



**PROGRAMA DE DOCTORADO EN BIOCENCIAS
MOLECULARES**

**Pharmacological and biological
annotations enhance functional
residues prediction**

Paolo Maietta

Madrid 2017

Departamento de Biología Molecular
Facultad de Ciencias



Pharmacological and biological annotations enhance functional residues prediction

Memoria presentada por Paolo Maietta
Licenciado en Biotecnologías Agrarias Vegetales
por la Università degli Studi di Milano (IT)

Director de Tesis:

Dr. Michael Liam Tress
Centro Nacional de Investigaciones Oncológicas (CNIO)

SUMMARY

El continuo desarrollo de las plataformas de secuenciación masiva ha dado lugar a un incremento vertiginoso en la cantidad de datos genómicos generados y, consecuentemente, a un aumento en la deposición en repositorios públicos de secuencias proteicas sin función conocida. Por esta razón, en los últimos años se han publicado diferentes algoritmos cuyo objetivo es la predicción automática de función de proteínas.

La predicción de función en proteínas presenta algunas características que la convierten en un reto científico apasionante. Para empezar, debido a la existencia de diferentes niveles de complejidad, la misma definición de función no es unívoca. Además se ha demostrado que la relación entre secuencia, estructura y función de proteínas no es lineal, lo que implica que su inferencia mediante homología de secuencia y/o estructura puede dar lugar a errores. Por otra parte, las proteínas pueden llevar a cabo más de una función dependiendo de ciertas condiciones externas, lo que también incrementa la dificultad.

La información funcional más específica se encuentra a menudo asociada a los aminoácidos directamente involucrados en su desempeño. En el laboratorio de Biología Computacional del Centro Nacional de Investigaciones Oncológicas (CNIO) se han desarrollado previamente diferentes herramientas para la caracterización y predicción de estos aminoácidos. En particular, esta tesis utiliza dos de dichas herramientas como punto de partida: FireDB es un repositorio de anotaciones de residuos funcionales y catalíticos extraídos de las estructuras depositadas en el Protein Data Bank (PDB), así como del Catalytic Site Atlas (CSA); *firestar* es un algoritmo que, utilizando esta información, es capaz de inferir sitios de unión a ligandos en proteínas con función desconocida, partiendo únicamente de su secuencia aminoacídica.

Esta memoria describe los desarrollos realizados en el contexto de FireDB y *firestar*. En primer lugar, en la base de datos FireDB se ha realizado un gran trabajo de anotación química y funcional de todos los compuestos no proteicos, así como un estudio detallado de su relevancia biológica. Así mismo, se ha estudiado la relevancia biológica de los sitios de unión a ligando. Dicho estudio ha resultado en el establecimiento un protocolo automático de clasificación que complementa el sistema de evaluación previo. Todos los cambios descritos se han incorporado en FireDB, cuya estructura ha sido mejorada.

En segundo lugar, la incorporación de un nuevo algoritmo de búsqueda de homólogos ha permitido mejorar la sensibilidad de *firestar*. Además, el análisis de los resultados obtenidos en el ámbito de las diferentes ediciones del experimento llamado CASP (Critical Assessment of protein Structure Prediction) ha permitido establecer estrategias para mejorar la especificidad de *firestar*. Los desarrollos derivados de estos trabajos han sido evaluados en un nuevo experimento, CASP10. Un análisis crítico de los resultados obtenidos se presenta en esta memoria.

Finalmente, esta Tesis describe la aplicación de FireDB y *firestar* en tres proyectos con diversos objetivos. En concreto, *firestar* se utilizó en combinación con SIAM, un algoritmo de predicción de función basado en homología, en la segunda edición del experimento CAFA (*Critical Assessment of protein Function Annotation*) con el objetivo de aprovechar las sinergias predictivas de ambos métodos. Además, FireDB y *firestar* se han empleado para estudiar la coherencia funcional de las familias de proteínas definidas por la base de datos Pfam. Por último, ambas herramientas se han integrado como parte del protocolo de construcción de APPRIS, un repositorio que contiene anotaciones sobre isoformas alternativas e identifica la isoforma principal para genes codificantes.

The recent exponential growth of Next Generation Sequencing data has led to a noticeable increase of deposited protein sequences without annotated function. As a result of this, a number of computational methods to automatically infer function have been published in the recent years. However, function prediction is complicated by a number of factors. To begin with, the definition of function itself is not straightforward. Furthermore it has been demonstrated that the relationship between sequence, structure and function is not linear and this means that automatic inference using global sequence homology and/or structure homology is not easily applicable. In addition, many proteins have multiples roles that depend on different factors.

