix $M^{(\mathrm{P})} \equiv \left(\beta^{(\mathrm{P})}\right)^{-1}\gamma^{(\mathrm{P})}\left(\beta^{(\mathrm{A})}\right)^{-1}\gamma^{(\mathrm{A})}$. The effective competition matrix $C^{(\mathrm{P})}$ can be written in matrix notation as $C^{(\mathrm{P})} \equiv \beta^{(\mathrm{P})}(I-M)$, where I is the identity matrix. Since the direct competition matrix $\beta^{(\mathrm{P})}$ is positive, if $C^{(\mathrm{P})}$ is positive then all eigenvalues of M must fulfill $\lambda(M) < 1$ (see Appendix A). Together with Eq.(12), this implies that the solutions of the fixed point equation must have $N_i^{(\mathrm{P})} \leq 0$

doi: 10.1038/nature07950 SUPPLEMENTARY INFORMATION

for all i. The same applies to $N^{(\mathrm{A})}$ if $C^{(\mathrm{A})}$ is positive. Therefore, if the weak mutualism fixed point is stable, no stable strong mutualism fixed point can exist.

If we relax the mean field assumption, the behavior of the strong mutualism regime changes dramatically. To get some flavour of this, we examined the simplest possible system, where all species are below the weak mutualism threshold and interact with coefficients $\gamma_0^{(\mathrm{P})}$ and $\gamma_0^{(\mathrm{A})}$ such that $\gamma_0^{(\mathrm{P})}\gamma_0^{(\mathrm{A})} < \beta_0^{(\mathrm{P})}\beta_0^{(\mathrm{A})}\rho^{(\mathrm{P})}\rho^{(\mathrm{A})}$, but one pair of species, animal species 1 and plant species 1, have a strong mutualistic interaction with interaction coefficients $\gamma_1^{(\mathrm{P})}\gamma_1^{(\mathrm{A})} > \beta_0^{(\mathrm{P})}\beta_0^{(\mathrm{A})}$. Solving the fixed point equations and considering all possible cases, it can be shown that in this case there is no possible fixed point where the species below the strong mutualism threshold have positive biomass. It follows from this analysis that, if one pair of species overcomes the strong mutualism threshold, no other species below the threshold can coexist with them at any fixed point. Notice that we have considered the best possible interaction matrices γ_{ij}, since all species below the threshold are assumed to interact with all other species, including the strong interacting ones. To allow coexistence, it would be necessary to relax the hypothesis that all other species are directly competing with the strong interacting species.

Therefore, the model predicts that, when a pair of species overcomes the strong mutualism threshold while the other species remain below it, all species below the threshold become extint, pointing out the interesting possibility of mutualism-induced extinctions. We will analyze this regime in more detail in a forthcoming work.

3 Structural stability and biodiversity

Dynamic stability is an important requirement for the fixed point of a model ecosystem to represent properties of a real ecological community. Despite being neither a necesary, nor a sufficient requirement, local dynamic stability is a simple analytic criterion that has

doi: 10.1038/nature07950 SUPPLEMENTARY INFORMATION

been widely used in theoretical ecology, giving very interesting insights on the properties of real communities. However, we believe that other interesting insights can be gained by considering structural stability, i.e., the stability with respect to modifications in the parameters of the dynamical system. In this section, structural stability is meant as the volume in parameter space compatible with positive densities at the fixed point. Interestingly, for competitive systems structural stability in this meaning is negatively correlated with the number of species in the system, so that, by considering a minimum variance of the parameters compatible with the environmental variability, we can predict the maximum number of species that can coexist in the system[4]. We show here that it is possible to extend this analytic insight also to communities in which competition and mutualism coexist. Notice that the existence of a direct term of interspecific competition, $\rho^{(\mathrm{P})}$, is an essential characteristic of our model with respect to other models in the literature, which alters completely its properties of structural stability.

3.1 Effective competition and structural stability

We consider here a species community in which the fixed point equations can be written in the form $\sum_j C_{ij} N_j = p_i$. We refer to C_{ij} as the effective competition matrix and p_i as the effective productivity vector. This formulation is rather general. It is suitable to represent a purely competitive system, in which $C_{ij} = \beta_{ij}$ and $p_i = \alpha_i$, a system with predation (see ref. 4) or a system with weak mutualism, in which the effective competition matrix and the effective productivity vector are given by equations (5) and (4), respectively.

It is convenient to normalize the effective competition matrix as

$$B_{ij} = \frac{C_{ij}}{\sqrt{C_{ii} C_{jj}}}\,, \tag{13}$$

in such a way that $B_{ii} = 1$. From the main eigenvalue of this matrix, $\lambda_1(B)$, we can derive

doi: 10.1038/nature07950 SUPPLEMENTARY INFORMATION

the effective interspecific competition parameter $\tilde{\rho}$ as

$$\tilde{\rho} \equiv \frac{\lambda_1(B) - 1}{S - 1}. \tag{14}$$

If the effective competition matrix is a direct competition matrix of mean field type, $B_{ij} = \rho + (1 - \rho)\delta_{ij}$, it holds $\lambda_1(B) = S\rho + (1 - \rho)$ and, consequently, $\tilde{\rho} = \rho$. Thus, the quantity $\tilde{\rho}$ measures the effective interspecific competition, generalizing the mean field parameter ρ. Positivity of all equilibrium densities imposes more and more stringent conditions on the effective productivity parameters $\{p_i\}$ for increasing number of species and interspecific competition $\tilde{\rho}$ (see ref. 4). This result generalizes the mean field result in refs. 2, 5 and 6. In other words, the larger is $\tilde{\rho}$, the less structurally stable the system is, in the sense that the productivity vectors must be fine tuned in order to get positive equilibrium densities. Assuming that the fluctuations of the productivity vector are limited from below by the environmental variability Δ, we obtain the following limit to the maximum biodiversity S (ref. 4)

$$S \leq 1 + \left(\frac{1 - \tilde{\rho}}{\tilde{\rho}}\right)\left(\frac{\lambda_2(B)/(1 - \tilde{\rho}) - \Delta}{\Delta}\right). \tag{15}$$

For mean field competition matrices, it holds $\tilde{\rho} = \rho$ and $\lambda_2(B) = 1 - \rho$, whence $S \leq \overline{S}(1 - \Delta)/\Delta$. Therefore, we define the maximum biodiversity parameter

$$\overline{S} \equiv \frac{1 - \tilde{\rho}}{\tilde{\rho}} \tag{16}$$

which sets the scale for the maximum biodiversity that a competitive community can host.

In this work, since we use direct competition matrices of mean field type, we will use the notation $\overline{S} = (1 - \rho)/\rho$ for the maximum biodiversity for the purely competitive system,

doi: 10.1038/nature07950 SUPPLEMENTARY INFORMATION

$\rho_{\rm mut}$ for the effective interspecific competition in the presence of mutualistic interactions, and $\overline{S}_{\rm mut} = (1-\rho_{\rm mut})/\rho_{\rm mut}$ for the maximum biodiveristy in the presence of mutualistic interactions.

3.2 Weak mutualism: mean field

The above calculations remain valid in the weak mutualist regime. As we have seen, if the mutualistic interactions and the direct competition matrix are of mean field type, the effective competition matrix is also mean field, and the effective interspecific competition parameter can be analytically computed as in Eq.(8). We see from this equation that mean field mutualism reduces the interspecific competition, thereby increasing the number of species that can stably stay in the system, which is now given by

$$\overline{S}^{\rm (P)}_{\rm mut} \equiv \frac{1-\rho^{\rm (P)}_{\rm mut}}{\rho^{\rm (P)}_{\rm mut}} = \frac{\overline{S}^{\rm (P)}}{1-a^{\rm (P)}/\rho^{\rm (P)}}\,. \tag{17}$$

3.3 Weak mutualism beyond the mean field: nestedness

Now we relax the mean field assumption that plant and animal species interact mutualistically with all species in the other group, but for mathematical simplicity, we mantain the assumption that the strength of all existing mutualistic interactions are equal. We will refer to this model as the *soft mean field*. Therefore, we can define a binary matrix g_{ik} whose elements are one if the link is present and zero otherwise, such that $\gamma_{ik} = \gamma_0 g_{ik}$. It holds that $g^{\rm (P)}_{ik} = g^{\rm (A)}_{ki}$. We further denote the number of links of plant i as $n^{\rm (P)}_i = \sum_k g^{\rm (P)}_{ik}$ and the number of common links of plants i and j as

$$n^{\rm (P)}_{ij} \equiv \sum_k g^{\rm (P)}_{ik} g^{\rm (P)}_{jk} \tag{18}$$

which we call the *overlap matrix*.

doi: 10.1038/nature07950 SUPPLEMENTARY INFORMATION

The elements of the overlap matrix fulfill the inequalities $n_{ij} \leq \min(n_i, n_j)$. We say that the interaction matrix g is maximally nested if all the elements of the overlap matrix take their maximum possible value, namely if $n_{ij} = \min(n_i, n_j)$. This definition coincides with the definition of a maximally nested matrix in terms of the nesting algorithm[7]. This algorithm proceeds by what we will refer to from now on as *nesting steps*. Each nesting step tries to exchange two matrix elements with a common index, either a column or a row, for instance γ_{ik} and γ_{jk}. The move is accepted if the nonzero element is moved to a row (column) whose number of links after the move is larger than the number of links in the original row (column). No row (column) is allowed to be left without any link. A maximally nested matrix is a matrix which can not be changed anymore through this algorithm. This is the case when $n_{ij} = \min(n_i, n_j)$. We therefore define the nestedness of the matrix g with respect to plants as

$$\eta^{(\mathrm{P})} = \frac{\sum_{i<j} n_{ij}^{(\mathrm{P})}}{\sum_{i<j} \min\left(n_i^{(\mathrm{P})}, n_j^{(\mathrm{P})}\right)}. \tag{19}$$

The symmetric definition holds for the nestedness with respect to animals. It is easy to see that the nestedness defined above is zero if $n_{ij}^{(\mathrm{P})} \equiv 0$, which we define as anti-nested interactions, and one for perfect nestedness $n_{ij}^{(\mathrm{P})} \equiv \min\left(n_i^{(\mathrm{P})}, n_j^{(\mathrm{P})}\right)$. For random networks with the same number of independent interactions as in the real network, the average nestedness is $\eta_{\mathrm{rand}}^{(\mathrm{P})} = \sum_i n_i^{(\mathrm{P})} / \left(S^{(\mathrm{P})} S^{(\mathrm{A})}\right)$.

3.4 Weak mutualism beyond the mean field: soft mean field

In order to get a simple analytical formula that explicitates the influence of the network architecture, we introduce here the soft mean field model, in which all parameters are equal but the mutualistic network is not fully connected as in the mean field case. In this model, the direct competition matrix β_{ij} is of mean field type and all non-zero mutualistic

doi: 10.1038/nature07950 SUPPLEMENTARY INFORMATION

interactions are equal, $\gamma_{ij}^{(\mathrm{P})} = \gamma_0^{(\mathrm{P})} g_{ij}^{(\mathrm{P})}$, where the binary matrix $g_{ij}^{(\mathrm{P})}$ is the adjacency matrix of the mutualistic network and $g_{ij}^{(\mathrm{A})}$ is the transpose of $g_{ij}^{(\mathrm{P})}$.

We still have to specify the intrinsic growth rates α_i (or death rates, if they are negative). For this purpose, we explicitate the effective productivity vector from Eq. (4),

$$p_i^{(\mathrm{P})} = \alpha_i^{(\mathrm{P})} + \left(\frac{\gamma_0^{(\mathrm{P})}}{\beta_0^{(\mathrm{A})}(1-\rho^{(\mathrm{A})})} \right) \left[\sum_j g_{ij}^{(\mathrm{P})} \alpha_j^{(\mathrm{A})} - \frac{S^{(\mathrm{A})}}{S^{(\mathrm{A})} + \overline{S}^{(\mathrm{A})}} \overline{\alpha^{(\mathrm{A})}} \right], \tag{20}$$

where $\overline{\alpha^{(\mathrm{A})}}$ is the average growth rate (or death rate, if α is negative) of species of type (A).

As we have seen, a necessary condition for species coexistence under competition is that the effective productivity vector has a narrow distribution. Therefore, we assume that the evolutionary process building the community leads to a narrow distribution of effective productivities, and that its dispersion Δ, which appears in Eq. (15), is the smallest one compatible with the unavoidable environmental variability, and it does not change in the presence or in the absence of mutualism. This assumption, which has to be justified through an explicit model of network assembly, implies that the α_i must be chosen negatively correlated to the number of mutualistic links, and it allows us to concentrate the focus of our analytic computation on the effective competition matrix This depends on the network architecture but it does not depend on the α_i, which will not play any role in the following analytic computation.

Alternatively, we could formulate the soft mean field model in such a way that the more mutualistic links a species has, the weaker these links are, defining the mutualistic parameters as $\gamma_{ij} = \gamma_0 / f(n_i) g_{ij}$, where $f(n_i)$ is a growing function of the number of links. This equation assumes that specialist species are more efficient than generalist species in dealing with their mutualistic partner, which is a quite plausible assumption. In

doi: 10.1038/nature07950 SUPPLEMENTARY INFORMATION

this formulation, the function $f(n)$ should be chosen such that the effective productivity vector given by Eq.(4) is uncorrelated with the number of mutualistic interactions n_i. This formulation of the soft mean field model would lead to different expressions for the effective competitivity matrix from the one that we present below, and we will study this formulation in following work.

In the weak mutualism regime, the normalized competition matrix $B_{ij}^{(\mathrm{P})}$ is defined through

$$\frac{C_{ij}^{(\mathrm{P})}}{\beta_0^{(\mathrm{P})}(1-\rho^{(\mathrm{P})})} = \delta_{ij} + \frac{1}{\overline{S}^{(\mathrm{P})}} + R\left(\frac{n_i^{(\mathrm{P})} n_j^{(\mathrm{P})}}{S^{(\mathrm{A})} + \overline{S}^{(\mathrm{A})}} - n_{ij}^{(\mathrm{P})}\right), \tag{21}$$

$$B_{ij}^{(\mathrm{P})} = \frac{C_{ij}^{(\mathrm{P})}}{\sqrt{C_{ii}^{(\mathrm{P})} C_{jj}^{(\mathrm{P})}}}, \tag{22}$$

where

$$R = \frac{\gamma_0^{(\mathrm{P})}\gamma_0^{(\mathrm{A})}}{\beta_0^{(\mathrm{P})}\beta_0^{(\mathrm{A})}(1-\rho^{(\mathrm{P})})(1-\rho^{(\mathrm{A})})}. \tag{23}$$

Notice that the matrix $B^{(\mathrm{P})}$ depends only on three numerical parameters, R, $\overline{S}^{(\mathrm{P})}$ and $\overline{S}^{(\mathrm{A})}$.

To get more analytic insight on how mutualism influences biodiversity, we computed the derivative of the main eigenvalue of the normalized effective competition matrix, $\lambda_1\left(B^{(\mathrm{P})}\right)$, with respect to the mutualistm-to-competition ratio R at the point $R = 0$ (absence of mutualism) This calculation shows that the effective interspecific competition decreases with the nestedness of the mutualist interaction matrix for a given distribution of number of links $\{n_i^{(\mathrm{P})}\}$ and fixed parameters. Since the maximum predicted biodiversity $\overline{S}_{\mathrm{mut}}^{(\mathrm{P})}$ increases with decreasing effective competition, the model predicts that, for perfectly nested mutualist networks, the effective competition is weakest and the maximum biodiversity is largest. Therefore, nested mutualist interactions favor biodiversity.

doi: 10.1038/nature07950 SUPPLEMENTARY INFORMATION

Specifically, calculating the derivative of $\lambda_1\left(B^{(\mathrm{P})}\right)$, we can easily obtain the derivative of the maximum biodiversity $\overline{S}^{(\mathrm{P})}_{\mathrm{mut}} = \left(1-\rho^{(\mathrm{P})}_{\mathrm{mut}}\right)/\rho^{(\mathrm{P})}_{\mathrm{mut}}$, where $\rho^{(\mathrm{P})}_{\mathrm{mut}} = \left(\lambda_1\left(B^{(\mathrm{P})}\right)-1\right)/(S^{(\mathrm{P})}-1)$ is the effective interspecific competition parameter, with respect to the mutualism-to-competition ratio. This measures the relative increment of the maximum biodiversity due to mutualism, and is equal to

$$\frac{1}{\overline{S}^{(\mathrm{P})}_{\mathrm{mut}}}\left.\frac{\partial \overline{S}^{(\mathrm{P})}_{\mathrm{mut}}}{\partial R}\right|_{R=0} = \left(1+\frac{1}{\overline{S}^{(\mathrm{P})}}\right)\langle n^{(\mathrm{P})}\rangle \quad \left[\overline{S}^{(\mathrm{P})}\left(\hat{\eta}^{(\mathrm{P})}-\frac{\langle n^{(\mathrm{P})}\rangle}{S^{(\mathrm{A})}+\overline{S}^{(\mathrm{A})}}\right)-(1-\hat{\eta}^{(\mathrm{P})}) + \frac{\langle (n^{(\mathrm{P})})^2\rangle-\langle n^{(\mathrm{P})}\rangle^2}{\langle n^{(\mathrm{P})}\rangle(S^{(\mathrm{A})}+\overline{S}^{(\mathrm{A})})}\left(\frac{S^{(\mathrm{P})}+\overline{S}^{(\mathrm{P})}}{S^{(\mathrm{P})}-1}\right)\right], \qquad (24)$$

where $\langle n^{(\mathrm{P})}\rangle = \sum_i n_i^{(\mathrm{P})}/S^{(\mathrm{P})}$ and $\langle \left(n^{(\mathrm{P})}\right)^2\rangle = \sum_i \left(n_i^{(\mathrm{P})}\right)^2/S^{(\mathrm{P})}$ are the mean and mean square number of mutualistic interactions per plant species. The parameter $\hat{\eta}^{(\mathrm{P})} = \sum_{i\neq j} n_{ij}^{(\mathrm{P})}/\left((S^{(\mathrm{P})}-1)\sum_i n_i^{(\mathrm{P})}\right)$ is very strongly correlated with the nestedness defined in Eq.(19) (for real networks, the correlation coefficient between nest and the nestedness parameter is 0.97).

The derivative in the above equation is not bound to be positive. In particular, the derivative is typically negative if there are few shared interactions (small $\hat{\eta}$) together with strong direct competition (small $\overline{S}^{(\mathrm{A})}$), so that the term $\hat{\eta}-\langle n^{(\mathrm{P})}\rangle/(S^{(\mathrm{A})}+\overline{S}^{(\mathrm{A})})$ is negative. This result shows that mutualism can also increase the effective competition and hinder biodiversity. Although it looks counter-intuitive, this result can be easily understood by considering that, if plant species i and j do not share any animal species ($n_{ij}^{(\mathrm{P})}=0$), the direct competition between the animals interacting with them has the net effect to increase the effective competition that i and j experience. This illustrates how the direct competition for resources explicitly described by the β_{ij} terms in Eq. (29) is now mediated by the use of a common set of mutualismstic partners.

A second more stringent condition for mutualism to enhance the maximum biodiversity is that the reduction in interspecific effective competition $C_{i\neq j}$ must be larger than

doi: 10.1038/nature07950 SUPPLEMENTARY INFORMATION

tice that, for networks in which $\hat{\eta}$ attains the maximum possible value $\hat{\eta} = 1$, as the fully connected mean field network, the increment of biodiversity Eq.(24) is always positive, independent of the parameters $\overline{S}^{(\mathrm{A})}$ and $\overline{S}^{(\mathrm{P})}$.

4 Summary of the stability analysis

We have shown here that (1) the weak mutualism fixed point is stable if and only if all equilibrium densities are positive and both matrices $C^{(\mathrm{A})}$ and $C^{(\mathrm{P})}$ are positive definite; (2) the strong mutualist fixed point can not have positive densities if the matrices $C^{(\mathrm{A})}$ and $C^{(\mathrm{P})}$ are positive definite; (3) when the system is in the weak mutualism regime, $\lambda_1(B)$ is positive. The limitation to biodiversity imposed by competition becomes less stringent as $\lambda_1(B)$ decreases, and they disappear when $\lambda_1(B) = 1$, implying through Eq.(14) that $\rho_{\mathrm{mut}} = 0$. The maximum biodiversity that each group (plants or animals) can attain is controlled by the main eigenvalue of the normalized effective competition matrix B, and it is larger, the smaller is this eigenvalue, $\lambda_1(B)$. As soon as $\lambda_1(B)$ becomes negative, the weak mutualism fixed point looses its stability, and the strong mutualism fixed point can become stable. However, if only one pair of species overcomes the strong mutualism threshold while the other species still remain below it, the model predicts that only the species above threshold will eventually survive, whereas the other species will go extinct, thus suggesting the interesting possibility of massive extinctions caused by mutualism.

doi: 10.1038/nature07950 SUPPLEMENTARY INFORMATION

5 Assembly of mutualistic networks

Consider the arrival of a new animal species into a community in the weak mutualism regime (for plants, the mathematical treatment would be exactly symmetric). We will assume that the new species, labelled as 1, is specialist, i.e., it can interact only with one plant species, also labeled as 1. We will show that, if plant species 1 is generalist, the animal species 1 will experience the lowest competitive load, and it will be incorporated most likely in the community.