Often the most interesting functional information is to be found at the residue level. We have developed tools to predict functional residues in the CNIO and this thesis takes as starting point two of these, FireDB and *firestar*. FireDB is a database that extracts information about ligand binding sites and catalytic residues directly from the Protein Data Bank (PDB) and Catalytic Site Atlas (CSA). *firestar* is a tool that takes advantage of this structured data to predict binding sites for proteins of unknown function and/or structure.

This thesis describes the many improvements applied to these two tools. For the FireDB database, functional and chemical information has been added for all PDB binding compounds. Ligands have been manually annotated for their biological relevance. In addition the biological relevance of every conserved binding site has been analysed automatically, complementing the existing evaluation schema. All these changes have been included in the revised schema of the database.

The sensitivity of *firestar* functional residue prediction has been increased by the addition of a new search method. At the same time, specificity has been improved by examining the data generated in the Critical Assessment of protein Structure Prediction (CASP) experiments. The new parameters were tested in the tenth CASP edition and the final results are presented.

Finally this thesis describes the incorporation of FireDB and *firestar* into other tools. *firestar* was used in conjunction with a SIAM, a function prediction algorithm based on homology, in the context of the second edition of the Critical Assessment of protein Function Annotation (CAFA) experiment. FireDB and *firestar* were used for a study of the functional coherence of the Pfam database protein families. Finally the two methods have been integrated along with other computational methods in the APPRIS database and web services, which provide annotations for alternative splice isoforms and identify principal isoforms for protein coding genes

TABLE OF CONTENTS

SUMMARY.....	III
TABLE OF CONTENTS	VII
ABBREVIATIONS	XI
1 INTRODUCTION.....	1
1.1 Guess what	3
1.2 Proteins.....	4
1.3 Protein representation.....	4
1.3.1 Sequence.....	4
1.3.2 Structure.....	5
1.3.3 Function.....	6
1.4 Sequence, structure and function relationships.....	7
1.4.1 Evolution and protein function	7
1.4.2 Similarity and homology	8
1.4.3 Sequence and structure space relationship.....	8
1.4.4 Sequence and function space relationship.....	9
1.4.5 Bringing together the three spaces.....	10
1.4.6 The search for homologs in sequence databases.....	10
1.4.7 Function annotation databases	11
1.5 Functional residues	11
1.5.1 Small ligands binding sites	12
1.5.2 Catalytic sites.....	12
1.5.3 Sources of functional residue annotations	13
1.5.4 Sources of catalytic residues annotations.....	14
1.5.5 FireDB.....	14
1.6 Chemical compounds in the PDB.....	15
1.7 Functional site annotation	16
1.7.1 Structure based methods	16
1.7.2 Sequence based methods	17
1.7.3 <i>firestar</i>	17
1.7.4 Evaluation of function prediction methods.....	18
2 MOTIVATION AND OBJECTIVES	20
2.1 Motivation.....	22
2.2 Specific Objectives.....	22
3 MATERIALS AND METHODS	24
3.1 Sequence analysis.....	26
3.2 Molecular Visualization.....	26
3.3 Compound Matching.....	26
3.4 Databases	27
3.4.1 Primary databases.....	27
3.4.2 Databases used for sequence analysis.....	28
3.4.3 Chemical databases	28
3.5 Statistical Methods.....	29
3.5.1 Programming, databases and web services.....	30
3.6 SQUARE. Assessing reliability of pairwise alignments	31
3.6.1 Reliability derived from template's profile	31
3.6.2 Reliability of functional regions and functional transfer	31
3.6.3 Profile generation.....	33
3.7 FireDB.....	34
3.8 <i>firestar</i>.....	36

4	RESULTS	37
	FireDB	
4.1	Compound Annotation	39
4.1.1	The biological relevance of compound.....	39
4.1.2	Ambiguous compounds	40
4.1.3	Metallic compounds	41
4.1.4	Metal binding site conservation	42
4.1.5	Compound cross-references	43
4.1.6	Database mapping.....	44
4.1.7	Bio-activity annotations	45
4.2	Binding sites biological relevance in FireDB	46
4.2.1	Improvements in biological relevance assessment.....	47
4.3	Final database schema and public accessibility	51
	<i>firestar</i>	
4.4	Consensus Predictions	54
4.4.1	Candidate search and filtering.....	54
4.4.2	Candidate merge and generation of a consensus prediction.....	55
4.4.3	New output web page	56
4.4.4	Site reliability score.....	58
4.5	Improvements in <i>firestar</i> algorithm	59
4.5.1	Introduction of HHsearch sequence search method.....	61
4.5.2	Metal Binding Sites	62
4.5.3	Non Metal Binding Sites.....	63
4.5.4	Merging per-residue frequency filter	64
4.5.5	The effect of the filters and new CASP8 dataset assessment.....	65
4.5.6	CASP10 experiment.....	70
	Applications	
4.6	Applications in large-scale collaborative projects	78
4.6.1	Human proteome sites annotation and selection of gene principal isoform.....	78
4.6.2	Pfam domain analysis.....	81
4.6.3	GO terms prediction for large scale annotation projects	82
5	DISCUSSION	87
5.1	Ligand annotation	90
5.2	Biologically relevant binding sites	91
5.3	Availability and future database developments	92
5.4	Functional residues prediction	92
5.5	<i>firestar</i> performance analysis	92
5.5.1	Source information	92
5.5.2	Homologous search methods and alignment quality.....	93
5.5.3	Alignments quality and position conservation.....	94
5.6	The effect of filters on <i>firestar</i> predictions	94
5.7	Applications of FireDB and <i>firestar</i> in large-scale projects	95
6	CONCLUSIONS	97
7	BIBLIOGRAPHY	101
8	APPENDIX	114

ABBREVIATIONS

3D	Three dimensional
BLAST	Basic Local Alignment Search Tool
CAFA	Critical Assessment of protein Function Annotation
CASP	Critical Assessment of protein Structure Prediction
CSA	Catalytic Site Atlas
EC	Enzyme Commission
EM	Electron microscopy
FN	False Negative
FP	False Positive
GO	Gene Ontology
HMM	Hidden Markov Model
InChi	INternational CHEmical Identifier
IUBMB	International Union of Biochemistry and Molecular Biology
IUPAC	International Union of Pure and Applied Chemistry
MCC	Matthews Correlation Coefficient
mmCIF	macromolecular Crystallographic Information File
MS	Master Sequence
MSA	Multiple Sequence Alignment
MSS	Master Sequence binding Site
NMR	Nuclear Magnetic Resonance
PDB	Protein Data Bank
PSI-BLAST	Position-Specific Iterated Basic Local Alignment Search Tool
PSSM	Position Specific Scoring Matrix
RMSD	Root Mean Square Deviation
SMILE	Simplified Molecular-Input Line-Entry System
TN	True Negative
TP	True Positive

1 INTRODUCTION

1.1 Guess what ...

I still remember a simple didactic game we played during a high school science lesson. The teacher gave everyone a square black box with a variable number of holes in 4 of the surfaces, the same number of holes in each surface. Long wooden dowels passed through the holes to create an invisible net inside the box and plastic rings were placed at different points of the net, either on the rods or at the intersection between two rods. The challenge was to guess the number and position of the rings in as little time as possible. It was possible to shake the box, but extracting one or more rods resulted in points being subtracted from your total score. At the end of the game we were allowed to open the boxes and to check how good our prediction was. The game was highly addictive, and while the idea was simple, it brought together a lot of fundamental scientific concepts.

In science we are continuously facing unknowns (the position and numbers of rings) that we cannot directly confirm (the black box) and that we are forced to predict based on a limited amount of (un)connected data (the rods and the sounds). The extraction of the dowels simulates the experimental part of our work. It is possible to experiment, but at a cost; points in the game, but expense and time in reality.

Although the investigation of a single box is intriguing, the computational biologist will want to know if there is a model able to predict the position and number of all the rings in every box. Furthermore I intentionally omitted to mention one detail: the rings were colored. And after the opening our teacher told us that even if we were able to guess perfectly the position and the numbers, there was no possibility to determine the color of the rings inside the black box. And I think this is the most stimulating part of science. Even when you discover something, there are always new questions arising and new aspects that you have not considered before.

So, many years later, I am repeating the game we played in school. Only this time the boxes are proteins and the rings are functionally important amino acid residues.