To prove our thesis, let us consider the effective competition matrix elements for the new animal species 1:

$$C_{1j}^{(\mathrm{A})} = \beta_{1j}^{(\mathrm{A})} - \gamma_{11}^{(\mathrm{A})} \sum_k \left(\beta^{(\mathrm{P})}\right)_{1k}^{(-1)} \gamma_{kj}^{(\mathrm{P})}, \tag{25}$$

where we have explicitly used the fact that the new insect species 1 is specialist. Let us now consider for the sake of simplicity a direct competition matrix $\beta^{(\mathrm{P})}$ of mean field type. The analytic expression for the inverse matrix is $\left(\beta^{(\mathrm{P})}\right)^{-1} = 1/\beta_0^{(\mathrm{P})}(1-\rho^{(\mathrm{P})})\left(\delta_{ij} + 1/(S^{(\mathrm{P})} + \overline{S}^{(\mathrm{P})}\right)$, whence

$$C_{1j}^{(\mathrm{A})} = \beta_{1j}^{(\mathrm{A})} - \frac{\gamma_{11}^{(\mathrm{A})}}{\beta_0^{(\mathrm{P})}(1-\rho^{(\mathrm{P})})}\left(\gamma_{1j}^{(\mathrm{P})} - \frac{1}{S^{(\mathrm{P})} + \overline{S}^{(\mathrm{P})}} \sum_k \gamma_{kj}^{(\mathrm{P})}\right). \tag{26}$$

Summing over all animal species j, we find

$$\sum_j C_{1j}^{(\mathrm{A})} = \sum_j \beta_{1j}^{(\mathrm{A})} - \frac{\gamma_{11}^{(\mathrm{A})}}{\beta_0^{(\mathrm{P})}(1-\rho^{(\mathrm{P})})}\left(\sum_j \gamma_{1j}^{(\mathrm{P})} - \frac{1}{S^{(\mathrm{P})} + \overline{S}^{(\mathrm{P})}} \sum_{jk} \gamma_{kj}^{(\mathrm{P})}\right). \tag{27}$$

The only term that depends on the plant species 1 with which the new animal interacts is $\sum_j \gamma_{1j}^{(\mathrm{P})}$. The larger is this term, the smaller the competition experienced by the new species. Now, although we expect that individual interaction coefficients $\gamma_{1j}^{(\mathrm{P})}$ tend to be larger for specialist species than for generalist species, we also expect that the sum of all

doi: 10.1038/nature07950 SUPPLEMENTARY INFORMATION

interactions $\sum_j \gamma_{1j}^{(\mathrm{P})}$ is largest if plant species 1 is generalist. Therefore, a specialist species is favoured if it interacts with a generalist species, producing nested interaction patterns. This result confirms a recent suggestion that we have to move beyond competition to predict ecosystem invasibility[8], and provides an analytical framework to quantify such an effect of positive interactions.

6 Robustness of our analytic results

Our previous results are based on an analytical solution of our model, i.e., on an analysis of the equilibrium. To obtain such analytic results we have had to make a series of assumptions. To begin with, the model only considers mutualistic and competitive interactions, and therefore one can wonder how robust are our results when other interaction types such as predation are considered. Second, our analyses are based on a mean field assumption whereby the values of competitive coefficients (β), for example, are the same across species. Similarly, we use a soft mean field approach to deal with mutualistic coefficients: while we address the real network of interactions, and therefore some interactions are zero, the observed interactions have the same value of mutualistic strength (γ). Finally, the emphasis on equilibrium precludes an analysis of the transient time before reaching this solution or other dynamic properties. In this section we explore the robustness of our results when these assumptions are relaxed and briefly address these other questions.

6.1 Introducing predation

The formalism of the effective competition matrix allows an analytic treatment of a generalized system including predation. This will allow us to test whether our results are qualitatively unchanged when another interaction type is considered.

doi: 10.1038/nature07950 SUPPLEMENTARY INFORMATION

We consider four groups of species: plants (P), animal pollinators or seed dispersers (A), herbivores (H) and consumers that predate animal mutualists (e.g., insectivorous birds that predate over pollinator insects) (C). The groups A and P are related through mutualistic interactions, the groups A and C are related through prey-predator interactions, and the same holds for the groups P and H. To simplify the mathematical treatment, we assume that no inteaction occurs between groups C and H (this assumption can be easily relaxed). Species within each group compete between each other. Assuming for simplicity a linear predator functional relationship of Lotka-Volterra type, the dynamic equations for plants and herbivore species are

$$\frac{1}{N_i^{(\mathrm{P})}}\frac{\mathrm{d}N_i^{(\mathrm{P})}}{\mathrm{d}t} = \alpha_i^{(\mathrm{P})} - \sum_{j\in\mathbf{P}} \beta_{ij}^{(\mathrm{P})} N_j^{(\mathrm{P})} + \sum_{k\in\mathbf{A}} \frac{\gamma_{ik}^{(\mathrm{P})} N_k^{(\mathrm{A})}}{1 + h^{(\mathrm{P})} \sum_{l\in\mathbf{A}} \gamma_{il}^{(\mathrm{P})} N_l^{(\mathrm{A})}} - \sum_{k\in\mathbf{H}} \delta_{ik}^{(\mathrm{P})} N_k^{(\mathrm{H})} \quad (28)$$

$$\frac{1}{N_i^{(\mathrm{H})}}\frac{\mathrm{d}N_i^{(\mathrm{H})}}{\mathrm{d}t} = \alpha_i^{(\mathrm{H})} - \sum_{j\in\mathbf{H}} \beta_{ij}^{(\mathrm{H})} N_j^{(\mathrm{H})} + \sum_{k\in\mathbf{P}} \delta_{ik}^{(\mathrm{H})} N_k^{(\mathrm{P})} \,. \quad (29)$$

The equations for mutualistic insect populations and for insectivorous can be written in a symmetric form interchanging the indices A and P and C and H, respectively. Here we use the same notation as in the paper: superscripts indicate the group of species, α is the vector of intrinsic growth rates, positive for plants and negative for animals, the matrix β represents intra-group competition, the matrix γ represents mutualistic interactions, and the matrix δ represents predator-prey relationships. Notice that in this way all possible kinds of pairwise ecological interactions are represented in the model.

In the weak mutualism regime, and in the small h approximation, the fixed point equations can be written, after some algebra, in the form

doi: 10.1038/nature07950 SUPPLEMENTARY INFORMATION

$$\sum_{ij} C_{ij}^{(\mathrm{P})} N_j^{(\mathrm{P})} = \pi_i^{(\mathrm{P})} , \tag{30}$$

$$\sum_{ij} C_{ij}^{(\mathrm{H})} N_j^{(\mathrm{H})} = \pi_i^{(\mathrm{H})} , \tag{31}$$

again, we only show plant and herbivore species since the equations for insects and carnivors can be obtained by permutation of indices. In matrix notation, the effective competition matrices C are given by

$$\widetilde{C}^{(\mathrm{P})} = I^{(\mathrm{P})} + \widetilde{\delta}^{(\mathrm{P})}\widetilde{\delta}^{(\mathrm{H})} - \widetilde{\gamma}^{(\mathrm{P})} \left(I^{(\mathrm{A})} + \widetilde{\delta}^{(\mathrm{A})}\widetilde{\delta}^{(\mathrm{C})} \right)^{-1} \widetilde{\gamma}^{(\mathrm{A})} \tag{32}$$

$$\widetilde{C}^{(\mathrm{H})} = I^{(\mathrm{H})} + \widetilde{\delta}^{(\mathrm{H})} \left[I^{(\mathrm{P})} - \widetilde{\gamma}^{(\mathrm{P})} \left(I^{(\mathrm{A})} + \widetilde{\delta}^{(\mathrm{A})}\widetilde{\delta}^{(\mathrm{C})} \right)^{-1} \widetilde{\gamma}^{(\mathrm{A})} \right]^{-1} \widetilde{\delta}^{(\mathrm{P})} , \tag{33}$$

where $\widetilde{C} = \beta^{-1}C$, $\widetilde{\gamma} = \beta^{-1}\gamma$ and $\widetilde{\delta} = \beta^{-1}\delta$ are competition reduced interaction matrices. Through a development similar to the case of mutualism without predation it is possible to show that the equilibrium points are stable if and only if (1) all the equilibrium biomasses are positive, and (2) the effective competition matrices are positive definite for all four groups of species.

Furthermore, the structural stability is related to the effective interspecific competition parameter (which in turn can be obtained from the maximum eigenvalue of the normalized effective competition matrix, as discussed in the paper). This relationship determines that a system with a smaller interspecific competition parameter will be more structurally stable and it will sustain stable equilibrium points for a broader range of productivity parameters π_i. This, in turn, will allow on average a larger number of coexisting species.

Through Taylor expansion, we can compute the effective competition matrix $C^{(\mathrm{P})}$ as follows:

doi: 10.1038/nature07950 SUPPLEMENTARY INFORMATION

$$\tilde{C}^{(\mathrm{P})} = I^{(\mathrm{P})} + \widetilde{\delta}^{(\mathrm{P})}\widetilde{\delta}^{(\mathrm{H})} - \widetilde{\gamma}^{(\mathrm{P})}\widetilde{\gamma}^{(\mathrm{A})} - \sum_{k=1}^{\infty}(-1)^k\widetilde{\gamma}^{(\mathrm{P})}\left(\widetilde{\delta}^{(\mathrm{A})}\right)^k\left(\widetilde{\delta}^{(\mathrm{C})}\right)^k\widetilde{\gamma}^{(\mathrm{A})}\,. \qquad (34)$$

The terms in the sum are sub-dominant, since stability of the fixed points requires that the matrices $\widetilde{\gamma}^{(\mathrm{P})}\widetilde{\gamma}^{(\mathrm{A})}$, $\widetilde{\delta}^{(\mathrm{P})}\widetilde{\delta}^{(\mathrm{H})}$ and $\widetilde{\delta}^{(\mathrm{A})}\widetilde{\delta}^{(\mathrm{C})}$ have eigenvalues smaller than one. Taking into account only the dominant terms, i.e., omitting the sum, and using the soft mean field approximation in which the elements of the mutualistic interaction matrix γ and the predatory interaction matrix δ are either zero or they are all equal, it is possible to relate structural stability with the architecture of ecological interactions. In particular, by analogy with the case with mutualism and competition, we see that the effective interspecific competition parameter is reduced if mutualistic interactions are nested and predatory interactions are antinested, therefore increasing structural stability and favouring biodiversity. The correlations between the mutualistic network and the predatory network introduces another interesting level at which we can study the architecture of the community.

6.2 Numerical results

We tested our analytic theory through numerical simulations. In particular, we wanted to test the following key aspects: whether the model ecosystems attain fixed points, how rapidly they reach equilibrium and whether some interesting dynamical behavior is observed in the transients. These issues were examined considering fully connected ecosystems. We chose the growth rates α from uniform distributions with variable width, in order to test our predictions that the width of the growth rates distribution limits the maximum possible biodiversity. The competition and mutualistic coefficients β, and γ were also chosen from a uniform distribution, in order to test the robustness of our mean field results with respect to noise in the parameters. Simulations were performed by

doi: 10.1038/nature07950 SUPPLEMENTARY INFORMATION

integrating the system of ordinary diffrential equations using a fourth-order Runge-Kutta method with small integration step.

For purely competitive systems, it can be shown analytically that the stable equilibrium points are also globally stable, in the sense that all initial conditions converge to the equilibrium point. In the presence of mutualistic and predatory interactions global stability can not be proven in general, and interesting dynamical behaviors like limit cycles or even chaos may in principle occurr. Therefore, we tested numerically the convergence to equilibrium. Our numerical results suggest that the direct competition matrix favors fast convergence to the equilibrium points even in the presence of mutualism and predation. In all cases that we simulated, including those with predatory interactions, the system attained a fixed point after a short transient, in which extinction of some species can occurr. Supplementary Fig. 1 below shows an example of the dynamics of this system. These results were robust with respect to fluctuations in the α, β, and γ parameters and confirmed the expected dependence of biodiversity on the width of the distribution of α and the expected increase of species abundance due to mutualism. The simulations not only confirmed our analytic mean field results, but also provided the new observation that the convergence to equilibrium becomes faster for a mutualistic system with respect to a purely competitive one.

Simulating ecosystems that are not fully connected requires further choices of the parameters, which we will explore systematically in future work. As a preliminary observation, we notice that in this case there must be a trade-off between the three types of parameters present in our model, i.e. the growth rates α_i, the direct competition coefficients β_{ij} and the mutualistic interactions γ_{ij}. This can be seen in the following way: The coexistence condition imposes that the main eigenvector c_i of the effective competition matrix, Eq.(5), must be almost parallel to the effective productivity vector, Eq.(4), which

doi: 10.1038/nature07950 SUPPLEMENTARY INFORMATION

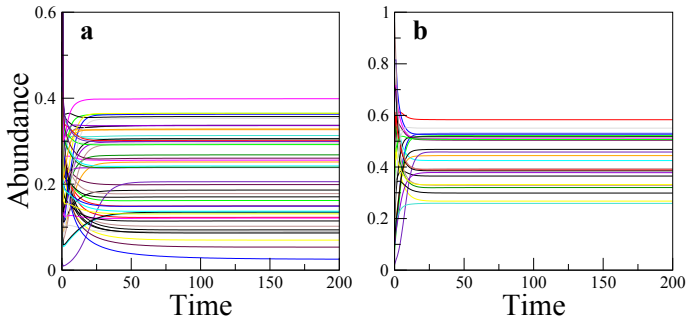


Figure 1: *Temporal dynamics of the abundance of plants (a) and animals (b) in a fully connected mutualistic community with 50 plant species and 25 animal species. The same system without mutualism leads to the extinction of one animal and 17 plant species. Parameter values are as follows: α_i are taken from a uniform distribution (0.85, 1.1); β_{ii} and β_{ij} are taken from a uniform distribution (0.99, 1.01) and (0.22, 0.24), respectively; γ_{ij} are taken from a uniform distribution (0.19, 0.21). $h^{(\mathrm{P})} = h^{(\mathrm{A})} = 0.1$. Initial population densities are taken from a uniform distribution (0, 1). This parameter combination corresponds to the weak mutualism regime. Qualitatively similar results are obtained for the strong mutualism regime, in which transients are even shorter and abundances higher.*

means that species that effectively compete more should be able to effectively grow faster in the absence of competition in order to survive. We verified through simulations (not shown here) that mutualistic interactions favour biodiversity when such correlations are implemented in the model, even if the number of mutualistic links is broadly distributed, provided that mutualism is weak enough to remain in the weak mutualistic regime.

We can think of these correlations either as the product of some physiological trade-off or as the product of an evolutionary process in which the ecosystem is slowly assembled. They may be achieved in real ecosystems in various ways, for instance through a trade-off between the number and the strength of mutualistic interactions, which decrease the effective competition and at the same time increase the effective productivity. In other words, specialist species must interact more strongly than the generalists. Another way

doi: 10.1038/nature07950 SUPPLEMENTARY INFORMATION

to obtain suitable correlations is a compensation between the growth rates and the mutualistic coefficients, so that species that interact mutualistically with more species have a smaller growth rate

In order to study the effect of mutualism systematically in networks with broad degree distributions, we will adopt in further work the following procedure: (1) Consider direct competition of mean field type, which is more demanding for allowing coexistence; (2) Extract the mutualistic strengths in such a way that there is a trade-off between the number of links and their strength; (3) Compute the effective competition matrix and its main eigenvector c_i; (4) The optimal distribution (i.e. the one that best promotes coexistence) of effective productivities P_i is predicted to be proportional to the eigenvector c_i. (5) From this optimal P_i, we can compute analytically the bare growth rates α_i that best favour coexistence. Our analytic prediction is that weak mutualism makes the system more structurally stable, in the sense that it allows more noise on the parameters α_i with respect to their optimal value. We will test this prediction in future extensive numerical work.

doi: 10.1038/nature07950 SUPPLEMENTARY INFORMATION

Appendix A: proof of the stability of the weak mutualism fixed point

We prove here that the fixed point in the weak mutualist regime is stable if and only if the effective competition matrices $C^{(\mathrm{P})}$ and $C^{(\mathrm{A})}$ are positive definite and all densities are positive.

If there is a fixed point of Eq. (29), at order zero in h, with all positive densities, it will be stable if and only if its Jacobian matrix J is negative definite, i.e., for any vector z one must have $(z, Jz) < 0$, where the brackets denote scalar product, and (z, Jz) is a generic quadratic form of the matrix J

$$J = \begin{pmatrix} -\beta^{(\mathrm{P})} & \gamma^{(\mathrm{P})} \\ \gamma^{(\mathrm{A})} & -\beta^{(\mathrm{A})} \end{pmatrix}. \tag{35}$$

We use here the matrix notation for the complete community, where the diagonal elements $\beta^{(\mathrm{P})}$ and $\beta^{(\mathrm{A})}$ are matrices acting on plant indices and animal indices, respectively, the upper right element $\gamma^{(\mathrm{P})}$ is a matrix going from plant to animal indices, and the lower left element $\gamma^{(\mathrm{A})}$ is a matrix going from animal to plant indices. With this notation, the notation C will denote the matrix formed with the two matrices $C^{(\mathrm{P})}$ and $C^{(\mathrm{A})}$ as diagonal elements.

We now show that positivity of $C^{(\mathrm{P})}$ and $C^{(\mathrm{A})}$ is necessary for the stability of the fixed point. More precisely, we will show that if J is negative definite then C must be positive definite. In order to do this, let us introduce two column vectors $x = (x^{(\mathrm{P})}, x^{(\mathrm{A})})$ and $y = (y^{(\mathrm{P})}, y^{(\mathrm{A})})$ (here $x^{(\mathrm{P})}$ ($x^{(\mathrm{A})}$) denotes the projection of the vector x on the P (A) subspace respectively), such that $x^{(\mathrm{A})} = \gamma^{(\mathrm{A})} x^{(\mathrm{P})}$, and $y^{(\mathrm{P})} = \gamma^{(\mathrm{P})} y^{(\mathrm{A})}$. These vectors have the property that $(Jx)^{(\mathrm{P})} = -C^{(\mathrm{P})} x^{(\mathrm{P})}$, $(Jx)^{(\mathrm{A})} = 0$, $Jy^{(\mathrm{A})} = -C^{(\mathrm{A})} y^{(\mathrm{A})}$, $(Jy)^{(\mathrm{P})} = 0$. Using these properties, we can see that for a generic vector $z = x + y$, it holds

$$-(z, Jz) = \left(x^{(\mathrm{P})}, C^{(\mathrm{P})} x^{(\mathrm{P})}\right) + \left(y^{(\mathrm{A})}, C^{(\mathrm{A})} y^{(\mathrm{A})}\right) + \left(x^{(\mathrm{A})}, C^{(\mathrm{A})} y^{(\mathrm{A})}\right) + \left(y^{(\mathrm{P})}, C^{(\mathrm{P})} x^{(\mathrm{P})}\right) . \tag{36}$$

doi: 10.1038/nature07950 SUPPLEMENTARY INFORMATION

From this it is immediate to see that if J is negative definite C must be positive. In fact, if $C^{(\mathrm{P})}$ is not positive, there is a vector such that $\left(x^{(\mathrm{P})}, C^{(\mathrm{P})}x^{(\mathrm{P})}\right) \leq 0$. Choosing $y^{(\mathrm{A})} = 0$, we find $(z, Jz) = -\left(x^{(\mathrm{P})}, C^{(\mathrm{P})}x^{(\mathrm{P})}\right) \geq 0$, contrary to the assumption.