1.2 Proteins

DNA, RNA and proteins are the principal biopolymers of the cells. The information encoded in DNA sequence controls the genetic makeup of organisms, while RNA is appointed to carry out the instructions encoded in DNA. RNA has different fundamental roles, but the most important is its translation as part of protein synthesis. So genetic information mainly flows from DNA to proteins, which finally fulfill genetic instructions.

A protein is a linear combination of amino acids that folds into a specific tridimensional structure to carry out one or more tasks. We can identify in proteins specific conserved and functionally independent subunits, called domains: these domains often fold into compact, clearly recognizable three-dimensional modules. Many proteins consist of several structural domains and each domain may appear in a variety of different proteins. We can consider domains as building blocks that may be recombined in different arrangements to create proteins with different functions.

Proteins variability and versatility are unmatched among other biomolecules, since they participate directly or indirectly in almost all biological processes. Mechanical support, movement, signaling, regulation and catalysis are just a few examples of the many tasks known to be carried out by proteins. For example, collagen is a structural protein: three protein chains wound together in a tight triple helix, creating molecular cables that strengthen connective tissues. Myosin is the protein responsible for movement, a molecule-sized muscle that uses chemical energy, in form of adenosine-3-phosphate (ATP), to perform a deliberate motion. Hemoglobin is the most abundant protein in red blood cells and its combined subunits, in association with the iron contained in the heme group cofactor, are able to bind and transport different gases. Insulin is a peptide hormone, responsible for controlling the sugar level in blood. When insulin binds another protein, the insulin-receptor, a complex cascade of reactions occurs, leading to the transcription of different proteins that promote the intake of glucose in the cell. Phosphoglucose isomerase (PGI) represents another fundamental class of proteins, the enzymes. PGI catalyzes the inter-conversion of glucose-6-phosphate and fructose 6-phosphate, a critical step in two fundamental cellular pathways as glycolysis and gluconeogenesis. This reaction is driven by the relative concentrations of these sugars in the cytoplasmic matrix of the cell.

Understanding how proteins work, individually and from a systemic point of view, is fundamental not only for molecular biology, but also for biomedicine, since the possibility to finely modulate them has a direct impact onto organisms' physiology.

1.3 Protein representation

From a reductionist point of view, individual proteins can be represented by their sequence, structure and function.

1.3.1 Sequence

Proteins are polymers, made up of amino acids linked together by peptide bonds. This chain of amino acids is the protein sequence. The length of amino acid chains can be very variable. Formally, below 50 amino acids they are called oligopeptides, while above that, polypeptides or proteins. Eukaryotic proteins have an average size of 472 residues¹, but the variations around this value can be important: for example titin, a component of the muscle sarcomere, reaches a total length of more than 18.000 amino acids².

Genetic material (DNA and RNA) information is translated to proteins through genetic code³. There are 20 standard amino acids and their vast permutation makes the number of possible proteins virtually unlimited.

In addition non-standard amino acids can also be found. Some are incorporated directly during biosynthesis of the peptide chain and are specific to some protein families or organisms^{4,5}. Others come from post-translational modifications: these modified amino acids are often essential for the function or regulation of a protein.

Protein sequences can be obtained directly by Edman degradation reaction or mass spectrometry. Despite the technical improvements of the last few years, especially in mass spectrometry⁶, both techniques are still time and money consuming. Due to these limitations, protein sequences are mainly obtained from *in-silico* DNA/mature mRNA translation. The flourishing of genomic and metagenomic sequencing projects⁷, pushed by the rapid costs decrease of Next Generation Sequencing (NGS) techniques, has rapidly generated a vast amount of genetic data that is driving the growth in protein sequences databases. In August 2016 TrEMBL, the UniProtKB⁸ section containing translations of all coding regions extracted from the DDBJ⁹, the European Nucleotide Archive and the GenBank¹⁰ databases, contained more than 65 million sequences. Of these, only 1.8% have their existence confirmed at transcript or protein level.

1.3.2 Structure

Protein structure is the three-dimensional arrangement of the amino acid chain. It is commonly accepted that for a given sequence there is only one possible way of folding under native conditions¹¹. While sequence is the main determinant of the final three-dimensional scaffold, many other factors, like pH, solvent, temperature, presence of chaperones influence the dynamics, speed and the correct folding of a protein¹². Furthermore proteins are not rigid bodies¹³: events like post-transcriptional modifications¹⁴, or the binding of a ligand¹⁵ or a modulator¹⁶ can cause either small conformational changes, such as the movement of one loop, or larger rearrangements, such as the displacement of entire domains of the protein¹⁷. It is worth mentioning here a particular group, the intrinsically disordered proteins. On their own, these proteins lack defined globular structure; but they can undergo a structural transition to a folded form upon interacting with their targets: this characteristic makes them well suited to associate with multiple partners¹⁸.

Protein structure can be experimentally determined using nuclear magnetic resonance (NMR), electron microscopy (EM) and X-ray diffraction, alone or in combination. Unfortunately each of them presents limitations; for NMR the main obstacle is the size, since current technologies are unable to resolve medium or large proteins. The upper limit is around 35KDa¹⁹ (approximately 300 aa), but most structures resolved by NMR are much smaller. EM Achilles heel is resolution; despite improvements over the last decade²⁰, it is still around 5-10 Å, much lower than other techniques. X-ray crystallography is the most widely used method; here the difficulty is the obtaining of a diffractable crystal, since there is no standard protocol applicable to all cases.

On top of that, independently of the technique used, experimental structure determination also entails great cost in terms of time and money, and a final result, if any, may take years to obtain.

Once a structure is resolved, it is deposited in the Protein Data Bank or PDB²¹. The PDB started as a repository of formatted plain text files, but over the years it has become a fully structured database, a unique reference for the crystallographic and structural community. Among its different tasks, the PDB consortium makes structural data easily accessible and establishes standards for data deposition and representation.

The number of individual protein chains stored in the database in August 2016 was more than 367,000, although if the sequences are clustered at 90% redundancy this drops to just 45,000. When compared to the more than 65 million sequences in the databases, it

is clear there is a huge gap between available sequence information and the deposited structural information.

While protein sequence can vary a lot, comparisons between PDB deposited structures shows that these sequences all fold into a limited number of basic structural domains^{22,23}. So in principle protein structure variability can be decomposed into a combination of basic individual folds. This observation gave birth to different structural domain databases, of which probably the most used are SCOP^{24,25} and CATH²⁶. Using as starting point the evolutionary relationship between proteins, they continuously analyze the PDB content and organize it in hierarchical representation of protein structure variability.

1.3.3 Function

The semantic definition of protein function is complex. The function of a protein may be defined by the effect it has on the substrate the protein binds (for example, phosphorylation of another protein), or by the role that it plays as part of one (or more) higher-level process, such as *apoptosis* or *mitosis*. Proteins can also perform more than a single function, depending on different factors, such as location, cellular condition or domain organization.

The translationally controlled tumor protein (*TCTP*) is a good example of multifunctional protein. This protein is highly conserved through its phylogenetic tree and it has been extensively studied. It has been suggested that it regulates the organization of microtubules and the centrosome²⁷, but also that it can work as transcription factor²⁸ and can be secreted as messenger that stimulates histamine release²⁹. Phosphoglucose isomerase (*PGI*) is an example of how location can affect function: in the cytoplasm it is involved in glycolysis and gluconeogenesis, while outside the cell it functions as a neurotrophic factor³⁰, promoting survival of skeletal motor neurons and sensory neurons, and as a lymphokine that induces immunoglobulin secretion³¹. Some intrinsically disordered proteins are able to interact with different targets in the cell regulatory network thanks to their structural adaptability³², and can play a role in different cell pathways.

The scientific community has long worked on controlled vocabularies or ontologies for standardizing protein function definition and classification. The initiative started by Enzyme Commission (EC) is worthy of mentioning in this work. This is a controlled nomenclature³³ following the recommendations of IUBMB (International Union of Biochemistry and Molecular Biology) for one specific subgroup of proteins, the enzymes. Basically, enzymatic catalytic activities are represented in a 4-digit code, which has the form x.x.x.x; the first two levels are hierarchical, whereas the third and fourth levels are specific to each group. For example the EC number 1.1.1.x is associated to alcohol dehydrogenases, which catalyze the oxidation of an alcohol to a ketone or an aldehyde, using NAD⁺ as acceptor. The number can be dissected as follow:

1 →	Class	oxidoreductase (redox catalysis reactions)
1.1 →	Subclass	acts on CH-OH group of donors bonds
1.1.1		uses NADH or NADP as acceptor
1.1.1.1		uses substrate glycerol as donor