We now show that the positivity of C is also a sufficient condition for stability, i.e., if C is positive then J is negative. For this proof it is convenient to rewrite the Jacobian matrix in terms of the competition reduced Jacobian $\widetilde{J}$,

$$J = \begin{pmatrix} \beta^{(\mathrm{P})} & 0 \\ 0 & \beta^{(\mathrm{A})} \end{pmatrix} \begin{pmatrix} -I^{(\mathrm{P})} & \widetilde{\gamma}^{(\mathrm{P})} \\ \widetilde{\gamma}^{(\mathrm{A})} & -I^{(\mathrm{A})} \end{pmatrix} \equiv \beta\, \widetilde{J}. \tag{37}$$

where $I^{(\mathrm{P,A})}$ denotes the identity matrix in the plant (animal) space, respectively, and $\widetilde{\gamma}^{(\mathrm{P})} = (\beta^{(\mathrm{P})})^{-1}\gamma^{(\mathrm{P})}$, $\widetilde{\gamma}^{(\mathrm{A})} = (\beta^{(\mathrm{A})})^{-1}\gamma^{(\mathrm{A})}$. We similarly define the matrix $\widetilde{C}$, satisfying

$$C = \begin{pmatrix} \beta^{(\mathrm{P})} & 0 \\ 0 & \beta^{(\mathrm{A})} \end{pmatrix} \begin{pmatrix} \widetilde{C}^{(\mathrm{P})} & 0 \\ 0 & \widetilde{C}^{(\mathrm{A})} \end{pmatrix} \equiv \beta\, \widetilde{C}\,, \tag{38}$$

where $\widetilde{C}^{(\mathrm{A})} = I^{(\mathrm{A})} - M^{(\mathrm{A})}$, $\widetilde{C}^{(\mathrm{P})} = I^{(\mathrm{P})} - M^{(\mathrm{P})}$, and the mutualistic matrix M can be written as

$$M = \begin{pmatrix} \widetilde{\gamma}^{(\mathrm{P})}\widetilde{\gamma}^{(\mathrm{A})} & 0 \\ 0 & \widetilde{\gamma}^{(\mathrm{A})}\widetilde{\gamma}^{(\mathrm{P})} \end{pmatrix}. \tag{39}$$

Since β is positive, J (C) will be positive definite if and only if $\widetilde{J}$ ($\widetilde{C}$) is positive definite. We now proceed to show that the statements

1. The effective competition matrix $\widetilde{C}$ is positive definite.

2. The reduced Jacobian $\widetilde{J}$ is negative definite.

are equivalent. It is crucial to note that M can be written as the square of a matrix, $M = \sqrt{M}\sqrt{M}$, with

$$\sqrt{M} \equiv \begin{pmatrix} 0 & \widetilde{\gamma}^{(\mathrm{P})} \\ \widetilde{\gamma}^{(\mathrm{A})} & 0 \end{pmatrix}. \tag{40}$$

doi: 10.1038/nature07950 SUPPLEMENTARY INFORMATION

Therefore $\sqrt{M}$ and M will have the same basis of eigenvectors and their eigenvalues will be related by $\lambda(M) = \left[\lambda\left(\sqrt{M}\right)\right]^2$. Furthermore note that the competition reduced matrices can be written as

$$\widetilde{J} = \quad -I \quad +\sqrt{M}\,, \tag{41}$$

$$\widetilde{C} = \quad I \quad -M\,. \tag{42}$$

Thus $\widetilde{J}$, $\widetilde{C}$, M and $\sqrt{M}$ can be diagonalized in the same basis, and their eigenvalues satisfy

$$\begin{aligned} \lambda(\widetilde{J}) = &\quad -1 \quad +\lambda(\sqrt{M}) \\ \lambda(\widetilde{C}) = &\quad 1 \quad -\lambda(M)\,. \end{aligned} \tag{43}$$

If $\lambda(\sqrt{M})$ is imaginary then it must be pure imaginary since the $\lambda(\widetilde{C})$ are real and positive by hypothesis. Therefore, in this case it must follow that $\mathrm{Re}(\lambda(\widetilde{J})) < 0$, so that the fixed point is stable. On the other hand, if $\lambda(\sqrt{M})$ is real, it follows that $\lambda(M)$ can not be negative. If we now assume that $\widetilde{C}$ is positive definite, it follows that $0 \leq \lambda(M) < 1$, which implies $-1 < \lambda\left(\sqrt{M}\right) < 1$ and consequently $\lambda(\widetilde{J}) < 0$, i.e., the fixed point is stable.

In this way, we have demonstrated that positivity of the effective competition matrix and the equilibrium biomasses are necesary and sufficient conditions for stability of the weak mutualism fixed points.

As a corollary, we see that if C is positive definite then $\lambda(M) < 1$, a result that was used in a previous section.

doi: 10.1038/nature07950 SUPPLEMENTARY INFORMATION

Appendix B: numerical calculation

For a given mutualistic interaction network $\{g_{ik}\}$ and given parameters R, $\overline{S}^{(\mathrm{P})}$ and $\overline{S}^{(\mathrm{A})}$, we can compute numerically the effective competition coefficient $\rho^{(\mathrm{P})}_{\mathrm{mut}}$ through Eq.(21), Eq.(22) and Eq.(14), where $\lambda_1(B^{(\mathrm{P})})$ is the main eigenvalue of the normalized competition matrix $B^{(\mathrm{P})}$. The corresponding maximum predicted biodiversity is given by $\overline{S}^{(\mathrm{P})}_{\mathrm{mut}} = (1-\rho^{(\mathrm{P})}_{\mathrm{mut}})/\rho^{(\mathrm{P})}_{\mathrm{mut}}$, and it characterizes the biodiversity of the model ecosystem. For $R=0$ (pure competition) it holds that $\rho^{(\mathrm{P})}_{\mathrm{mut}} = \rho^{(\mathrm{P})}$ and $\overline{S}^{(\mathrm{P})}_{\mathrm{mut}} = \overline{S}^{(\mathrm{P})}$,which can be interpreted as the maximum biodiversity of the system in the absence of mutualism

For this computation, we use the same parameters R, $\overline{S}^{(\mathrm{A})}$, $\overline{S}^{(\mathrm{P})}$ for all systems. The parameters $\overline{S}^{(\mathrm{A})}$, $\overline{S}^{(\mathrm{P})}$ are chosen large enough so that mutualism favors biodiversity in all real networks, i.e., $\rho^{(\mathrm{P})}_{\mathrm{mut}}$ decreases with the mutualism-to-competition ratio R when this is close to zero. The parameter R should be small enough so that all real networks are in the weak mutualism regime (the matrix $B^{(\mathrm{P})}$ is positive for all real networks). To eliminate the dependence on this parameter, we compute numerically the derivative with respect to R of the relative increase in biodiversity due to mutualism, using a very small value of R:

$$r^{(\mathrm{P})} \equiv \frac{1}{\overline{S}^{(\mathrm{P})}} \left.\frac{\partial \overline{S}^{(\mathrm{P})}_{\mathrm{mut}}}{\partial R}\right|_{R=0} \approx \frac{\overline{S}^{(\mathrm{P})}_{\mathrm{mut}}\left(R, \overline{S}^{(\mathrm{P})}, \overline{S}^{(\mathrm{A})}, \{g_{ik}\}\right) - \overline{S}^{(\mathrm{P})}}{R\overline{S}^{(\mathrm{P})}}. \qquad (44)$$

We verified that this computation agrees within the numerical precision with the analytical calculation reported in Eq.(24).

For each model constructed from a real mutualist network $\{g_{ik}\}$ we considered an ensemble of random networks with the same number of species and the same number of mutualistic links and different overlap matrix n_{ij}, so that their nestedness is different, and we computed $r^{(\mathrm{P})}$ both for the real network and for the ensemble of random networks.

doi: 10.1038/nature07950 SUPPLEMENTARY INFORMATION

Appendix C: Supplementary Table 1. Real networks

Pollination networks

Network	Plants	Animals	Links	Connectivity	Latitude
Arroyo et al 1982	84	101	361	0.0426	Temperate
Arroyo et al 1982	43	64	196	0.0712	Temperate
Arroyo et al 1982	36	25	81	0.09	Temperate
Elberling & Olesen 1999	24	118	242	0.0855	Arctic
Elberling & Olesen 1999	31	76	456	0.1935	Arctic
Hocking 1968	29	81	179	0.0762	Arctic
Kakutani et al l990	113	315	772	0.0217	Temperate
Kato & Miura 1996	64	187	430	0.0359	Temperate
Kato et al 1990	91	679	1193	0.0193	Temperate
Kato et al 1993	90	356	865	0.027	Temperate
Kevan 1970	20	91	190	0.1044	Arctic
McMullen 1993	10	22	27	0.1227	Tropical
Mosquin & Martin 1967	11	18	38	0.1919	Arctic
Percival 1974	61	36	178	0.0811	Tropical
Primack 1983	49	118	346	0.0598	Temperate
Primack 1983	41	139	374	0.0656	Temperate
Primack 1983	18	60	120	0.1111	Temperate
Petanidou 1991	131	666	2931	0.0336	Mediterranean
Ramirez 1989	47	46	151	0.0698	Tropical
Schemske et al 1978	7	33	65	0.2814	Temperate
Herrera 1988	26	179	412	0.0885	Mediterranean
Olesen unp.	10	12	30	0.25	Tropical
Olesen unp.	10	40	72	0.18	Temperate
Olesen unp.	8	42	79	0.2351	Temperate
Olesen unp.	29	55	145	0.0909	Tropical
Olesen unp.	26	82	248	0.1163	Temperate
Inoue et al 1990	112	840	1872	0.0199	Temperate
Inoue & Pyke 1988	36	81	252	0.0864	Temperate
Eskildsen et al unp.	14	13	52	0.2857	Tropical

doi: 10.1038/nature07950 SUPPLEMENTARY INFORMATION

Seed dispersal networks

Network	Plants	Animals	Links	Connectivity	Latitude
Baird 1980	21	7	50	0.3401	Temperate
Beeheler 1983	31	9	119	0.4265	Tropical
Jordano unp.	25	33	154	0.1867	Mediterranean
Crome 1975	71	7	142	0.2857	Tropical
Frost 1980	16	10	110	0.6875	Subtropical
Guitian 1983	12	7	40	0.4762	Temperate
Jordano unp.	16	17	121	0.4449	Mediterranean
Kantak 1979	5	27	86	0.637	Tropical
Lambert 1989	25	61	511	0.3351	Tropical
Wheelwright et al. 1984	169	40	666	0.0985	Tropical
Jordano unp.	18	28	129	0.256	Mediterranean
Tutin et al 1997	19	8	75	0.4934	Tropical
Noma 1997	15	8	38	0.31367	Temperate
Sorensen 1981	7	6	22	0.5238	Temperate
Galetti & Pizo 1996	7	18	38	0.3016	Tropical
Galetti & Pizo 1996	35	29	146	0.1438	Tropical
Snow & Snow 1971	50	14	234	0.3343	Tropical
Herrera 1984	14	10	65	0.4643	Mediterranean
Silva et al 2002; unp.	207	110	1120	0.0492	Tropical
Snow & Snow 1988	11	14	47	0.3052	Temperate
Jordano unp.	3	3	6	0.66667	Mediterranean
Jordano unp.	12	4	31	0.64583	Mediterranean
Jordano unp.	8	5	26	0.65	Mediterranean
Jordano unp.	21	6	58	0.4603	Mediterranean
Jordano unp.	11	6	36	0.5455	Mediterranean
Jordano unp.	4	5	10	0.5	Mediterranean
Jordano unp.	5	4	11	0.55	Mediterranean

For the list of references of Table S1 see ref. 9.

doi: 10.1038/nature07950 SUPPLEMENTARY INFORMATION

Supplementary References

1. Holling, C.S. Some characteristics of simple types of predation and parasitism. *Can. Ent.* **91**, 385-398 (1959).

2. Lässig, M., Bastolla, U., Manrubia, S.C. & Valleriani, A. The shape of ecological networks. *Phys. Rev. Lett.* **86**, 4418-4421 (2001).

3. Bascompte, J., Jordano, P. & Olesen, J.M. Asymmetric mutualistic networks facilitate biodiversity maintenance. *Science* **312**, 431-433 (2006).

4. Bastolla, U., Lässig, M., Manrubia, S.C. & Valleriani, A. Biodiversity in model ecosystems, I: Coexistence conditions for competing species. *J. Theor. Biol.* **235**, 521-530 (2005).

5. Chesson, P. Multispecies competition in variable environments. *Theor. Pop. Biol.* **45**, 227-276 (1994).

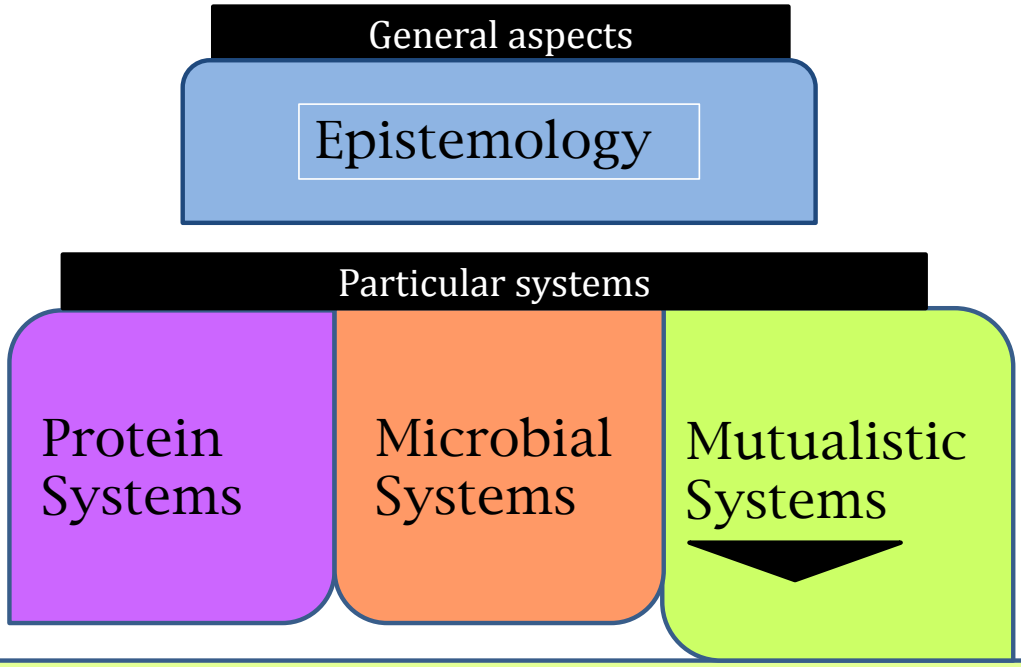
6. Chesson, P. Mechanisms of maintenance of species diversity. *Annu. Rev. Ecol. Syst.* **31**, 343-366 (2000).

7. Medan, D., Perrazzo, P.J., Devoto, M., Burgos, E., Zimmermann, M.G., Ceva, H. & Delbue, A.M. Analysis and assembling of network structure in mutualistic systems. *J. Theor. Biol.* **246**: 510-521 (2007).

8. Bulleri, F., Bruno, J.F. & Benedetti-Cecchi, L. Beyond competition: incorporating positive interactions between species to predict ecosystem invasibility. *PLoS Biology* **6**, e162 (2008).

9. Bascompte, J., Jordano, P., Melián, C.J. & Olesen, J.M. The nested assembly of plant-animal mutualistic networks. *Proc. Natl. Acad. Sci. USA* **100,** 9383-9387 (2003).

A.2. Supplementary Material Article [MUT-2]



1) Bastolla U, Fortuna MA, Pascual-García A, Ferrera A, Luque B, Bascompte J. (2009) The architecture of mutualistic networks minimizes competition and increases biodiversity. *Nature.* 458(7241):1018-20

2) Pascual-García A ., (2010) Explorando el rol de la Competición, el Mutualismo y la Arquitectura en Redes Ecológicas: ¿Qué podemos decir sobre la Biodiversidad?
Published in: *Evolución y Adaptación: 150 años después del origen de las especies.* Editors: Hernán Dopazo and Arcadi Navarro.
ISBN 978-84-92910-06-9

3) Pascual-García A., Ferrera A., and Bastolla, U. (2014) Does mutualism hinder biodiversity? arXiv preprint arXiv:1409.1683

4) Pascual-García A ., Ferrera A., and Bastolla, U. (2015) Effective competition determines the structural stability of model ecosystems. *Under revision.*

5) Ferrera A., Pascual-García A ., and Bastolla, U. (2015) Effective competition determines the global stability of model ecosystems. *Under revision.*

6) Pascual-García A., Bastolla U., (2015) The complexity-stability relation of mutualistic systems reconciles MacArthur and May. ***Under revision.***

Extended Data for the paper: The complexity–stability relation of mutualistic systems reconciles MacArthur and May

Alberto Pascual-García[(1)] and Ugo Bastolla[(1,2)]

[(1)] Centro de Biologia Molecular "Severo Ochoa"
CSIC-UAM Cantoblanco, 28049 Madrid, Spain

[(2)] E-mail: ubastolla@cbm.csic.es

Contents

10. **Interspecific competition and network properties**

11. **Relation between structural and dynamical stability**

Supplementary figures

1. Nestedness versus connectance for the pool of simulated networks.

2. Dynamical stability: Maximum eigenvalue of the Jacobian matrix of the dynamical equations at the equilibrium point (green line) and minimum eigenvalues of the effective competition matrices for plants (black) and animals (red) as a function of the mutualistic strength γ_0.

3. Interspecific effective competition parameter ρ^{eff} versus nestedness for various values of γ_0 and ρ.

4. Critical interspecific competition of plants versus nestedness in the weak mutualistic regime for facultative mutualism.

5. Structural stability, effective interspecific competition parameter and propagation of perturbations versus network architecture (connectance and nestedness) in various regimes.

6. Lower critical mutualistic threshold $\gamma_0^{(1)}$ versus the effective competition parameter.

7. Critical mutualistic threshold $\gamma_0^{(1)}$ and $\gamma_0^{(2)}$ versus nestedness and connectance, respectively.

8. Structural stability Δ_c versus dynamical stability $\gamma_0^{(1)}$ for facultative weak mutualism with weak competition $\rho = 0.05$ (left) and strong competition $\rho = 0.23$ (right). Each point represents a network with different connectance and nestedness.

1 Model

1.1 Dynamical system

As in Refs. [1–3], the multi-species population dynamics is governed by the equations

$$\frac{1}{N_i^{(\mathrm{P})}}\frac{\mathrm{d}N_i^{(\mathrm{P})}}{\mathrm{d}t} = \alpha_i^{(\mathrm{P})} - \sum_{j\in\mathbf{P}} \beta_{ij}^{(\mathrm{P})} N_j^{(\mathrm{P})} + \sum_{k\in\mathbf{A}} \frac{\gamma_{ik}^{(\mathrm{P})} N_k^{(\mathrm{A})}}{1 + h_i^{(\mathrm{P})} \sum_{l\in\mathbf{A}} \gamma_{il}^{(\mathrm{P})} N_l^{(\mathrm{A})}} \,. \tag{1}$$

where $N_i^{(\mathrm{P})}$ denotes the abundance of plant species i, $\alpha_i^{(\mathrm{P})}$ is its bare growth rate in the absence of other species, $\beta_{ij}^{(\mathrm{P})}$ is the direct competition matrix and $\gamma_{ik}^{(\mathrm{P})}$ is the mutualistic matrix, which vanishes if the link a_{ik} is absent. The equations for animals are obtained interchanging the superscripts P and A and they will not be presented.

We adopt the so-called soft mean field model, which assumes that all equivalent interaction parameters are uniformly distributed. This is the simplest assumption that one can make without additional hypothesis. The competition parameters are uniformly distributed within two classes, intraspecific and interspecific. We consider two scales of biomass, $\hat{N}^{(\mathrm{P})}$ for plants populations and $\hat{N}^{(\mathrm{A})}$ for animal populations. The biomass scale of competition is determined by the biomass scale of the corresponding guild, and the intraspecific competition is chosen as $1/\hat{N}$, i.e. the time scale is set by the intraspecific competition:

$$\beta_{ij}^{(\mathrm{P})} = \frac{b_{ij}}{\hat{N}^{(\mathrm{P})}} \left(\rho^{(\mathrm{P})} + \delta_{ij}(1-\rho^{(\mathrm{P})})\right) \,, \tag{2}$$

where $0 \leq \rho^{(\mathrm{P})} \leq 1$ is the interspecific competition parameter, δ_{ij} is Kronecker's delta and b_{ij} are dimensionless numbers uniformly distributed in $[1-\delta_b, 1+\delta_b]$.

We then parameterize the mutualistic interactions as

$$\gamma_{ik}^{(\mathrm{P})} = a_{ik} \frac{\gamma_0 c_{ik}^{(\mathrm{P})}}{\sqrt{\hat{N}^{(\mathrm{P})}\hat{N}^{(\mathrm{A})}}} \tag{3}$$

where a_{ik} is the binary adjacency matrix of the mutualistic network, the dimensionless parameters $c_{ik}^{(\mathrm{P})}$ are uniformly distributed between $1-\delta_c$ and $1+\delta_c$ if $a_{ik}=1$ and are zero if $a_{ik}=0$. The handling times are set to $h_i = H^{(\mathrm{A})}$ for animals and $h_i = H^{(\mathrm{P})}$ for plants.