The Gene Ontology (GO) Consortium started in 2000 and is the biggest effort to systematically classify protein functions to date³⁴. The goal is to produce a dynamic controlled vocabulary applicable to all organisms and that can be easily adapted as the knowledge about cellular protein roles grows and becomes more accurate. The structure of the GO terms can be described in terms of a graph, where all the ontologies (or GO terms) are nodes, connected with arcs representing the different relationships between them. GO relationships are directional and the graph is acyclic, meaning that cycles are

not allowed. Ontologies resemble a hierarchy, as child terms are more specialized than parent terms, but unlike a hierarchy, a term may have more than one parent. All terms can trace their parentage to the root term, but actually there are three unrelated roots:

cellular component: groups information relating to the cellular compartment or extracellular environment;

molecular function: brings together the elemental activities of a gene product at the molecular level, such as binding or catalysis;

biological process: gathers sets of molecular events with a defined beginning and end, pertinent to the functioning of integrated living units: cells, tissues, organs, and organisms.

1.4 Sequence, structure and function relationships

The total set of sequences so far discovered can be defined as sequence space, just as the whole set of structures or functions can be referred to respectively as the structural and functional spaces. The exact definition of the characteristics of these spaces is still challenging (especially for the functional space), and in this task converge efforts from different groups and public initiatives. Intra-space similarities are informative but inter-space relationships are more interesting to the purposes of this work. Fully understanding of how these spaces are connected and to what extent two sequences that share certain sequence similarity also share structural and functional features is one of the ambitious goals of biologists. It is important for the full comprehension of proteins biology and the underlying mechanisms of their evolution, but also because it is the basis of prediction algorithms that are used to transfer functional and structural information from one protein to another.

1.4.1 Evolution and protein function

Biological evolution is the continuous process of transformation of the species through changes in successive generations, and is reflected in the change of DNA allele frequencies. These changes affect the genetic material, since only germinal genetic information is transmitted from one generation to another (epigenetic changes can be also transmitted³⁵, but they are environment-susceptible). When we talk about molecular evolution, reference is made to the evolution at the molecular level and more specifically to specific changes in nucleic acids sequences and their primary products, RNA and proteins.

Gene variability is due to events that directly change their nucleotide sequence, such as mutations, insertions, deletions, etc. Nevertheless, the reason we do not observe all the possible combinations of DNA sequence in genes is because of selection. In "*The Origin of Species*" Darwin first proposed the theory of natural selection. It states that those modifications that provide adaptive advantages to the organism will prevail. Since natural selection did not explain perfectly the allelic distribution observed in living populations, Kimura in 1983 enunciated the neutral theory of molecular evolution. According to this theory, the prevalence of gene sequences that do not involve adaptive changes occurs as a result of chance, often as the result of genetic drift. This means that without selective forces (or genetic bottlenecks) genes will evolve gradually over time, accumulating changes.

From the next section, we are going to discuss how we can investigate the evolutive relationship between two proteins, and what this implies considering their sequence, structure and function.

1.4.2 Similarity and homology

It is appropriate at this point to define protein similarity and homology and the relationship between them. Similarity and homology are often used in an interchangeable way in scientific literature, but they are actually two distinct concepts. Two proteins are homologous when their encoding genes have a common ancestor. Unlike similarity, homology between proteins is not measurable: it only implies an evolutionary relationship. Similarity and homology are linked, but not always directly. Similar proteins are not necessarily homologous, similarity may arise from convergent evolution events or by chance for sequences with low complexity, while homologous proteins might have little detectable sequence similarity if their common ancestor is very distant in evolutionary terms.

Two aligned protein sequences can be considered similar if they have a significant number of identical (or at least with compatible characteristics) amino acids in corresponding positions. This evaluation is obviously alignment-dependent, meaning that the quantification of the similarity can be different depending on the algorithm used to generate the alignment between two sequences. This is a classical problem in computational biology and over the years several solutions have been proposed^{36,37}

Structural similarity can be calculated using a similar approach. Given two aligned protein structures, their structural similarity depends on the spatial closeness of corresponding atoms in the aligned structures. The traditional similarity measure used is the root mean square deviation (RMSD), the measure of the average distance between the backbone alpha carbon's atoms. RMSD is reliable but it is sensitive to flexible and divergent protein regions. In order to overcome this limitation, different algorithms^{38,39} have been developed that compute the optimal local superposition of structures.

Assessing function similarity is more complicated. In many cases, functional descriptors are based on the available experimental techniques or have been used for historical reasons. Nevertheless, thanks to the GO consortium's efforts to standardize function annotation, different methods^{40,41} have been developed that use GO term graph structure to calculate the distance between the annotated functions of two proteins; a new method⁴² has recently been proposed that tries to bring together all the available functional information, albeit scattered and unrelated, in order fill GO annotation gaps.

1.4.3 Sequence and structure space relationship

The first work to study the relationship between sequence and structure was presented by Chothia and Lesk in 1986⁴³ once a sufficient number of solved structures and the tools for reliable quantification of structure similarity became available. They compared percentage sequence identities with RMSD for homologous pairs of proteins. They concluded that differences in structures correlate directly with sequence changes in a non-linear way; they also claimed that below 20% sequence identity structures might differ substantially.

Other groups⁴⁴⁻⁴⁶ investigated whether there was a percentage sequence identity limit below which structures were not similar. It was found that there was not a well-defined cut-off, but instead there was a blurred range (defined as the twilight zone) from 20 to 35% where the correlation with structural similarity goes down dramatically. Even under this limit there are still many similar structures with very low sequence similarity⁴⁷, some of which may have arisen from events of convergent evolution.

In order to unveil the relation between sequence and structure, a key point is to understand whether structure folding is driven by interactions of a small number of residues or whether it requires the contribution of many residues. Previously cited papers (Chothia and Lesk, Chung and Subbiah, Rost^{43,44,46}) support a "local" model, in which protein folding is driven mainly by a few critical residues, while changes in residues other

than these hotspots have almost no effect on the general tridimensional arrangement. Other groups support a “global” model, in which every change in sequence is followed by a structural change; the magnitude of this change can vary by position and has to be evaluated in the evolutionary context of the protein family. Wood and Pearson in 1999⁴⁸ presented results that showed how sequence similarities and RMSD statistical significances were linearly related. Koehl and Levitt suggested⁴⁹ that variations and not identities correlated linearly and recent work taking advantage of new data and new metrics has supported this hypothesis^{50,51}. However, there are cases that cannot be completely explained either by the “global” model, (such as proteins that share the same fold despite very low percentage identities⁵²), or by the “local” model (such as the results of some designed mutational studies)⁵³. It seems that neither model is able to explain the whole picture alone, and that probably other phenomena, such as coevolution⁵⁴, should be taken into account.

1.4.4 Sequence and function space relationship

Determining functional similarity and its relationship with sequence similarity is complicated. Protein function can be defined by a combination of local (ligand binding sites, protein-protein interfaces, post translational modifications, etc.) and global (folding, stability, conformational changes, etc.) features. There is no standard way to define function, in part because of the breadth of functional space.

We have already mentioned that until the beginning of the GO Consortium project, there was no standardization in functional annotation. The number of manually curated proteins stored in the databases is also really small compared with the total (in UniprotKB, the manually curated part SwissProt⁵⁵ represents just 0.01% of the whole database) and, apart from that, the type of functional annotations generated is strongly influenced by the leading technology. A study⁵⁶ published in 2013 showed how results from high-throughput technologies (such as Mass-spectrometry, RNAi) dominate recent functional annotation. These studies generate information bias: while they determine the cellular component and some specific biological process terms, they provide almost no contribution to molecular function.

Within this tangled panorama, different groups have tried to explore the relationship between sequence and functional spaces. Todd⁵⁷ in 2001 published a work on homologous enzyme superfamilies, taking into account sequence, structure and function (EC numbers). They observed how the four EC numbers rarely change above 40% sequence identity, and that above 30% identity was possible to predict three of the four numbers with a 90% accuracy. Below this threshold structural information had to be taken into account to understand potential functional changes. Devos and Valencia in 2000 published a comprehensive study⁵⁸ to shed a light on how functional conservation correlates with sequence identity. The selected functional features were: EC numbers, keywords extracted from SwissProt, cellular function class according to TIGR⁵⁹ classification and conservation of ligand binding residues. They also took into account the structural changes that occur, using the FSSP domain classification⁶⁰. Comparing the minimum percentage of sequence identities required for the conservation of 50% of functional or structural characteristics, they found that binding site, keywords, and functional class annotations were less conserved than EC numbers, and all of them in turn were less conserved than protein structure.