1.2 Feasibility condition

For each realization of the ecological interactions β_{ij} and γ_{ik}, the initial growth rates $\overline{\alpha_i^{(\mathrm{A})}}$ are chosen in such a way that the equilibrium abundances are equal to $\overline{N}_i > 0$, which automatically guarantees the feasibility of the equilibrium:

$$\overline{\alpha_i^{(\mathrm{P})}} = \sum_{j\in\mathbf{P}} \beta_{ij}^{(\mathrm{P})} \overline{N}_j^{(\mathrm{P})} - \sum_{k\in\mathbf{A}} \frac{\gamma_{ik}^{(\mathrm{P})} \overline{N}_k^{(\mathrm{A})}}{1 + h_i^{(\mathrm{P})} \sum_{l\in\mathbf{A}} \gamma_{il}^{(\mathrm{P})} \overline{N}_l^{(\mathrm{A})}} \,. \tag{4}$$

1.3 Obligatory and facultative mutualism

We define obligatory mutualism when the growth rates are always negative for animals and positive for plants, and facultative mutualism when all growth rates are positive. From Eq.(4), this implies that the inverse of the handling time of animals, which is the maximum mutualistic growth rate, must be larger than the abundance loss at equilibrium due to competition,

$$\frac{1}{h_i^{(\mathrm{A})}} > (S^{(\mathrm{A})} - 1)\rho^{(\mathrm{A})} + 1 \,. \tag{5}$$

However, h_i must be limited, otherwise the equilibrium would be dynamically unstable [4], and it produces an important trade-off between the number and the strength of mutualistic interactions, since when the number of mutualistic interaction is large they saturate and their effective strength is reduced. To fulfill these conditions, for obligatory mutualism we set $H^{(\mathrm{A})} = 0.75/\left(S^{(\mathrm{A})}\rho^{(\mathrm{A})} + 1 - \rho^{(\mathrm{A})}\right)$ for animals and $H^{(\mathrm{P})} = 0.25$ for plants, for which there is not such a constraint.

Assuming that all equilibrium abundances are equal and using Eq.(3), the conditions that growth rates are positive for plants and negative for animals translate into the in-

equalities

$$\gamma_0\sqrt{\frac{\hat{N}^{(\mathrm{A})}}{\hat{N}^{(\mathrm{P})}}}\sum_k c_{ik}^{(\mathrm{P})} < \frac{S^{(\mathrm{P})}\rho^{(\mathrm{P})}+1-\rho^{(\mathrm{P})}}{1-h_i^{(\mathrm{P})}\left(S^{(\mathrm{P})}\rho^{(\mathrm{P})}+1-\rho^{(\mathrm{P})}\right)} \quad \forall i$$

$$\gamma_0\sqrt{\frac{\hat{N}^{(\mathrm{P})}}{\hat{N}^{(\mathrm{A})}}}\sum_k c_{ik}^{(\mathrm{A})} > \frac{S^{(\mathrm{A})}\rho^{(\mathrm{A})}+1-\rho^{(\mathrm{A})}}{1-h_i^{(\mathrm{A})}\left(S^{(\mathrm{A})}\rho^{(\mathrm{A})}+1-\rho^{(\mathrm{A})}\right)} \quad \forall i$$

that require that the ratio of the abundances between plant and animal populations $\hat{N}^{(\mathrm{P})}/\hat{N}^{(\mathrm{A})}$ must be large,

$$\sqrt{\frac{\hat{N}^{(\mathrm{P})}}{\hat{N}^{(\mathrm{A})}}} > \max\left(\gamma_0\,\frac{\max\left(d_i^{(\mathrm{P})}(1-\tilde{h}_i^{(\mathrm{P})})\right)}{\left(S^{(\mathrm{P})}\rho^{(\mathrm{P})}+1-\rho^{(\mathrm{P})}\right)}, \sqrt{\frac{\hat{N}^{(\mathrm{P})}}{\hat{N}^{(\mathrm{A})}}} > \frac{1}{\gamma_0}\frac{\left(S^{(\mathrm{A})}\rho^{(\mathrm{A})}+1-\rho^{(\mathrm{A})}\right)}{\min\left(d_i^{(\mathrm{A})}(1-\tilde{h}_i^{(\mathrm{A})})\right)}\right). \quad (6)$$

where $d_i^{(\mathrm{A})}=\sum_k c_{ik}^{(\mathrm{A})}$ is the weighted degree of animal i and equivalent for plants, and $\tilde{h}_i^{(\mathrm{A})} = h_i/\left(S^{(\mathrm{A})}\rho^{(\mathrm{A})}+1-\rho^{(\mathrm{A})}\right) < 1$. The most stringent condition is the second one, imposed by animal growth rates. If the smallest mutualistic degree is one, the number of animal species is $S^{(\mathrm{A})}=50$, the competition is strong ($\rho^{(\mathrm{A})}=0.25$) and the system is in the weak mutualist regime ($\gamma_0=0.1$) we obtain $\hat{N}^{(\mathrm{P})}/\hat{N}^{(\mathrm{A})} > 2.5\cdot 10^5$. This value is consistent with empirical estimates [5]. Nevertheless, to achieve obligatory mutualism in a large range of parameters, we use $\hat{N}^{(\mathrm{P})}/\hat{N}^{(\mathrm{A})} = 7\cdot 10^7$.

1.4 Metaparameters

With these assumptions, the dynamical equations depend on the mutualistic network, in particular its number of species $S^{(\mathrm{A})}$ and $S^{(\mathrm{P})}$ and the adjacency matrix, which we characterize in terms of its connectivity and its overlap (related to the nestedness measure), on six meta-parameters that determine the average values of the interaction parameters: $\rho^{(\mathrm{A})}$, $\rho^{(\mathrm{P})}$, γ_0, $H^{(\mathrm{A})}$, $H^{(\mathrm{P})}$, and the ratio $\hat{N}^{(\mathrm{A})}/\hat{N}^{(\mathrm{P})}$, on two parameters that determine the equilibrium abundances $\overline{N}^{(\mathrm{P})}$ and $\overline{N}^{(\mathrm{A})}$ and, through Eq.(4), the ideal growth rates, and

Parameter	A	B	C	D	E	F	G	H
$S^{(\mathrm{A})}$, $S^{(\mathrm{P})}$	46, 47							
α	Facultative				Obligatory			
$\rho^{(\mathrm{A})}$, $\rho^{(\mathrm{P})}$	0.05		0.23		0.05		0.23	
γ_0	0.15				0.05	100	0.15	10^3
$\overline{N}$	1	100	1	1000	1			
$H^{(\mathrm{A})}$	0.1				0.23		0.066	
$H^{(\mathrm{P})}$	0.1				0.25			
$\hat{N}^{(\mathrm{P})}/\hat{N}^{(\mathrm{A})}$	1				$7 \cdot 10^7$			
δ_b	0.15							
δ_c	0.15							
δ_N	0.15							

Table 1: Metaparameters regimes presented in the figures.

on the 2 parameters δ_b and δ_c that control the broadness of the distributions of interaction parameters, which we assume to be the same for plants and animals.

Since a systematic exploration of all combinations of meta-parameters is unaffordable, we studied several distinct regimes whose properties are summarized in Table 1. To reduce the number of combinations, we use the same metaparameters for plants and animals except for the H and $\hat{N}$ parameters in obligatory mutualism.

1.5 Construction of mutualistic networks

For each set of meta-parameters we consider 125 mutualistic networks with different combinations of connectance and nestedness. The properties of the simulated networks are depicted in Fig.1, where one can see that almost all networks have nestedness larger than connectance.

The definitions of these structural properties are reported here for completeness. Denoting by a_{ik} the adjacency matrix for plants, whose transpose is the adjacency matrix for animals, and by $d_i^{(\mathrm{P})} = \sum_k a_{ik}$ the degree (number of mutualistic interactions) of plant species i, and analogous for animal, the connectance κ of the network is the number of

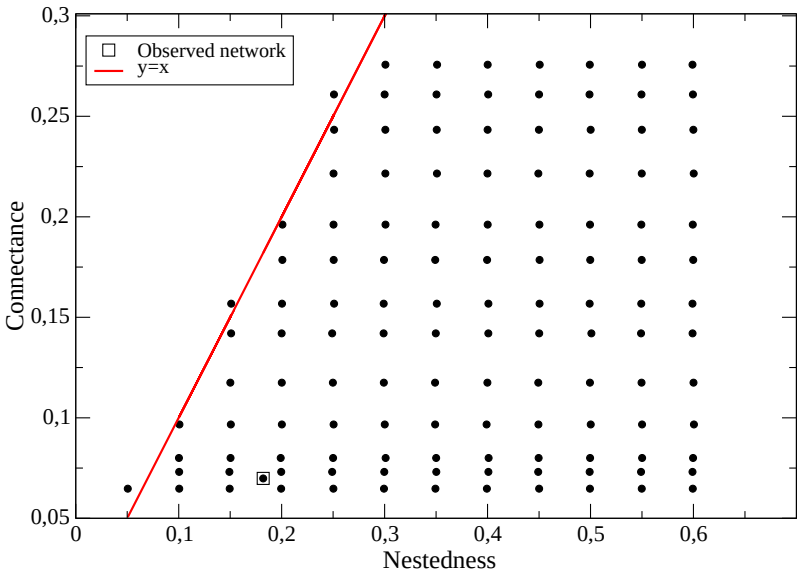


Figure 1: Nestedness versus connectance for the pool of simulated networks. The number of species is $S^{(\mathrm{P})} = 46$, $S^{(\mathrm{A})} = 47$. The observed network from which the simulated networks are derived has connectance 0.073 and nestedness 0.149.

links L divided by all possible links,

$$\kappa = \frac{L}{S^{(\mathrm{A})}S^{(\mathrm{P})}} = \frac{1}{S^{(\mathrm{A})}}\sum_i \frac{d_i^{(\mathrm{A})}}{S^{(\mathrm{P})}} = \frac{1}{S^{(\mathrm{P})}}\sum_k \frac{d_k^{(\mathrm{P})}}{S^{(\mathrm{A})}}\,. \tag{7}$$

We adopt the definition of nestedness ν proposed in [1], namely the average number of shared links between species of the same group normalized such that the maximum possible nestedness is one. It is easy to see that the nestedness is correlated with degree heterogeneity:

$$\nu^{(\mathrm{P})} = \frac{\sum_{i<j}\sum_k a_{ik}a_{kj}^T}{\sum_{i<j}\min\left(d_i^{(\mathrm{P})}, d_j^{(\mathrm{P})}\right)} = \frac{\frac{1}{S^{(\mathrm{A})}}\sum_k \left(d_k^{(\mathrm{A})}/S^{(\mathrm{P})}\right)^2 - \kappa/S^{(\mathrm{P})}}{\frac{1}{S^{(\mathrm{P})}}\sum_i \frac{d_i^{(\mathrm{P})}}{S^{(\mathrm{A})}}\frac{i}{S^{(\mathrm{P})}} - \kappa/S^{(\mathrm{P})}}\,. \tag{8}$$

The simulated networks are generated as follows. We start from an observed mutualistic network with $S^{(\mathrm{A})} = 46$ animals and $S^{(\mathrm{P})} = 47$ plants present at Cainama, Venezuela [6], and obtain different connectances by extracting links at random with a modification of

the null model proposed by Bascompte et al [7], in which the probability that there is a link between animal i and plant k is modelled as

$$p_{ik} = f\frac{d_i^{\mathrm{P,obs}} + d_k^{\mathrm{A,obs}}}{S^{(\mathrm{A})} + S^{(\mathrm{P})}} \tag{9}$$

Networks extracted with $f = 1$ have an average connectance equal to the one of the observed network and a degree distbution that interpolates between the one of the observed network and the one of a random network with uncorrelated links. The average connectance can be changed by modifying the parameter f. For each value of connectance, we obtain different values of the nestedness by applying the algorithm by Medan et al. [8] that swaps links maintaining the degree. Each swapping is selected with a Metropolis criterion that enforces the target value of the nestedness. Convergence is typically achieved after $20,000$ swaps.

1.6 Numerical experiments

Our numerical experiments proceed through the following steps: First, we set the system at a feasible and stable equilibrium; Second, we randomly perturbate the intrinsic growth rates, so that feasibility is not guaranteed anymore; Third, we simulate the ecological dynamics untill a new equilibrium is reached and record extinctions if any. For each set of metaparameters and each network, we determine the critical perturbation Δ_c as the perturbaiton such that the probability that there is at least one extinction is equal to 0.5. These steps are briefly detailed below.

First of all, for each set of metaparameters and each network we determine the maximum and minimum values of γ_0 such that the equilibrium is locally stable (see below). For the weak-strong regime of obligatory mutualism the equilibrium is stable for all values of γ_0, as analytically predicted (see section 10). For each combination of metaparameters and each one of the 125 networks, we randomly generate 50 realizations of the interaction

variables b_{ij} and c_{ik} and the equilibrium abundances $\overline{N}_i$. For each realization, we determine initial intrinsic growth rates $\overline{\alpha}_i$ through Eq.(4). We then consider several values of the perturbation parameter Δ. For each value of Δ, we generate 100 random perturbations of the growth rates as $\alpha_i = \overline{\alpha}_i\,(1 + \Delta r_i)$, where r_i is a random number uniformly distributed between -1 and 1, we draw the initial condition and and integrate the ecological dynamics Eq.(1) with the Bulirsch-Stoer algorithm with adaptive step [9] until a new equilibrium is reached, considering extinct species whose abundance falls below 10^{-8} of the initial value. In this way, for each set of metaparameters and each network we compute the probability that at least one species is extinct as a function of Δ. The value of Δ_c is obtained through interpolation.

2 Types of stability

There are several types of stability of the dynamical system Eq.(1). Dynamical stability refers to perturbations in the dynamical variables N_i, and it can be can be local or global. Local stability against small perturbations of the dynamical variables around their equilibrium values is analytically studied considering the linearized dynamical system Eq.(10). Global stability refers to perturbations of whatever size of the dynamical variables. Goh has shown that, if an interaction matrix A is diagonally positive, meaning that there is a diagonal and positive matrix D such that the symmetric matrix $DA + A^T D$ is positive definite, then the Lotka-Volterra system defined by this matrix, $\frac{1}{N_i}\frac{\mathrm{d}N_i}{\mathrm{d}t} = \sum_j A_{ij}N_j + \alpha_i$ is globally stable provided that the equilibrium is feasible, $\overline{N}_i > 0 \forall i$ [12]. This implies that, for whatever perturbation of the dynamical variable, no species will get extinct provided that the parameters of the system allow feasibility. In this situation a more interesting type of stability is structural stability.

A dynamical system is said to be structurally stable if its qualitative properties do not

change when its parameters suffer a perturbation. If the interaction matrix is diagonally stable, then structural stability against perturbations of the growth rates can be defined as the maximum perturbation that maintains a feasible equilibrium. We can define in an analogous way the structural stability with respect to changes in the interaction matrix, but in this case we have also to test that the matrix remains diagonally positive.

3 Equivalent Lotka-Volterra system and local stability

Close to a dynamical equilibrium, the dynamical stability for small perturbations of the abundances is determined by the equivalent Lotka-Volterra system

$$\frac{1}{N_i^{(\mathrm{P})}}\frac{\mathrm{d}N_i^{(\mathrm{P})}}{\mathrm{d}t} = \alpha_i^{(\mathrm{eff,P})} - \sum_{j\in\mathbf{P}} \beta_{ij}^{(\mathrm{P})} N_j^{(\mathrm{P})} + \sum_{k\in\mathbf{A}} \gamma_{ik}^{(\mathrm{eff,P})} N_k^{(\mathrm{A})} \,. \tag{10}$$

The effective interaction and growth rates parameters are obtained by differentiating the full dynamical equations (1) at the equilibrium point, and they are

$$\gamma_{ik}^{(\mathrm{eff,P})} = \frac{\gamma_{ik}^{(\mathrm{P})}}{(1+z_i)^2} \tag{11}$$

$$\alpha_i^{(\mathrm{eff,P})} = \alpha_i^{(\mathrm{P})} + h_i^{(\mathrm{P})} \left(\frac{z_i}{1+z_i}\right)^2 . \tag{12}$$

$$z_i = h_i^{(\mathrm{P})} \sum_{l\in\mathbf{A}} \gamma_{il}^{(\mathrm{P})} \overline{N}_l^{(\mathrm{A})}$$

We see from this equation that for each species there are two regimes of parameters: weak mutualism, in which the equilibrium mutualistic benefit is far from saturation ($z_i \ll 1$) and strong mutualism, in which the saturation is reached ($z_i \gg 1$). The effective mutualistic strength increases with γ_0 in the weak mutualistic regime and decreases in the strong regime, reaching a maximum in between.

The local stability of the equilibrium point can be tested from the linearized system Eq.(10), which we rewrite as $\frac{1}{N_i}\frac{\mathrm{d}N_i}{\mathrm{d}t} = \sum_j A_{ij}N_j + \alpha_i$. The equilibrium is locally stable if

and only if all eigenvalues of the Jacobian matrix $J_{ik} = \overline{N}_i A_{ik}$ have negative real part. This requires that the matrices γ^{eff} are small. From Eq.(11), it is clear that this happens both for small and for large values of γ_0, i.e. the equilibrium is stable if $\gamma_0 < \gamma_0^{(1)}$ or $\gamma_0 > \gamma_0^{(2)}$, which define the lower and upper critical mutualistic strengths and the weak and strong mutualistic regimes, respectively. We determine numerically the critical strengths for each given network and given realization of the interaction parameters by diagonalizing the matrix J_{ik}, and obtain analytic insight on them using the effective competition matrix defined below.

4 Effective competition

The effective competition matrix represents the interactions between species in the same group, either P or A, both due to their direct interaction (in this case, competition) and to their interaction with species in the other group (in this case mutualism). It allows to decouple the equations for computing equilibrium abundances as

$$\overline{N}^{(\mathrm{P})} = \left(C^{(\mathrm{P})}\right)^{-1} p^{(\mathrm{P})} \qquad\qquad \overline{N}^{(\mathrm{A})} = \left(C^{(\mathrm{A})}\right)^{-1} p^{(\mathrm{A})} \tag{13}$$

where the effective competition matrices C and the effective productivity vectors p are defined as

$$C^{(\mathrm{A})} = \beta^{(\mathrm{A})} - \gamma^{(\mathrm{eff,A})} \left(\beta^{(\mathrm{P})}\right)^{-1} \gamma^{(\mathrm{eff,P})}, \qquad p^{(\mathrm{A})} = \alpha^{(\mathrm{eff,A})} + \gamma^{(\mathrm{eff,A})} \left(\beta^{(\mathrm{P})}\right)^{-1} \alpha^{(\mathrm{eff,P})} \tag{14}$$

and analogous for plants [10].

4.1 Effective interspecific competition parameter

The usefulness of the effective competition matrix stems from the fact that, while the full interaction matrix A has both positive and negative components, we expect that most elements of the matrix $C^{(\mathrm{P})}$ and $C^{(\mathrm{A})}$ are positive in the regime in which the system is

dynamically stable (for simplicity, we omit here the superscripts since the properties that we describe are common to both matrices). In particular, we assume that the components of C are non-negative and that the matrix is irreducible, i.e. the complete space is the only spaces that is invariant under the action of C, so that we can apply the Perron-Frobenius theorem [11] that states that each matrix C has a dominant eigenvalue λ_1 of order S (the dimension of the space) associated with left and right eigenvectors u^1 and v^1 whose elements are all positive, while all other eigenvectors have at least one negative or complex component.

The main eigenvalue λ_1 can be interpreted as the weighted average of the matrix C with weights given by the main eigenvectors, $\lambda_1 = \sum_{ij} C_{ij} u_i^1 u_j^1 = \sum_{ij} C_{ij} v_i^1 v_j^1$. Although in previous publications we normalized the effective competition matrix in such a way that the intraspecific competition is equal to one, i.e. $\tilde{C}_{ij} = \frac{C_{ij}}{\sqrt{C_{ii}C_{jj}}}$and $\tilde{p}_i = \frac{p_i}{\sqrt{C_{ii}}}$, which is equivalent to changing the units of abundance and productivity of each species in such a way that $\tilde{N}_i = \sqrt{C_{ii}} N_i$, here we found that the numerical results are slightly more accurate without this normalization. Possibly this happens because the equilibrium abundances are chosen to be uniform, while the rescaled abundances $\tilde{N}_i$ are not uniform. Therefore, we provide here more general equations that hold for non normalized effective connectivity. The formulas reported in [10] are recovered as a special case when $C_{ii} = 1$. The main eigenvalue $\lambda_1 = \sum_{ij} \tilde{C}_{ij} v_i^1 v_j^1$ represents the average competition in the system, weighting each species with weights v^1. We express it in the form

$$\lambda_1(C) = \left(\sum_i C_{ii}/S \right) \left((S-1)\rho + 1 \right) , \tag{15}$$

where the interspecific competition parameter ρ^{eff} represents the ratio between interspecific and intraspecific competition, and is given by

$$\rho^{\text{eff}} = \frac{1}{S-1} \left(\frac{\lambda_1(C)}{\sum_i C_{ii}/S} - 1 \right) . \tag{16}$$

Since the sum of the eigenvalues is equal to the trace, $\sum_\alpha \lambda_\alpha(C) = \sum_i C_{ii}$, $1 - \rho^{\text{eff}}$ is related with the average value of the minor eigenvalues,

$$\frac{1}{S-1}\sum_{\alpha>1}\lambda_\alpha\left(\tilde{C}\right) = \left(1-\rho^{\text{eff}}\right)\frac{S}{\sum_i C_{ii}}\,. \tag{17}$$

The effective interspecific competition parameter plays a key role in determining the dynamical stability, structural stability and abundance of the model ecosystem, as shown in previous works [10, 13, 14] and reminded below for completeness.

5 Effective competition and dynamical stability

The effective competition matrix C provides analytic insight on dynamical stability. Diagonal positivity of the matrix C, i.e. the fact that we can find a positive diagonal matrix D such that $DC + C^T D$ is positive definite, is a necessary condition for diagonal stability of the linearized dynamical system Eq.(10) [13], which is in turn a sufficient condition for its global stability [12], i.e. for stability with respect to whatever perturbation of the dynamical variables $N_i^{(\text{P})}$ and $N_i^{(\text{A})}$. Global stability of the linearized system is necessary for global stability of the complete dynamical system, and we conjecture that it is even sufficient, given that the non linear terms have a stabilizing effect. Furthermore, diagonal positivity of C plus a mild symmetry condition is sufficient for global stability of the linearized system [13]. We conjecture that, for the class of systems that we study, for which the direct competition matrix is close to symmetric, diagonal positivity of C is necessary and sufficient for global stability, which means that no species will get extinct provided that the growth rates guarantee that the equilibrium is feasible (all $\overline{N}_i$ are positive), as Eq.(4) guarantees by construction.

Diagonal positivity is not simple to test numerically, because we have to find a suitable matrix D or to rule out its existence. However, for the model that we simulated the interaction matrix is almost symmetric, since the direct competition matrices are symmetric

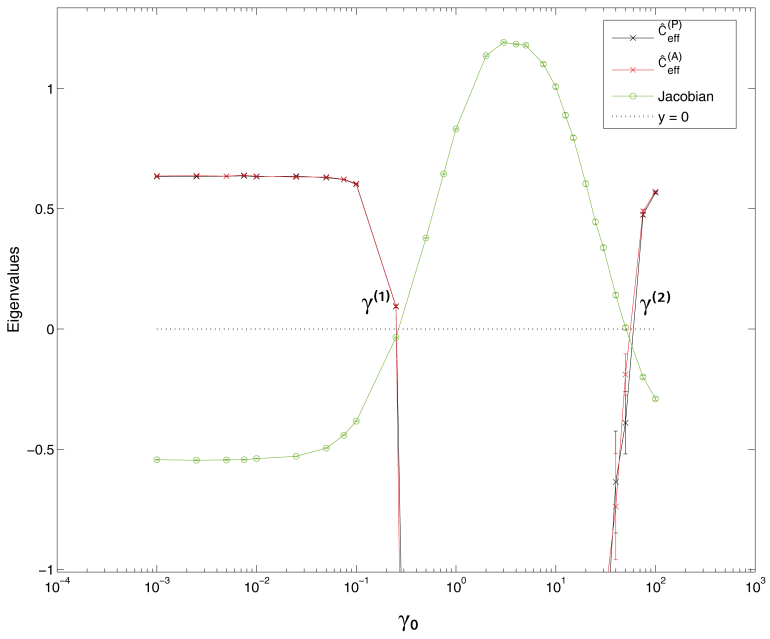


Figure 2: Maximum eigenvalue of the Jacobian matrix of the dynamical equations at the equilibrium point (green line) and minimum eigenvalues of the effective competition matrices for plants (black) and animals (red) as a function of the mutualistic strength γ_0. The system is dynamically unstable when the eigenvalue of the Jacobian is positive. Each point represents the average over 50 realizations of the interaction parameters b_{ij}, c_{ij} and e_i for one particular network with $S^{(\mathrm{P})} = 46$, $S^{(\mathrm{A})} = 47$, connectance 0.073 and nestedness 0.149, $\hat{N}^{(\mathrm{A})}/\hat{N}^{(\mathrm{P})} = 1$, $\rho^{(\mathrm{P})} = \rho^{(\mathrm{A})} = 0.23$.

and the mutualistic interaction coefficients are the same for plants and animals. In case of a symmetric interaction matrix, positivity of C and not diagonal positivity is a necessary and sufficient condition for positivity of the interaction matrix. Therefore, we conjecture that for our model positivity of C is almost necessary and sufficient for global stability. We tested numerically that the threshold value of γ_0 obtained through the condition that C is positive almost coincides with the threshold value obtained from the local stability condition that $\mathrm{Re}\left(\lambda\left(J_{ik}\right)\right) < 0$ (see Fig.2).

Because of Eq.(17), the smaller is $\rho^{\rm eff}$, the larger is the average of the minor eigenvalues of C and the less likely it is that the minimum eigenvalue λ_S is negative and the system is dynamically unstable. Therefore, dynamical stability is inversely related with $\rho^{\rm eff}$, in particular the lower mutualistic threshold $\gamma_0^{(1)}$ is inversely related with $\rho^{\rm eff}$.

6 Effective competition and structural stability

The structural stability with respect to changes in the productivities is in large extent determined by the effective interspecific competition $\rho^{\rm eff}$ [10, 14] (in this section, we omit the superscripts P and A to simplify the notation). Necessary condition for all species having positive abundance is that all productivities fulfill the inequality

$$\eta \equiv \max_i \left(\eta_i \equiv 1 - \frac{p_i}{v_i^1 p^1} \right) \leq \frac{S^{\rm eff}}{S + S^{\rm eff}} \left(1 - \frac{n_c}{\langle N \rangle} \right) , \tag{18}$$

In this section we omit superscripts to simplify the notation. Here v^1 is the main eigenvector of the effective competivity matrix C, $p^1 = \sum_i p_i v_i^1$ is the projection of the productivity vector onto v^1 (note that the weighted average of η_i with weights $(v_i^1)^2$ is one) and $S^{\rm eff}$ is a natural biodiversity scale set by the effective competition,

$$S^{\rm eff} = \frac{1 - \rho^{\rm eff}}{\rho^{\rm eff}} . \tag{19}$$

If $S^{\rm eff}/S$ is small (either large $\rho^{\rm eff}$ or large S) all of the η_i must be close to one, i.e. the productivity vector must be almost parallel to v^1, so that the feasibility condition is very demanding and even small perturbations of the productivities can violate it.

7 Effective competition and equilibrium abundances

In the limit of vanishing $S^{\rm eff}/S$, the productivity vector must be directed along v^1, $p_i = p^1 v_i^1$, which implies, through the equilibrium equation $N = (C)^{-1} p$, that the optimally

stable equilibrium abundances are also directed along v^1:

$$\overline{N}_i^{(\mathrm{P})} = \frac{v_i^1 p^1}{\rho^{\mathrm{eff}}(S^{(\mathrm{P})}-1)+1}\frac{S}{\sum_i C_{ii}}\,. \tag{20}$$

Thus for large systems with large S/S^{eff} whose productivities are strongly constrained, the equilibrium abundances are inversely related with ρ^{eff}.

8 Critical perturbation

The structural stability of the system with respect to changes in the intrinsic growth rates α_i depends on how these changes propagate into changes of productivities and, from them, changes in the η_i. This computation is complicated by the fact that a change in α_i modifies the equilibrium abundances, and consequently the saturation factors z_i and, through them, the effective growth rates and mutualistic interactions, Eq.(12) and Eq.(11). Therefore, we would need to explicitly compute the perturbed abundances and z_i, which would make the feasibility condition Eq.(18) useless.

Nevertheless, we can get analytic insight by noticing that, when mutualistic interactions are far or close to saturation ($z_i = h_i \sum_l \gamma_{il}\overline{N}_l$ either small or large), the change in $\gamma^{(\mathrm{eff,P})}$ and $\alpha^{(\mathrm{eff,P})}$ due to a change in equilibrium abundances is small and it can be neglected, except for obligatory mutualism (see below). Thus, we assume that the perturbation of growth rates does not modify $\gamma^{(\mathrm{eff,P})}$ and the effective competition matrix $C^{(\mathrm{P})}$ and that the perturbed $\alpha^{(\mathrm{eff,P})}$ is simply given by $\alpha_i^{(\mathrm{eff,P})}(\Delta) = \alpha_i^{(\mathrm{eff,P})}(\Delta=0) + \Delta\alpha_i$. Consequently, we compute the perturbed productivity vector $p^{(\mathrm{P})}(\alpha(\Delta))$ from Eq.(14), we project it onto the main eigenvector of the effective competition matrix and we compute the perturbed η as

$$\eta^{(\mathrm{P})}(\Delta) = \min_i \left(\frac{p_i^{(\mathrm{P})}(\alpha(\Delta))}{v_i^{(P),1} p^{(P),1}(\alpha(\Delta))} \right)\,. \tag{21}$$

Since the perturbation of the productivity is not correlated with the value of the productivity of the uperturbed system with $\Delta = 0$, when Δ is large enough we expect that the

minimum over i of $p_i(\Delta)$ only depends on Δ and not on the unperturbed value $p_i(\Delta = 0)$, i.e.

$$\eta^{(\mathrm{P})}(\Delta) \approx \Delta \eta'^{(\mathrm{P})} \,. \tag{22}$$

so that we can compute $\eta'^{(\mathrm{P})} \approx \eta^{(\mathrm{P})}(\Delta)/\Delta$. This equation clearly fails for $\Delta = 0$, since $\eta^{(\mathrm{P})}(\Delta = 0) \neq 0$, but it is sufficiently accurate for large Δ, and we perform the computation at the critical value of Δ expected for pure competition, $\Delta^{(\mathrm{P})} = (1 - \rho^{(\mathrm{P})})/\left(\rho^{(\mathrm{P})}(S^{(\mathrm{P})} - 1) + 1\right)$.

Nevertheless, the assumption that we can neglect the change in equilibrium abundances is not justified for obligatory mutualism. We consider the worst case of a plant species k that is the only feeding of the animal species i. Positivity of $N_i^{(\mathrm{A})}$ requires that

$$\overline{N}_i^{(\mathrm{A})} = \frac{1}{\beta_0}\left[\alpha_i^{(\mathrm{A})} + \frac{1}{h_i}\left(\frac{z_i'}{1+z_i'}\right) - \frac{\beta_0}{\hat{N}^{(\mathrm{A})}}(S^{(\mathrm{A})} - 1)\rho^{(\mathrm{A})}\left\langle \overline{N}^{(\mathrm{A})} \right\rangle\right] > 0 \tag{23}$$

where $z_i' = h_i \gamma_{ik}^{(\mathrm{A})} \overline{N}'^{(\mathrm{P})}_k$ is the saturation factor after the perturbation. We now assume that the term $\alpha_i^{(\mathrm{A})} - (\beta_0/\hat{N}^{(\mathrm{A})})(S^{(\mathrm{A})} - 1)\rho^{(\mathrm{A})}\left\langle \overline{N}^{(\mathrm{A})} \right\rangle$ does not change significantly from its value before the perturbation, which can be estimated as $(1/h_i)\,(z_i/(1+z_i)) + \left\langle N^{(\mathrm{A})} \right\rangle$, and obtain the inequality

$$-\beta_0 h_i^{(\mathrm{A})} \left\langle N^{(\mathrm{A})} \right\rangle < \frac{z_i'}{1+z_i'} - \frac{z_i}{1+z_i} \approx \frac{n_c^{(\mathrm{P})} - \left\langle N^{(\mathrm{P})} \right\rangle}{n_c^{(\mathrm{P})} \left\langle N^{(\mathrm{P})} \right\rangle}\left(\frac{1}{h_i^{(\mathrm{A})}\gamma_{ik}}\right) ,$$

where the abundance after the perturbation is the minimum plant abundance that can maintain the animal species, $n_c^{(\mathrm{P})}$, and the abundance before the perturbation is the average plant abundance $\left\langle N^{(\mathrm{P})} \right\rangle$. Thus, using $\gamma_{ik} = \gamma_0/\sqrt{\hat{N}^{(\mathrm{A})}\hat{N}^{(\mathrm{P})}}$ we can estimate the minimum plant abundance as

$$\frac{n_c^{(\mathrm{P})}}{\left\langle N^{(\mathrm{P})} \right\rangle} = \frac{1}{1 + \beta_0\gamma_0\sqrt{\frac{\hat{N}^{(\mathrm{P})}}{\hat{N}^{(\mathrm{A})}}}\,(h^{(\mathrm{A})})^2\left(\frac{\left\langle N^{(\mathrm{A})} \right\rangle}{\hat{N}^{(\mathrm{A})}}\right)\left(\frac{\left\langle N^{(\mathrm{P})} \right\rangle}{\hat{N}^{(\mathrm{P})}}\right)} = \frac{1}{1 + \gamma_0\sqrt{\frac{\hat{N}}{Pl/\hat{N}^{(\mathrm{A})}}}\,(h^{(\mathrm{A})})^2} \,, \tag{24}$$

since we choose units such that $\beta_0 = 1$ and we set the unperturbed abundances such that $\left\langle N^{(\mathrm{A})} \right\rangle / \hat{N}^{(\mathrm{A})} = \left\langle N^{(\mathrm{P})} \right\rangle / \hat{N}^{(\mathrm{P})} = 1$. Putting everything together, we obtain that

the maximum perturbation of growth rates compatible with the persistence of all species is the value of Δ that generates the minimum allowed perturbed η, Eq.(18), where the critical abundance is zero in the case of animals and it is Eq-(24) in the case of plants. Given that $\eta = \Delta\eta\prime$, we find

$$\Delta_c^{(\mathrm{A})} = \frac{1}{\eta\prime^{(\mathrm{A})}} \left[\left(\frac{S^{(\mathrm{eff,A})}}{S^{(\mathrm{A})} + S^{(\mathrm{eff,A})}} \right) \right] , \tag{25}$$

$$\Delta_c^{(\mathrm{P})} = \frac{1}{\eta\prime^{(\mathrm{P})}} \left[\left(\frac{S^{(\mathrm{eff,P})}}{S^{(\mathrm{P})} + S^{(\mathrm{eff,P})}} \right) \left(1 - f_1^{(\mathrm{P})} \frac{n_c^{(\mathrm{P})}}{\langle N^{(\mathrm{P})} \rangle} \right) \right] , \tag{26}$$

where $f_1^{(\mathrm{P})}$ is the fraction of plant species that are the only connection of at least one animal species in obligatory mutualism, and $\Delta_c = \min\left(\Delta_c^{(\mathrm{A})}, \Delta_c^{(\mathrm{P})}\right)$.

9 Propagation of perturbations

In the above formula, the propagation of perturbations $\eta\prime$ is numerically computed through Eq.(22). We can predict it analytically in the same approximation used above that the effective mutualistic interactions do not change after the perturbation. We consider relative perturbations of growth rates of size Δ, $\alpha_i \to \alpha_i(1 + \Delta r_i)$, where r_i are independent normal Gaussian variable. The effective growth rate is $\alpha_i^{\mathrm{eff}} = \alpha_i + m_i$ with $m_i = (1/h_i)\,(z_i/(1+z_i))^2$. Under our assumption, the mutualistic growth rate m_i does not change upon perturbation. The productivity Eq.(14) resulting from the perturbation is also a Gaussian variable with the same mean $p_i^{(\mathrm{A})}$ as the unperturbed productivity and variance

$$\overline{\left(\Delta p_i^{(\mathrm{P})}\right)^2} = \Delta^2 \left[\left(\alpha_i^{(\mathrm{P})}\right)^2 + \sum_k \left(G_{ik} \alpha_k^{(\mathrm{A})} \right)^2 \right] \tag{27}$$

where $G_{ik} = \gamma^{(\mathrm{eff,P})} \left(\beta^{(\mathrm{A})}\right)^{-1}$. To compute the term η in Eq.(21) we still need the unperturbed productivity $v_i^1 p^1 \approx \sum_i p_i/S$, and we have to take into account that it depends on

the effective growth rates $\alpha_i^{(\mathrm{eff,A})} = \alpha_i^{(\mathrm{A})} + m_i^{(\mathrm{A})}$. We obtain

$$\eta(\Delta) \equiv \min_i \left(\frac{p_i}{p^1 v_i^1} \right) \approx \Delta \frac{S^{(\mathrm{P})} \sqrt{\left(\alpha_i^{(\mathrm{P})}\right)^2 + \sum_k \left(G_{ik}\alpha_k^{(\mathrm{A})}\right)^2}}{\sum_j \left[\alpha_j^{(\mathrm{P})} + m_j^{(\mathrm{P})} + \sum_k G_{jk}\left(\alpha_k^{(\mathrm{A})} + m_k^{(\mathrm{A})}\right)\right]} \tag{28}$$

This formula is complex, and we prefer to compute $\eta\prime$ numerically; however, it makes clear two important qualitative points: (1) η' decreases with the number of links in the mutualistic network (i.e. the number of non-zero components G_{ik}), and (2) $\eta\prime$ is larger for obligatory mutualism, in which the terms α and m have opposite sign.

10 Interspecific competition and network properties

The effective interspecific competition can be analytically estimated at first order in the effective mutualistic strengths Eq.(11), which must be small to guarantee dynamical stability. We also assume that the direct competition matrix is fully connected and described by the mean-field matrix $\beta_{ij} = \rho + (1-\rho)\delta_{ij}$. Under these assumptions, the effective competition can be explicitly computed. We only give expressions for plants, since those for animals can be obtained interchanging the superscripts.

$$C_{ij}^{(\mathrm{P})} = (1-\rho^{(\mathrm{P})})\delta_{ij} + \rho^{(\mathrm{P})} - \mu_{ij}^{(\mathrm{P})}$$

with

$$\begin{aligned} \mu_{ij}^{(\mathrm{P})} &= \left(\gamma^{(\mathrm{eff,P})} \left(\beta^{(\mathrm{A})}\right)^{-1} \gamma^{(\mathrm{eff,A})}\right)_{ij} \\ &= \frac{\hat{N}^{(\mathrm{A})}\hat{N}^{(\mathrm{P})}}{(1-\rho^{(\mathrm{A})})} \left[\sum_k \gamma_{ik}^{(\mathrm{eff,P})}\gamma_{kj}^{(\mathrm{eff,A})} - \frac{1}{S^{(\mathrm{A})} + S_0^{(\mathrm{A})}} \sum_{kl} \gamma_{ik}^{(\mathrm{eff,P})}\gamma_{lj}^{(\mathrm{eff,A})}\right] . \end{aligned}$$

where

$$S_0^{(\mathrm{A})} = (1-\rho^{(\mathrm{A})})/\rho^{(\mathrm{A})} \tag{29}$$

is the biodiversity scale set by direct competition. Using the approximation $\lambda_1(C) \approx \frac{1}{S}\sum_{ij} C_{ij}$, the effective interspecific competition Eq.(16) can be computed as

$$\begin{aligned}\rho^{(\mathrm{eff,P})} - \rho^{(\mathrm{P})} &\approx \frac{1}{S^{(\mathrm{P})}-1}\left[\frac{1+(S^{(\mathrm{P})}-1)\rho^{(\mathrm{P})} - \sum_{ij}\mu_{ij}/S^{(\mathrm{P})}}{1-\sum_i \mu_{ii}/S^{(\mathrm{P})}}\right] \\ &= \rho^{(\mathrm{P})} - \frac{\sum_{i\neq j}\mu_{ij}}{S^{(\mathrm{P})}(S^{(\mathrm{P})}-1)} + \rho^{(\mathrm{P})}\frac{\sum_i \mu_{ii}}{S^{(\mathrm{P})}}\,. \end{aligned} \tag{30}$$

This formula shows that mutualistic interactions reduce the effective interspecific competition between plants, i.e. $\rho^{(\mathrm{eff,P})} < \rho^{(\mathrm{P})}$, only if the direct interspecific competition parameter $\rho^{(\mathrm{P})}$ is smaller than the critical value $\rho(\mathrm{P}),\mathrm{c}$ given by

$$\rho^{(\mathrm{P}),\mathrm{c}} = \frac{\sum_{i\neq j}\mu_{ij}^{(\mathrm{P})}}{(S^{(\mathrm{P})}-1)\sum_i \mu_{ii}^{(\mathrm{P})}}\,. \tag{31}$$

We can explicitly compute the matrix $\mu_{ij}^{(\mathrm{P})}$ in three situations: when mutualistic interactions are far from saturation for all species, close to saturation for all species, or close to saturation for animals and far from plants, as in obligatory mutualism. Two parameters that define the architecture of mutualistic networks, the connectance κ Eq.(7) and the degree heterogeneity $\left\langle \left(d_k^{(\mathrm{P})}/S^{(\mathrm{A})}\right)^2 \right\rangle$, which is related to nestedness Eq.(8), play a major role in determining the effective competition of mutualistic networks. For simplicity in this computations we neglect the variability of the interaction coefficients c_{ik}, and instead of them we use the binary adjacency matrix a_{ik}.

10.1 Weak mutualism

If all mutualistic interactions are far from saturation ($z_i \ll 1$, see Sec.3) we approximate the effective mutualistic strengthsas $\gamma_{ik}^{(\mathrm{eff,P})} \approx \left(\gamma_0/\sqrt{\hat{N}^{(\mathrm{A})}\hat{N}^{(\mathrm{P})}}\right) a_{ik}^{(\mathrm{P})}$. This is valid if the degree $d_i^{(\mathrm{P})}$ is smaller than the value $d_c^{(\mathrm{P})} = 1/\left(\gamma_0 h_i\right)\sqrt{\hat{N}^{(\mathrm{A})}/\hat{N}^{(\mathrm{P})}}$, and equivalent for animals. If all species are in the weak regime, it holds

$$\mu_{ij}^{(\mathrm{P})} = \frac{(\gamma_0)^2}{1-\rho^{(\mathrm{P})}}\left[\sum_k a_{ik}a_{kj}^T - \frac{d_i^{(\mathrm{P})}d_j^{(\mathrm{P})}}{S^{(\mathrm{A})}+S_0^{(\mathrm{A})}}\right]\,,$$

and a straightforward computation yields

$$\rho^{(\mathrm{eff,P})} - \rho^{(\mathrm{P})} \approx \frac{(\gamma_0)^2 S^{(\mathrm{A})}}{1-\rho^{(\mathrm{A})}} \left[\kappa - F^{(\mathrm{A})} \left\langle \left(\frac{d_i^{(\mathrm{P})}}{S^{(\mathrm{A})}} \right)^2 \right\rangle \right] \left(\rho^{(\mathrm{P})} - \rho^{(\mathrm{P}),\mathrm{c}} \right) \tag{32}$$

where $\kappa = L/(S^{(\mathrm{A})}S^{(\mathrm{P})})$ is the connectance, with L the total number of links, $\langle x_k^2 \rangle = \sum_k x_k^2/S$ and

$$F^{(\mathrm{A})} = \frac{S^{(\mathrm{A})}}{S^{(\mathrm{A})} + S_0^{(\mathrm{A})}} \tag{33}$$

with S_0 as in Eq.(29). $F^{(\mathrm{A})} \leq 1$ is a measure of the richness of animal species with respect to the biodiversity scale $S_0^{(\mathrm{A})}$ of the direct competition, and the critical competition above which mutualism increases the effective competition is

$$\rho^{(\mathrm{P}),\mathrm{c}} = \frac{\left\langle \left(d^{(\mathrm{A})}/S^{(\mathrm{P})} \right)^2 \right\rangle - F^{(\mathrm{A})}\kappa^2}{\kappa - F^{(\mathrm{A})} \left\langle \left(d^{(\mathrm{P})}/S^{(\mathrm{A})} \right)^2 \right\rangle} \left(\frac{S^{(\mathrm{P})}}{S^{(\mathrm{P})}-1} \right) - \frac{1}{S^{(\mathrm{P})}-1} \,. \tag{34}$$

Comparing this equation with Eq.(8), we expect that the critical competition increases with the nestedness. For fully connected networks, $\kappa = 1$, the mean square of the degree distribution is one, and $\rho^{(\mathrm{P}),\mathrm{c}} = 1$. Thus, we recover the result of [1] that fully connected mutualistic networks always decrease the effective competition, and $\rho^{(\mathrm{eff,P})} = \rho^{(\mathrm{P})} - \gamma_0^2(1-\rho^{(\mathrm{P})})S^{(\mathrm{A})}(1-F^{(\mathrm{A})})/(1-\rho^{(\mathrm{A})}) = \rho^{(\mathrm{P})} - \gamma_0^2(1-\rho^{(\mathrm{P})})F^{(\mathrm{A})}$, which coincides with the result reported in [1].

10.2 Strong mutualism

In the strong mutualism limit $z_i \gg 1$ the effective mutualistic interactions are approximately given by $\gamma_{ik}^{(\mathrm{eff,P})} \approx \frac{1}{\hat{N}^{(\mathrm{P})}} \frac{1}{\gamma_0 \left(h^{(\mathrm{P})} \right)^2} \left(\frac{\hat{N}^{(\mathrm{P})}}{\hat{N}^{(\mathrm{A})}} \right)^{3/2} \frac{a_{ik}^{(\mathrm{P})}}{\left(d_i^{(\mathrm{P})} \right)^2}$. In this regime, the mutualistic matrix is given by

$$\mu_{ij}^{(\mathrm{P})} = \frac{1}{\left(\gamma_0 h^{(\mathrm{P})} h^{(\mathrm{A})}\right)^2 \left(1-\rho^{(\mathrm{A})}\right)} \frac{1}{\left(d_i^{(\mathrm{P})} \right)^2} \left[\sum_k \frac{a_{ik}^{(\mathrm{A})} a_{kj}^{(\mathrm{P})}}{\left(d_k^{(\mathrm{A})} \right)^2} - \frac{d_i^{(\mathrm{P})}}{S^{(\mathrm{A})} + S_0^{(\mathrm{A})}} \sum_l \frac{a_{lj}^{(\mathrm{A})}}{\left(d_l^{(\mathrm{A})} \right)^2} \right] ,$$

which yields the effective competition

$$\begin{aligned}\rho^{(\mathrm{eff,P})} &- \rho^{(\mathrm{P})} \approx \frac{1}{\left(\gamma_0 h^{(\mathrm{A})} h^{(\mathrm{P})}\right)^2 \left(1-\rho^{(\mathrm{A})}\right)} \\ &\cdot \frac{1}{S^{(\mathrm{P})}} \sum_{ik} \frac{a_{ik}}{\left(d_i^{(\mathrm{P})} d_k^{(\mathrm{A})}\right)^2} \left(1 - F^{(\mathrm{A})} \frac{d_i^{(\mathrm{P})}}{S^{(\mathrm{A})}}\right) \left(\rho^{(\mathrm{P})} - \rho^{(\mathrm{P}),\mathrm{c}}\right)\end{aligned} \tag{35}$$

wher $F^{(\mathrm{A})}$ is given by Eq.(33) and the critical competition parameter is

$$\rho^{(\mathrm{P}),\mathrm{c}} = \frac{\frac{1}{S^{(\mathrm{P})}} \sum_{ik} \frac{1}{\left(d_i^{(\mathrm{P})}\right)^2} \frac{1}{d_k^{(\mathrm{A})}} \left(a_{ik} - F^{(\mathrm{A})} \frac{d_i^{(\mathrm{P})}}{S^{(\mathrm{A})}}\right)}{\sum_{ik} \frac{a_{ik}}{\left(d_i^{(\mathrm{P})} d_k^{(\mathrm{A})}\right)^2} \left(1 - F^{(\mathrm{A})} \frac{d_i^{(\mathrm{P})}}{S^{(\mathrm{A})}}\right)} \left(\frac{S^{(\mathrm{P})}}{S^{(\mathrm{P})}-1}\right) - \frac{1}{S^{(\mathrm{P})}-1}\,. \tag{36}$$

Once again, for fully connected networks it holds $\rho^{(\mathrm{P}),\mathrm{c}} = 1$, so that mutualism always decreases the effective competition. For sparse networks, the term at the denominator is always positive. Using the approximation $a_{ik} \approx d_i^{(\mathrm{P})} d_k^{(\mathrm{A})}/L$, we see that the term at the numerator is proportional to $\left\langle d_k^{(\mathrm{A})}/S^{(\mathrm{P})} \right\rangle^{-1} - F^{(\mathrm{A})} \left\langle \left(d_k^{(\mathrm{A})}/S^{(\mathrm{P})}\right)^{-1} \right\rangle$, which is negative unless $\kappa = 1$ or $F^{(\mathrm{A})}$ is small. This implies that in the strong mutualism regime mutualistic interactions often increase the effective interspecific competition.

10.3 Obligatory (weak-strong) mutualism

Finally, in obligatory weak-strong mutualism animals are in the strong regime and plants are in the weak regime and it holds

$$\begin{aligned}\mu_{ij}^{(\mathrm{A})} &= \frac{\hat{N}^{(\mathrm{A})}}{\hat{N}^{(\mathrm{P})}} \frac{1}{\left(h^{(\mathrm{A})}\right)^2 \left(1-\rho^{(\mathrm{P})}\right)} \frac{1}{\left(d_i^{(\mathrm{A})}\right)^2} \left[\sum_k a_{ki} a_{kj} - \frac{d_i^{(\mathrm{A})} d_j^{(\mathrm{A})}}{S^{(\mathrm{P})} + S_0^{(\mathrm{P})}}\right], \\ \mu_{ij}^{(\mathrm{P})} &= \frac{\hat{N}^{(\mathrm{A})}}{\hat{N}^{(\mathrm{P})}} \frac{1}{\left(h^{(\mathrm{P})}\right)^2 \left(1-\rho^{(\mathrm{A})}\right)} \left[\sum_k \frac{a_{ik} a_{jk}}{(d_k^{(\mathrm{A})})^2} - \frac{d_i^{(\mathrm{P})} \sum_k a_{jk}/(d_k^{(\mathrm{A})})^2}{S^{(\mathrm{A})} + S_0^{(\mathrm{A})}}\right],\end{aligned}$$

We see from this equation that the mutualistic matrix does not depend on γ_0 for weak-strong mutualism, and therefore the effective competition parameter does not depend on

γ_0 either,

$$\begin{aligned}
\rho^{(\mathrm{eff,A})} - \rho^{(\mathrm{A})} &\approx \frac{\hat{N}^{(\mathrm{A})}}{\hat{N}^{(\mathrm{P})}} \frac{1}{\left(h^{(\mathrm{A})}\right)^2 \left(1-\rho^{(\mathrm{P})}\right)} \\
&\cdot \frac{1}{S^{(\mathrm{A})}} \sum_i \frac{1}{d_i^{(\mathrm{A})}} \left[\rho^{(\mathrm{A})} \left(1 - F^{(\mathrm{P})} \frac{d_i^{(\mathrm{A})}}{S^{(\mathrm{P})}} \right) + \kappa F^{(\mathrm{P})} - \sum_k \frac{a_{ki} d_k^{(\mathrm{P})}}{d_i^{(\mathrm{A})} S^{(\mathrm{A})}} \right] \\
\rho^{(\mathrm{eff,P})} - \rho^{(\mathrm{P})} &\approx \frac{\hat{N}^{(\mathrm{A})}}{\hat{N}^{(\mathrm{P})}} \frac{1}{\left(h^{(\mathrm{A})}\right)^2 \left(1-\rho^{(\mathrm{A})}\right)} \\
&\cdot \frac{1}{S^{(\mathrm{P})}} \left[\left(\rho^{(\mathrm{P})} + \kappa F^{(\mathrm{A})}\right) \sum_k \frac{1}{d_k^{(\mathrm{A})}} - \frac{S^{(\mathrm{A})}}{S^{(\mathrm{P})}} - \rho^{(\mathrm{P})} F^{(\mathrm{A})} \frac{\sum_{ik} d_i^{(\mathrm{P})} a_{ik}}{S^{(\mathrm{A})} (d_k^{(\mathrm{A})})^2} \right]
\end{aligned} \tag{37}$$

The deviation from pure competition is proportional to $\hat{N}^{(\mathrm{A})}/\hat{N}^{(\mathrm{P})}$, and therefore it is very small.

10.4 Other ecological interactions

The computations presented above can be repeated in the same way for other types of ecological interactions. If the two groups of species compete with each other, the sign of the interaction parameter γ_0 would be negative, however γ_0^2 would not change. Therefore, all the results presented above also hold for competitive interactions.

On the other hand, if the two groups of species represent predators and preys, the interaction is positive for one group and negative for the other group, so that we have to substitute γ_0^2 with $-\gamma_0^2$. In particular, the expressions for the critical mutualistic strength remains the same, but its meaning changes: predatory iteractions reduce the effective competition, $\rho^{\mathrm{eff}} < \rho$, if the direct competition is above the critical value, $\rho > \rho^c$, thereby exerting a stabilizing effect, and they increase the effective competition if the direct competition is weak, $\rho < \rho^c$, which is exactly the opposite of what happens with mutualistic interactions.

Therefore, for weak direct competition mutualistic interactions decrease the effective

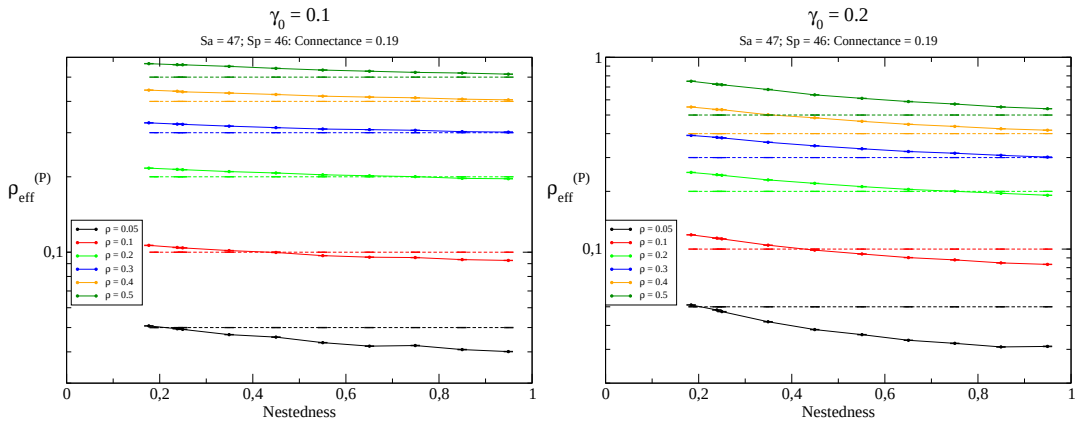


Figure 3: Interspecific effective competition parameter ρ^{eff} versus nestedness for various values of ρ and $\gamma_0 = 0.1$ (left) and 0.2 (right). If $\gamma_0 < \gamma_0^{(1)}$, the absolute value of the difference $\rho^{\text{eff}} - \rho$ increases with γ_0 but the critical interspecific competition at which $\rho^{\text{eff}} = \rho$ does not depend on γ_0.

competition and predatory interactions increase it, while the opposite holds for strong direct competition.

10.5 Numerical results

We show in Fig.3 the effective competition parameter ρ^{eff} versus nestedness for different values of γ_0 and ρ. In the weak mutualism regime ρ^{eff} decreases with nestedness, in agreement with the results of the previous section. The critical competition is given by the point where ρeff $= \rho$. We show in Fig.4 that the critical competition increases with nestedness in the weak mutualistic regime.

We show in Fig.5 how structural stability Δ_c, effective interspecific competition ρ^{eff} and propagation of perturbations η' depend on network properties for some regimes of parameters that are not represented in the main text.

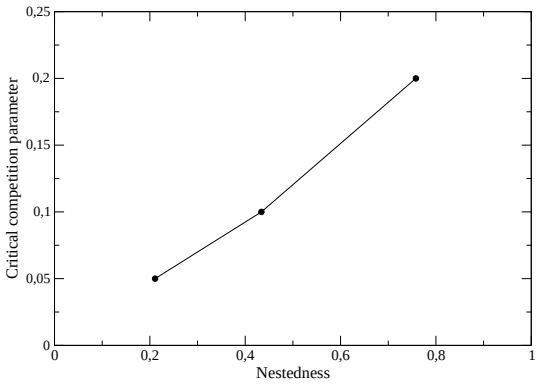


Figure 4: Critical interspecific competition of plants versus nestedness in the weak mutualistic regime for facultative mutualism. The connectance is $\kappa = 0.19$. Computations are performed at $\gamma_0 = 0.1$, which is not expected to influence ρ^c.

11 Relation between structural and dynamical stability

Finally, we look at the relationship between structural stability and dynamical stability. Dynamical stability is fulfilled when the mutualistic strength γ_0 is below and above the two critical mutualistic thresholds, $\gamma_0 < \gamma_0^{(1)}$ and $\gamma_0 > \gamma_0^{(2)}$. Note that the two thresholds can also be interpreted as a measure of structural stability with respect to changes of the strength of mutualistic interactions.

As discussed above, we conjecture that $\gamma_0^{(1)}$ can be estimated as the minimum value of γ_0 at which the effective competition matrix has a vanishing eigenvalue. Since the average of the minor eigenvalues of the normalized effective competition matrix is equal to $1 - \rho^{\text{eff}}$, we expect that the larger is ρ^{eff}, the smaller is $\gamma_0^{(1)}$. This relationship is well fulfilled numerically, comparing different network architectures (see Fig.6 and Fig.7, where we see that $\gamma_0^{(1)}$ is positively related with nestedness, and the inverse of the upper threshold $1/\gamma_0^{(2)}$ is positively related with connectance). Therefore, in the regime A in which Δ_c is negatively influenced by ρ^{eff}, we expect a negative correlation between $\gamma_0^{(1)}$,

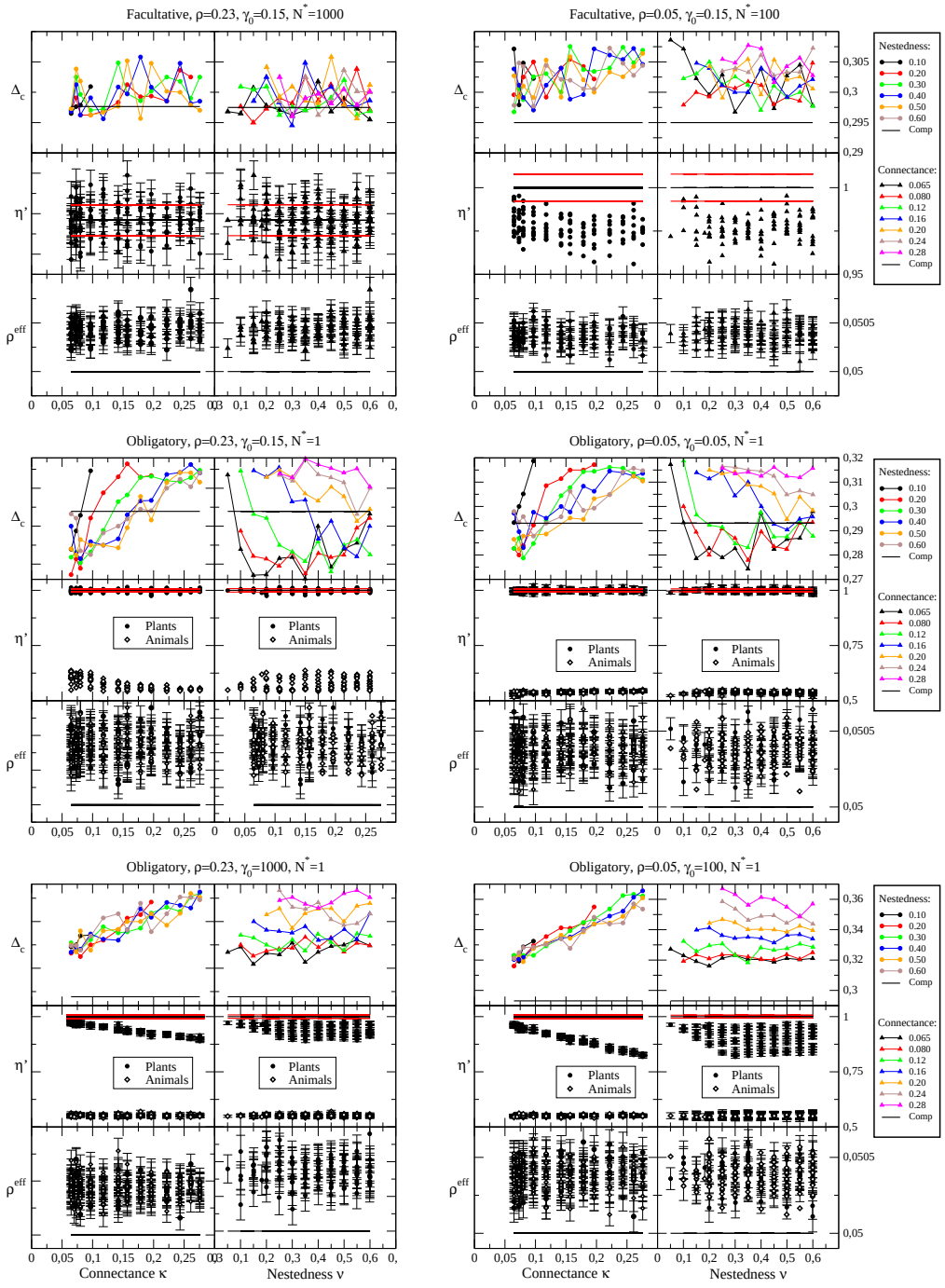


Figure 5: Structural stability, effective interspecific competition parameter and propagation of perturbations versus network architecture (connectance and nestedness) in various regimes. Top line: facultative mutualism, large equilibrium abundance, strong (left) and weak (right) direct competition. Intermediate line: obligatory mutualism, weak mutualistic strengths, strong (left) and weak (right) direct competition. Bottom line: obligatory mutualism, large mutualistic strength, strong (left) and weak (right) direct competition.

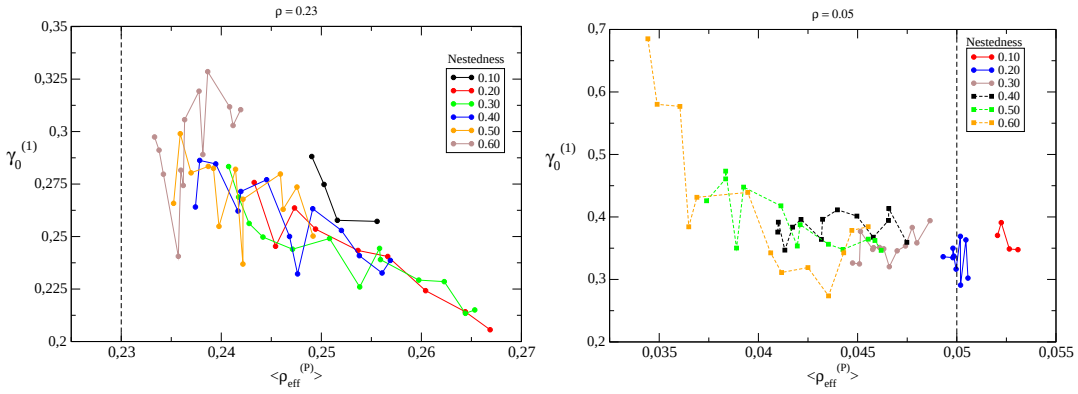


Figure 6: For facultative mutualism, the critical mutualistic threshold $\gamma_0^{(1)}$ decreases with the effective competition parameter ρ^{eff}.

which measures dynamical stability, and Δ_c, which measures the structural stability with respect to fluctuations of the intrinsic growth rates. Conversely, in the regimes in which Δ_c is mainly influenced by the propagation of perturbations, both Δ_c and $\gamma_0^{(1)}$ increase with the connectance of the mutualistic network and we expect that they are positively related. This implies that the structural stability with respect to variation in the mutualistic interactions and with respect to variations in the intrinsic growth rates go in the same direction in these regimes (see Fig.8).

References

[1] Bastolla, U., Fortuna, M.A., Pascual-García, A., Ferrera, A., Luque, B. & Bascompte, J. The architecture of mutualistic networks minimizes competition and increases biodiversity, *Nature* **458,** 1018-1020 (2009).

[2] James, A., Pitchford, J.W. & Plank, M.J. Disentangling nestedness from models of ecological complexity, *Nature* **487,** 227-230 (2012).

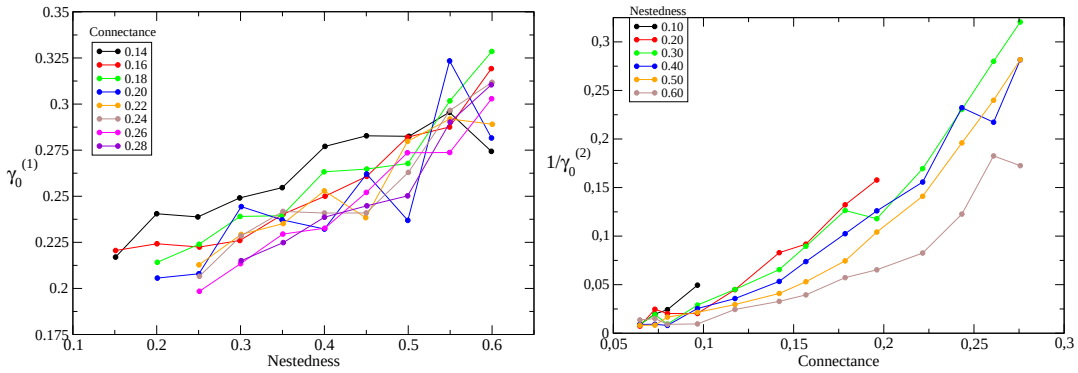


Figure 7: For facultative mutualism, the lower critical mutualistic threshold $\gamma_0^{(1)}$ increases with nestedness for high connectance (left) and it depends little of nestedness for low connectance (not shown). The upper threshold $1/\gamma_0^{(2)}$ increases with connectance (right).

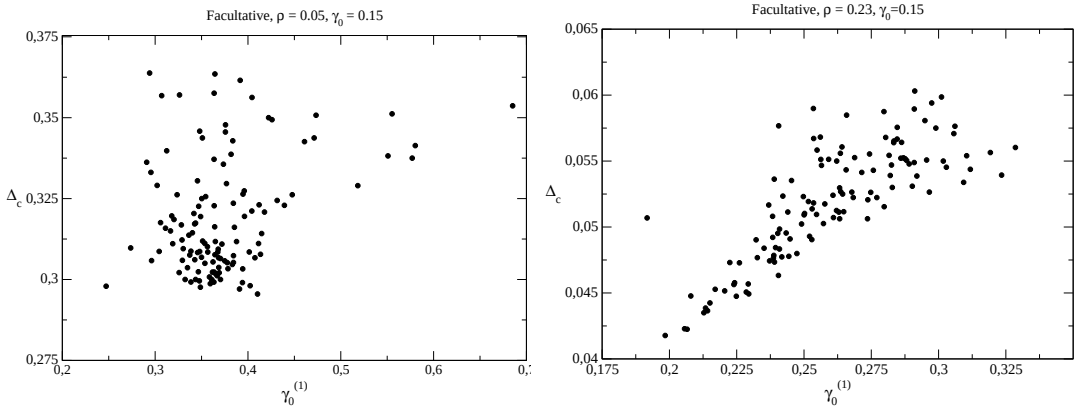


Figure 8: Structural stability Δ_c versus dynamical stability $\gamma_0^{(1)}$ for facultative weak mutualism with weak competition $\rho = 0.05$ (left) and strong competition $\rho = 0.23$ (right). Each point represents a network with different connectance and nestedness.

[3] Pascual-García A., Ferrera A. and Bastolla U., Comment to "Disentangling nestedness from models of ecological complexity" http://arxiv.org/abs/1409.1683

[4] Holland, J.N., DeAngelis, D.L. and Bronstein, J.L. (2002). Population dynamics and mutualism: functional responses of benefits and costs. Am. Nat., 159, 231244.

[5] Jorgensen, S. E., and Svirezhev, Y. M. (2004). Towards a thermodynamic theory for ecological systems. Elsevier.

[6] Ramirez, N. (1989). Biología de polinización en una comunidad arbustiva tropical de la alta Guyana Venezolana. Biotropica 21, 319-330.

[7] Bascompte, J., Jordano, P., Melián, C. J. and Olesen, J. M. (2003). The nested assembly of plantanimal mutualistic networks. Proc. Natl. Acad. Sci. U.S.A. 100, 9383-9387.

[8] Medan, D., Perazzo, R. P., Devoto, M., Burgos, E., Zimmermann, M. G., Ceva, H., and Delbue, A. M. (2007). Analysis and assembling of network structure in mutualistic systems. J. theor. biol. 246, 510-521.

[9] Press, W. H., Flannery, B. P., Teukolsky, S. A., and Vetterling, W. T. (1986). Numerical Recipes: The art of scientific computing (Cambridge University Press, Cambridge).

Numerical Recipes in Fortran 77. The Art of Scientific Computing, 2nd Edition, 1992, ISBN 0-521-43064-X. (Chapter 16.4)

[10] Bastolla, U., Lässig M., Manrubia, S.C. and Valleriani A. Biodiversity in model ecosystems, I: coexistence conditions for competing species. *J. Theor. Biol.* **235,** 521-530 (2005).

[11] Meyer, C. (2000) Matrix analysis and applied linear algebra. SIAM, ISBN 0-89871-454-0.

[12] Goh B.S. Global stability in many-species systems. The American Naturalist (1977) 111, 135-143.

[13] Ferrera A, Pascual-García A and Bastolla U (2014) Effective competition determines the global stability of model ecosystems. Submitted.

[14] Pascual-García A, Ferrera A and Bastolla U (2014) Effective competition determines the structural stability of model ecosystems. Submitted.

Bibliography

Y así, del mucho leer y del poco dormir,
se le secó el celebro de manera que vino
a perder el juicio.

Miguel de Cervantes Saavedra

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. Basic local alignment search tool. *Journal of molecular biology*, vol. 215(3), pp. 403–410, 1990.

Bar-Yam, Y. A mathematical theory of strong emergence using multiscale variety. *Complexity*, vol. 9(6), pp. 15–24, 2004. ISSN 1099-0526.

Bascompte, J., Jordano, P., Melián, C. J. and Olesen, J. M. The nested assembly of plant-animal mutualistic networks. *Proc Natl Acad Sci U S A*, vol. 100(16), pp. 9383–9387, 2003.

Bastolla, U. Detecting selection on protein stability through statistical mechanical models of folding and evolution. *Biomolecules*, vol. 4(1), pp. 291–314, 2014.

Bastolla, U., Fortuna, M. A., Pascual-García, A., Ferrera, A., Luque, B. and Bascompte, J. The architecture of mutualistic networks minimizes competition and increases biodiversity. *Nature*, vol. 458(7241), pp. 1018–1020, 2009.

Bastolla, U., Lässig, M., Manrubia, S. C. and Valleriani, A. Biodiversity in model ecosystems, i: coexistence conditions for competing species. *Journal of theoretical biology*, vol. 235(4), pp. 521–530, 2005.

Bedau, M. A. Weak emergence. *Noûs*, vol. 31(s11), pp. 375–399, 1997.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I. N. and Bourne, P. E. The protein data bank. *Nucleic acids research*, vol. 28(1), pp. 235–242, 2000.

BIALEK, W., CAVAGNA, A., GIARDINA, I., MORA, T., SILVESTRI, E., VIALE, M. and WALCZAK, A. M. Statistical mechanics for natural flocks of birds. *Proceedings of the National Academy of Sciences*, vol. 109(13), pp. 4786–4791, 2012.

BOCCALETTI, S., LATORA, V., MORENO, Y., CHAVEZ, M. and HWANG, D.-U. Complex networks: Structure and dynamics. *Physics Reports*, vol. 424(4-5), pp. 175–308, 2006. ISSN 0370-1573.

BOHM, D. *Causality and Chance in Modern Physics*. University of Pennsylvania Press, 1971. ISBN 9780812210026.

BOLKER, B. M., BROOKS, M. E., CLARK, C. J., GEANGE, S. W., POULSEN, J. R., STEVENS, M. H. H. and WHITE, J.-S. S. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in ecology & evolution*, vol. 24(3), pp. 127–135, 2009.

BONIOLO, G. and VALENTINI, S. Vagueness, kant and topology: a study of formal epistemology. *Journal of Philosophical Logic*, vol. 37(2), pp. 141–168, 2008. ISSN 0022-3611.

BORK, P., SANDER, C. and VALENCIA, A. Convergent evolution of similar enzymatic function on different protein folds: the hexokinase, ribokinase, and galactokinase families of sugar kinases. *Protein Science*, vol. 2(1), pp. 31–40, 1993.

BOSCHETTI, F. Causality, emergence, computation and unreasonable expectations. *Synthese*, vol. 181(3), pp. 405–412, 2011.

BROMHAM, L. and PENNY, D. The modern molecular clock. *Nature Reviews Genetics*, vol. 4(3), pp. 216–224, 2003. ISSN 1471-0056.

CAPITÁN, J. A. and CUESTA, J. A. Species assembly in model ecosystems, i: Analysis of the population model and the invasion dynamics. *Journal of theoretical biology*, vol. 269(1), pp. 330–343, 2011.

CHAO, A., CHAZDON, R. L., COLWELL, R. K. and SHEN, T.-J. Abundance-based similarity indices and their estimation when there are unseen species in samples. *Biometrics*, vol. 62(2), pp. 361–371, 2006.

CHOTHIA, C. Proteins. one thousand families for the molecular biologist. *Nature*, vol. 357(6379), page 543, 1992.

CHOTHIA, C. and LESK, A. M. The relation between the divergence of sequence and structure in proteins. *The EMBO Journal*, vol. 5(4), pp. 823–826, 1986. ISSN 0261-4189.

CHOTHIA, C. and MICHAEL, L. Structural patterns in globular proteins. *Nature*, vol. 261, pp. 552–558, 1976.

COHAN, F. What are bacterial species? *Annu Rev Microbiol*, vol. 56, pp. 457–487, 2002. ISSN 0066-4227.

CONNOR, E. and SIMBERLOFF, D. The assembly of species communities: Chance or competition? *Ecology*, vol. 60(6), pp. 1132–1140, 1979. ISSN 00129658.

CORNING, P. A. The re-emergence of emergence: A venerable concept in search of a theory. *Complexity*, vol. 7(6), pp. 18–30, 2002.

CRUMP, B. C., HOPKINSON, C. S., SOGIN, M. L. and HOBBIE, J. E. Microbial biogeography along an estuarine salinity gradient: combined influences of bacterial growth and residence time. *Applied and environmental microbiology*, vol. 70(3), pp. 1494–1505, 2004.

DELANO, W. L. The pymol molecular graphics system. 2002.

DEMETRIUS, L. and MANKE, T. Robustness and network evolution-an entropic principle. *Physica A: Statistical Mechanics and its Applications*, vol. 346(3-4), pp. 682–696, 2005. ISSN 0378-4371.

DEMETRIUS, L., MATTHIAS GUNDLACH, V. and OCHS, G. Complexity and demographic stability in population models. *Theoretical Population Biology*, vol. 65(3), pp. 211–225, 2004. ISSN 0040-5809.

DEMETRIUS, L. A. Boltzmann, darwin and directionality theory. *Physics reports*, vol. 530(1), pp. 1–85, 2013.

DIAMOND, J. M. The island dilemma: Lessons of modern biogeographic studies for the design of natural reserves. *Biological Conservation*, vol. 7(2), pp. 129–146, 1975. ISSN 0006-3207.

DOBZHANSKY, T. Nothing in biology makes sense except in the light of evolution. *The american biology teacher*, vol. 35(3), pp. 125–129, 1973.

DOUGHERTY, E. R. and BRAGA-NETO, U. Epistemology of computational biology: mathematical models and experimental prediction as the basis of their validity. *Journal of Biological Systems*, vol. 14(1), pp. 65–90, 2006.

FERRERA, A., PASCUAL-GARCÍA, A. and BASTOLLA, U. Effective competition determines the global stability of model ecosystems, 2015.

GAUSE, G. The struggle for existence. williams and wilkins, baltimore, maryland, usa. hardin, g. 1960. the competitive exclusion principle. *Science*, vol. 131, pp. 1292–1297, 1934.

GEORGESCU-ROEGEN, N. *The Entropy Law and the Economic Process*. Harvard University Press, first edition, 1971. ISBN 0674257804.

GOH, B. S. Stability in models of mutualism. *The American Naturalist*, vol. 113(2), pp. 261–275, 1979. ISSN 0003-0147.

GOTELLI, N. J. and MCGILL, B. J. Null versus neutral models: What's the difference? *Ecography*, vol. 29(5), pp. 793–800, 2006. ISSN 1600-0587.

GOTELLI, N. J. and ULRICH, W. Statistical challenges in null model analysis. *Oikos*, vol. 121(2), pp. 171–180, 2012.

GREEN, J. and BOHANNAN, B. J. Spatial scaling of microbial biodiversity. *Trends in Ecology & Evolution*, vol. 21(9), pp. 501–507, 2006.

GRISHIN, N. V. Fold change in evolution of protein structures. *Journal of Structural Biology*, vol. 134(2–3), pp. 167–185, 2001. ISSN 1047-8477.

HARCOMBE, W. R., RIEHL, W. J., DUKOVSKI, I., GRANGER, B. R., BETTS, A., LANG, A. H., BONILLA, G., KAR, A., LEIBY, N., MEHTA, P. ET AL. Metabolic resource allocation in individual microbes determines ecosystem interactions and spatial dynamics. *Cell reports*, vol. 7(4), pp. 1104–1115, 2014.

HARDIN, G. The competitive exclusion principle. *Science*, vol. 131, pp. 1292–1297, 1960. ISSN 0036-8075.

HAYDON, D. Pivotal assumptions determining the relationship between stability and complexity : an analytical synthesis of the stability-complexity debate. *The American naturalist*, vol. 144(1), pp. 14–29, 1994. ISSN 0003-0147.

HOLLAND, T. A., VERETNIK, S., SHINDYALOV, I. N. and BOURNE, P. E. Partitioning protein structures into domains: Why is it so difficult? *Journal of Molecular Biology*, vol. 361(3), pp. 562–590, 2006. ISSN 0022-2836.

HOLM, L. and SANDER, C. Dali/FSSP classification of three-dimensional protein folds. *Nucleic Acids Research*, vol. 25(1), pp. 231–234, 1997. ISSN 0305-1048, 1362-4962.

HORNER-DEVINE, M. C., CARNEY, K. M. and BOHANNAN, B. J. An ecological perspective on bacterial biodiversity. *Proceedings of the Royal Society of London B: Biological Sciences*, vol. 271(1535), pp. 113–122, 2004.

HUTCHINSON, G. E. Homage to Santa Rosalia or why are there so many kinds of animals? *The American Naturalist*, vol. 93(870), pp. 145–159, 1959. ISSN 0003-0147.

IVES, A. R. and CARPENTER, S. R. Stability and diversity of ecosystems. *science*, vol. 317(5834), pp. 58–62, 2007.

JAMES, A., PITCHFORD, J. W. and PLANK, M. J. Disentangling nestedness from models of ecological complexity. *Nature*, vol. 487(7406), pp. 227–230, 2012.

JONHSON, S., DOMINGUEZ-GARCIA, V. and MUNOZ, M. A. Factors determining nestedness in complex networks. *PloS one*, vol. 8(9), page e74025, 2013.

JUSTUS, J. Ecological and lyapunov stability. *Philosophy of Science*, vol. 75(4), pp. 421–436, 2008.

KARPLUS, M., PETSKO, G. A. ET AL. Molecular dynamics simulations in biology. *Nature*, vol. 347(6294), pp. 631–639, 1990.

KETTLER, G. C., MARTINY, A. C., HUANG, K., ZUCKER, J., COLEMAN, M. L., RODRIGUE, S., CHEN, F., LAPIDUS, A., FERRIERA, S., JOHNSON, J. ET AL. Patterns and implications of gene gain and loss in the evolution of prochlorococcus. *PLoS genetics*, vol. 3(12), page e231, 2007.

KLITGORD, N. and SEGRÈ, D. Environments that induce synthetic microbial ecosystems. *PLoS computational biology*, vol. 6(11), page e1001002, 2010.

KOONIN, E. V., MAKAROVA, K. S. and ARAVIND, L. Horizontal gene transfer in prokaryotes: quantification and classification 1. *Annual Reviews in Microbiology*, vol. 55(1), pp. 709–742, 2001.

KOONIN, E. V., WOLF, Y. I. and KAREV, G. P. The structure of the protein universe and genome evolution. *Nature*, vol. 420(6912), pp. 218–223, 2002. ISSN 0028-0836.

LEGENDRE, P. and LEGENDRE, L. *Numerical Ecology*. Elsevier, 2012. ISBN 9780444538697.

LEWIN, R. Santa Rosalia Was a Goat: Ecologists have for two decades made assumptions about the importance of competition in community organization; that idea is now under vigorous attack. *Science*, vol. 221(4611), pp. 636–639, 1983.

LEY, R. E., LOZUPONE, C. A., HAMADY, M., KNIGHT, R. and GORDON, J. I. Worlds within worlds: evolution of the vertebrate gut microbiota. *Nature Reviews Microbiology*, vol. 6(10), pp. 776–788, 2008.

LOVELL, S. C., DAVIS, I. W., ARENDALL, W. B., DE BAKKER, P. I., WORD, J. M., PRISANT, M. G., RICHARDSON, J. S. and RICHARDSON, D. C. Structure validation by cα geometry: ϕ, ψ and cβ deviation. *Proteins: Structure, Function, and Bioinformatics*, vol. 50(3), pp. 437–450, 2003.

LUPAS, A. N., PONTING, C. P. and RUSSELL, R. B. On the evolution of protein folds: Are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? *Journal of Structural Biology*, vol. 134(2–3), pp. 191–203, 2001. ISSN 1047-8477.

LUPYAN, D., LEO-MACIAS, A. and ORTIZ, A. R. A new progressive-iterative algorithm for multiple structure alignment. *Bioinformatics*, vol. 21(15), pp. 3255–3263, 2005.

MACARTHUR, R. Fluctuations of animal populations and a measure of community stability. *Ecology*, vol. 36(3), page 533, 1955. ISSN 00129658.

MARTINY, J. B. H., BOHANNAN, B. J., BROWN, J. H., COLWELL, R. K., FUHRMAN, J. A., GREEN, J. L., HORNER-DEVINE, M. C., KANE, M., KRUMINS, J. A., KUSKE, C. R. ET AL. Microbial biogeography: putting microorganisms on the map. *Nature Reviews Microbiology*, vol. 4(2), pp. 102–112, 2006.

MATURANA, H., LENNEBERG, E. and LENNEBERG, E. Biology of language, the epistomology of reality. In *Foundations of Language Development, a Multidisciplanary approach*, vol. 2. The UNESCO Press, 1975.

MAY, R. M. Will a large complex system be stable? *Nature*, vol. 238, pp. 413–414, 1972.

MAYR, E. The objects of selection. *Proceedings of the National Academy of Sciences*, vol. 94(6), pp. 2091–2094, 1997.

MAZZOCCHI, F. Complexity in biology. exceeding the limits of reductionism and determinism using complexity theory. *EMBO Reports*, vol. 9(1), pp. 10–14, 2008. ISSN 1469-221X.

MILLER, M. B. and BASSLER, B. L. Quorum sensing in bacteria. *Annual Reviews in Microbiology*, vol. 55(1), pp. 165–199, 2001.

MINATI, G., PESSA, E. and ABRAM, M. *Systemics of emergence: research and development*. Springer Science & Business Media, 2006.

MORRIS, B. E., HENNEBERGER, R., HUBER, H. and MOISSL-EICHINGER, C. Microbial syntrophy: interaction for the common good. *FEMS microbiology reviews*, vol. 37(3), pp. 384–406, 2013.

MORRIS, J. J., LENSKI, R. E. and ZINSER, E. R. The black queen hypothesis: evolution of dependencies through adaptive gene loss. *MBio*, vol. 3(2), pp. e00036–12, 2012.

MURZIN, A. G., BRENNER, S. E., HUBBARD, T. and CHOTHIA, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of molecular biology*, vol. 247(4), pp. 536–540, 1995. ISSN 0022-2836.

NAVARRO-ALBERTO, J. and MANLY, B. Null model analyses of presence-absence matrices need a definition of independence. *Population Ecology*, vol. 51, pp. 505–512, 2009. ISSN 1438-3896.

NIDO, G. S., ROMANO, L., BASTOLLA, U. and PASCUAL-GARCÍA, A. Structural bioinformatics within an snake puzzle, 2015.

O'MALLEY, M. A. The nineteenth century roots oféverything is everywhere'. *Nature Reviews Microbiology*, vol. 5(8), pp. 647–651, 2007.

ORENGO, C., MICHIE, A., JONES, S., JONES, D., SWINDELLS, M. and THORNTON, J. CATH - a hierarchic classification of protein domain structures. *Structure*, vol. 5(8), pp. 1093–1109, 1997. ISSN 0969-2126.

ORTH, J. D., THIELE, I. and PALSSON, B. Ø. What is flux balance analysis? *Nature biotechnology*, vol. 28(3), pp. 245–248, 2010.

PASCUAL-GARCÍA, A. *Explorando el rol de la Competición, el Mutualismo y la Arquitectura en Redes Ecológicas: Qué podemos decir sobre la Biodiversidad?*, chapter 6.5. Sociedad Española de Biología Evolutiva, 2009. ISBN 978-84-92910-06-9.

PASCUAL-GARCÍA, A. On the epistemology of complex networks theory. In *AIFBI proceedings*. 2012.

PASCUAL-GARCÍA, A. *Alineamiento de estructura de proteínas*. Publicación independiente, 2014. ISBN 978-84-617-1976-X.

PASCUAL-GARCÍA, A. Epistemology of complex biological systems: insights into dimensionality reduction, constraints identification and emergence from a topological approach, 2015.

PASCUAL-GARCÍA, A., ABIA, D., MÉNDEZ, R., NIDO, G. S. and BASTOLLA, U. Quantifying the evolutionary divergence of protein structures: The role of function change and function conservation. *Proteins: Structure, Function, and Bioinformatics*, vol. 78(1), pp. 181–196, 2010. ISSN 1097-0134.

PASCUAL-GARCÍA, A., ABIA, D., ORTIZ, Á. R. and BASTOLLA, U. Crossover between discrete and continuous protein structure space: Insights into automatic classification and networks of protein structures. *PLOS Computational Biology*, vol. 5(3), page e1000331, 2009. ISSN 1553-7358.

PASCUAL-GARCÍA, A., ANTONIO, F. and BASTOLLA, U. Does mutualism hinder biodiversity? *arXiv preprint*, 2014a.

PASCUAL-GARCÍA, A. and BASTOLLA, U. The complexity-stability relation of mutualistic systems reconciles macarthur and may, 2015.

PASCUAL-GARCÍA, A., FERRERA, A. and BASTOLLA, U. Effective competition determines the structural stability of model ecosystems, 2015.

PASCUAL-GARCÍA, A., TAMAMES, J. and BASTOLLA, U. Bacteria dialog with Santa Rosalia: Are aggregations of cosmopolitan bacteria mainly explained by habitat filtering or by ecological interactions? *BMC microbiology*, vol. 14(1), pp. 1–16, 2014b.

PIMM, S. L. The complexity and stability of ecosystems. *Nature*, vol. 307(5949), pp. 321–326, 1984.

RAMETTE, A. and TIEDJE, J. M. Biogeography: An emerging cornerstone for understanding prokaryotic diversity, ecology and evolution. *Microbial Ecology*, vol. 53, pp. 197–207, 2006.

REGENMORTEL, M. H. V. Reductionism and complexity in molecular biology. *EMBO Reports*, vol. 5(11), pp. 1016–1020, 2004. ISSN 1469-221X.

ROCAP, G., LARIMER, F. W., LAMERDIN, J., MALFATTI, S., CHAIN, P., AHLGREN, N. A., ARELLANO, A., COLEMAN, M., HAUSER, L., HESS, W. R. ET AL. Genome divergence in two prochlorococcus ecotypes reflects oceanic niche differentiation. *Nature*, vol. 424(6952), pp. 1042–1047, 2003.

ROHR, R. P., SAAVEDRA, S. and BASCOMPTE, J. On the structural stability of mutualistic systems. *Science*, vol. 345(6195), page 1253497, 2014.

RYAN, A. J. Emergence is coupled to scope, not level. *Complexity*, vol. 13(2), pp. 67–77, 2007. ISSN 1099-0526.

SACHS, J. L. and HOLLOWELL, A. C. The origins of cooperative bacterial communities. *mBio*, vol. 3(3), 2012. ISSN 2150-7511.

SAMBIN, G. Some points in formal topology. *Theoretical computer science*, vol. 305(1), pp. 347–408, 2003.

SCHRÖDINGER, E. *What is Life?: With Mind and Matter and Autobiographical Sketches*. Cambridge University Press, 1992. ISBN 9780521427081.

SELLA, G. and HIRSH, A. E. The application of statistical physics to evolutionary biology. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102(27), pp. 9541–9546, 2005. ISSN 0027-8424, 1091-6490.

SETH, A. K. Measuring autonomy and emergence via granger causality. *Artificial Life*, vol. 16(2), pp. 179–196, 2010. ISSN 1064-5462.

SKOLNICK, J., ARAKAKI, A. K., LEE, S. Y. and BRYLINSKI, M. The continuity of protein structure space is an intrinsic property of proteins. *Proceedings of the National Academy of Sciences*, vol. 106(37), pp. 15690–15695, 2009. ISSN 0027-8424, 1091-6490.

STEIN, R. R., BUCCI, V., TOUSSAINT, N. C., BUFFIE, C. G., RÄTSCH, G., PAMER, E. G., SANDER, C. and XAVIER, J. B. Ecological modeling from time-series inference: insight into dynamics and stability of intestinal microbiota. *PLoS computational biology*, vol. 9(12), page e1003388, 2013.

SUWEIS, S., SIMINI, F., BANAVAR, J. R. and MARITAN, A. Emergence of structural and dynamical properties of ecological mutualistic networks. *Nature*, vol. 500(7463), pp. 449–452, 2013.

TAMAMES, J., ABELLÁN, J. J., PIGNATELLI, M., CAMACHO, A. and MOYA, A. Environmental distribution of prokaryotic taxa. *BMC microbiology*, vol. 10(1), page 85, 2010.

VALLÈS, Y., ARTACHO, A., PASCUAL-GARCÍA, A., FERRÚS, M. L., GOSALBES, M. J., ABELLÁN, J. J. and FRANCINO, M. P. Microbial succession in the gut: directional trends of taxonomic and functional change in a birth cohort of spanish infants. *PLoS genetics*, vol. 10(6), page e1004406, 2014.

VELLEND, M. Conceptual synthesis in community ecology. *The Quarterly review of biology*, vol. 85(2), pp. 183–206, 2010.

VOLTERRA, V. Fluctuations in the abundance of a species considered mathematically. *Nature*, vol. 118, pp. 558–560, 1926.

Agradecimientos

Y toco apenas y tu bulto aprendo
y torpe sigo lo que tú me indicas.
Lo que no miro, lo que no comprendo,
tú multiplicas.

Letanía del ciego. Carlos Bousoño

En una tesis que habla de evolución, no podría agradecer de otro modo más que intentando entender lo que soy por lo que he sido, y lo que he sido por lo que me ha acompañado. Esto significa que no entraré en el detalle del nombre sino de lo que esconde, con la seguridad de que el lector es buen entendedor.

Desde pequeño en mi casa he encontrado, más allá de lo que dicen necesario para ser filósofo –a saber, el regalo de tener las necesidades básicas cubiertas–, desde Asimov hasta apuntes de álgebra. Siendo mis padres profesores y habiendo c̈oevolucionadoc̈on una empresa dedicada a enseñar a cómo tratar la información, creo que crecí ya en un vaso medio lleno.

De mi primer colegio sólo encuentro relevante la formación musical, a Antonio Brandi que casi podría decir que me enseñó lo poco que sé de biología, el que te hicera gracia llamarme Pitagorín y a Walter –curiosamente profesor de gimansia–, que nunca ha dejado de decirme que valgo para esto tras ganarme al ajedrez.

La cosa cambió en el Instituto San Juan Bautista. Allí recibimos tanta cera en ciencias que ha dejado un buen reguero de investigadores. Público por cierto. Especial recuerdo para Joaquín Hernández, profesor de matemáticas, por su pasión por la docencia, y a Concha Cogolludo, por Nietzsche. Como fue un momento mágico para el instituto también lo fue para el barrio y se materializó en El Flori (su epicentro), en donde se hablaba de ciencia a la vez que de la resaca de los 80 y se practicaba la de los 90. No tengo espacio para tantos nombres implicados.

La Complutense no tiene mucho mérito que llevarse. La gente del zulo sí –por no dejarme caer en la teoría–, la de rugby físicas también –por la realidad oculta de la universidad–, y el destello de algún profesor, puntual para tantos años. En la Universidad de Pavia encontré sin embargo en un

año una docencia tan diametralmente opuesta –¡y más fácil!– que, junto a Leopardi –que entendí básico para la física de la mano de Valentina–, retomé la ilusión por la física. El último año de carrera tuve mi primer contacto con la investigación gracias a Antonio Quiroga, quien me enseñó que lo mejor es enemigo de lo bueno –y que nunca aprendí, motivo por el cual le debo una cena (mínimo)–.

Como no sólo de ciencia vive el hombre allí ya estaba Ángela, que ha sostenido la mayor parte de este esfuerzo que llaman tesis de un modo en el que no encuentro metáfora, sino amor.

Entre tanto, en El Flori se seguía sumando y ese punto de inflexión colectivo que es el fin de carrera a mí me llevó a la biología, y de esto fueron muy culpables Pablo Mateos y Antonio Barbachano (y Schrödinger, claro).

En el Máster de Biofísica de la Autónoma no aprendí todo lo que buscaba pero encontré lo que no esperaba. En particular encontré gente con ilusión por enseñar, con la dedicación de Marisela Vélez y Raúl Guantes, la disponibilidad de Cristina Murga, la originalidad de Juan Poyatos o las collejas verbales de Gonzalo Polavieja. A Raúl le debo especialmente el que confiara en mí para impartir docencia.

Tras el Máster aprendí lo que es ser eficiente como científico gracias a Liset Menéndez. Mi estómago no supo entender todo el espacio que hay en un cerebro, pero todo lo demás creo que sí lo entendí Lis.

Entonces comencé una andadura azarosa en busca de un lugar para hacer la tesis que me llevó a encontrar una serie de personas enrelazadas –y que, de hecho, me presentaron el mundo de las redes complejas– cuya relevancia en esta historia es crítica. Bartolo Luque fue la puerta y allí estaba Lucas Lacasa, compañero de facultad y genial coincidencia, para luego dar con Juan Carlos Nuño, Jordi Bascompte y Susanna Manrubia, quien me enseñó la salida. En todo ese recorrido he visto que ser buen científico no es antagónico a cercano, sencillo e informal. Incluso muy informal.

En esa salida estaba Ugo Bastolla. Desde el punto de vista personal es difícil hablar de quien ha sido tu jefe y no caer en el tópico. Pero si tu jefe sólo se mosquea si le llamas jefe, debe ser verdad. Así que puedo decir con toda tranquilidad que lo que más he ganado con Ugo es una gran amistad. Y, bueno, si desde el punto de vista científico encontráis algo de clase, rigor, buen hacer y, en general, cualquier cosa que os haga despertar admiración hacia mí como científico, por favor, pensad en él. Pero si alguna vez os aburrís también, yo quise meter a Santa Rosalía por aligerar.

Ugo y yo pudimos empezar a funcionar gracias a Ángel R. Ortiz, fundador del laboratorio, quien me dijo que mi misión era ser el Linneo de las proteínas y me financió y codirigió en los primeros años. Era una persona exigente (aunque le gustaba el cutrelux) pero cierto es que el laboratorio estaba cerca de funcionar como un reloj. Por desgracia, Ángel falleció, y quiero recordar la lección de humanidad que él y su mujer Carme nos dieron en sus últimos

meses. También la del grupo de personas que hemos tenido la suerte de estar juntos en el laboratorio durante esos años. Quiero subrayar, en particular, el esfuerzo de Ugo, Antonio Morreale y David Abia, y el apoyo de Federico Gago.

Fue un momento piña único en el laboratorio y en el centro, en donde surgió –por qué no decirlo, también por mérito nuestro– la bien llamada Crema del CBM. Se auto-llama así porque desborda de la cantidad y calidad de los nombres que hay dentro, sin los que estos años hubieran tenido otro color. Baste decir que juntar al Flori con la Crema (y los alrededores de ambos, que llegan hasta Serbia y Bosnia) es de lo mejor que me ha pasado en la vida.

Durante la tesis he hecho una estancia en Buenos Aires y dos en Valencia. En Buenos Aires tengo que agradecer a Julián Echave muchos cafés con alfajores empapados de buena ciencia, política y vida (lo que viene a ser buena Argentina al quitar el psicoanálisis). A Sebas, Palermo y (sobre todo) a Eri les debo el aire, en una ciudad en la que este bien es escaso.

En Valencia, Andrés Moya y su gente no pudieron tratarme mejor en todos los sentidos y, quizá el mejor recuerdo científico, sean las comidas en la universidad y las noches en el Ruta. Caigo ahora en que no todo lo que veáis elegante se lo debo a Ugo. Porque si hay algo de estadística bien hecha se lo debemos a Juanjo Abellán, a quien le tengo que agradecer mucho tiempo de dedicación, y el enseñarme que contar bien es muy complicado.

Me queda por señalar a algunos colaboradores (aún no mencionados) como Javier Tamames –quien puso las bacterias en mi vida–, Antonio Ferrera –que puso las matrices– e Yvonne Vallés y Pilar Francino –muchas cacas de bebé–. I would like also to acknowledge to Silvio Valentini his pacience with me as an outsider in the (very) formal topological world, and to Lloyd Demetrius for his stimulating feedback.

Ya dije que no me gustan los nombres propios porque el olvido es poderoso e inapropapiado. Pero voy a caer con aquéllos que han leído (o al menos intentado) sin venir a cuento alguna de mis locuras (Jorge, Amigo, Guzmán, Duplá), me han seguido incluso en otras (Gon, Ludo, Fernando Rosas), o me han ayudado con las correcciones (Begoña Aguado y mi hermana Silvia).

También agradezco a Bill Gates el que me haya forzado a aprender Linux y a Fons, David y Jorge, que me hayan hecho funcionar con él (y con muchas otras cosas).

Para terminar, como Dios los cría y ellos se juntan, no puedo olvidarme de quien me ha aguantado en el día a día. En el primer tramo de la tesis viví en un piso digno de The Big Bang Theory. Muchas horas de nocturnidad científica con Matthew y George. Éstos alimentaron al monstruo y requeriría otra tesis el explicar su aportación. De ahí entré en un punto de inflexión con Erik, con el que aún compartía la inquietud científica pero nos dejábamos llevar –para compensar– al lado oscuro, que no es más que el motor de la

creatividad. Finalmente diluvió lado oscuro con Bufa, Moro y Raphael, con rituales Made in Cidamón donde un entretenimiento cualquiera podía ser quemar (literalmente) mi pizarra (ecuaciones incluidas).

Y sí, sufrirme no ha sido fácil en el último año. Se han repetido muchos días –como hoy– en los que esas tres, casi cuatro de la mañana han sembrado el sueño y el mal humor de mañana. Aunque también a veces –como hoy– me ha llegado un antídoto de Lara capaz de cambiarme el ánimo.

Así que sólo me queda disculparme por todos los daños colaterales compañeros y compañeras. También por el olvido a todo aquél que, por no ser suficientemente cuidadoso, no se sienta representado, aun formando parte de esta historia.

Agradezco por último al lector el haber llegado hasta aquí.

Y al azar, por todo lo demás.

–I know all about entropy, said Adell, standing on his dignity.

–The hell you do.

–I know as much as you do.

–Then you know everything's got to run down someday.

–All right. Who says they won't?

–You did, you poor sap. You said we had all the energy we needed, forever. You said "forever".

–It was Adell's turn to be contrary. "Maybe we can build things up again someday", he said.

–Never.

–Why not? Someday.

–Never.

–Ask Multivac.

–You ask Multivac. I dare you. Five dollars says it can't be done.

Adell was just drunk enough to try, just sober enough to be able to phrase the necessary symbols and operations into a question which, in words, might have corresponded to this: Will mankind one day without the net expenditure of energy be able to restore the sun to its full youthfulness even after it had died of old age?

Or maybe it could be put more simply like this: How can the net amount of entropy of the universe be massively decreased?

Multivac fell dead and silent. The slow flashing of lights ceased, the distant sounds of clicking relays ended.

Then, just as the frightened technicians felt they could hold their breath no longer, there was a sudden springing to life of the teletype attached to that portion of Multivac. Five words were printed: insufficient data for meaningful answer.

The last question
Isaac Asimov

VITA BREVIS:
ONtología OFF.

El museo de los números
Dimitris Calokiris