In a more recent study⁶¹, where functional similarity was inferred based on the distance in the Gene Ontology graph, it was observed that among homologous proteins the proportion of divergent functions increases below a sequence similarity of 50%. For very similar proteins (50% or more identical residues) the chance of completely different annotation is low; however, it is still non-zero. So it seems clear that there is a correlation between sequence and function similarity, but again it seems not to be linear and a

blurred area corresponding to the sequence-structure “twilight zone” is difficult to establish⁶².

1.4.5 Bringing together the three spaces

The studies presented so far clearly demonstrate, using different approaches and metrics, that there is a correlation between sequence, structural and functional similarities. However at the same time they also demonstrate that this correlation is not linear.

Among the three spaces, the structural space is the smallest, and it is commonly accepted that there is a limited group of folds that act as attractors⁶³, since different sequence families often have the same structure, and different functions can be carried out by proteins families with similar 3D arrangements⁵⁷.

In the last years, thanks to the increase of information available, sequence similarity can be calculated with ease, and can be used to infer homology, structural and functional similarities. But every case has to be evaluated individually, and possible effects of specific sequence changes corroborated through evolutionary information.

1.4.6 The search for homologs in sequence databases

The search for possible homologs of a protein of interest based on sequence similarity is a classical task in bioinformatics. It consists of the comparison of the amino acid sequence of the protein of interest with all the sequences stored in a given database and the generation of an optimal alignment between them. These alignments are scored according to preset rules (e.g. rewarding identical amino acids, penalizing gaps, ...) and finally the results are presented in the form of a ranked list.

Generating the optimal alignment between two sequences has a high computational cost³⁶ and if this operation is repeated many times (such as in the case of very large databases), it would be unfeasible. In order to deal with this limitation, the problem has been tackled from two different sides. At algorithmic level, the release of the Basic Local Alignment Search Tool (BLAST³⁷) in 1990 became a computational biology milestone and suppose an historic advance. Instead of exploring all possible solutions for an alignment between two sequences, BLAST uses a heuristic algorithm. It is several times faster than an exhaustive algorithm would be, and it is reasonably accurate. BLAST has become one of the most widely used programs in sequence analysis.

At database level, the speed of the analysis is related to the global size to the order of n^2 , where n is the total number of the sequences analyzed. The rapid growth of sequences databases in the last 10 years pushed the development of clustering algorithms, able to group sequences that share a given level of similarity, so reducing the size of the database with a minimal loss of informative power.

As we said before, proteins are biological entities that are evolving. As result of this process, sequences diverge. Direct comparison between two sequences cannot be sufficient to detect a distant evolutionary relationship and for this reason new algorithms have been developed. The main idea was to generate profiles as well as substitution matrices^{64,65} from pre-calculated alignments, able to capture the specific evolution of the different amino acid positions within a given protein family. The widely used Position Specific Iterative BLAST or PSI-BLAST was one of the first programs using profiles or Position Specific Score Matrices (PSSMs). Later on in 1998, Eddy released HMMER⁶⁶, another widely used program based on Hidden Markov Models (HMMs) profiles, that were more robust than PSSMs, especially for protein domains. HHsearch⁶⁷ algorithm represents a further development. While most methods generate profiles and use them to search against sequences, HHsearch generates a HMM profile and searches against a

pre-calculated database of profiles. This approach⁶⁸ has been demonstrated to be very sensitive and fast.

1.4.7 Function annotation databases

The main source of functional information is the scientific literature. This data is not structured, comes from thousands of detailed small-scale experiments and therefore it is not easy to access. Biocurators search for, bring together, filter and store this information in specialized databases. They are also in charge of updating this information. Input from bioinformatics can speed up the process⁶⁹, but human evaluation and validation is always needed and this makes the entire process time and money consuming.

As previously mentioned, most of the protein sequences in databases are obtained indirectly by translation of DNA sequences⁸. These proteins are hypothetical, meaning that even their *in vivo* expression has not been experimentally determined: almost always structure and function are also unknown. Their deposition rate keeps pace with nucleotide sequences deposition, something that accurate manual curation cannot possibly do. This trend can be easily verified if we look at UniprotKB: Swissprot, which contains the high quality manually annotated protein sequences (although there are functionally unannotated entries), is 130 times smaller than TrEMBL, which contains all automatically DNA-translated protein sequences. In order to narrow this gap, sequence databases automatically annotate proteins, transferring information from high-quality datasets. Electronic annotation is mainly based on the fact that homologous proteins often share similar structural and functional features.

One of the great challenges in bioinformatics is to develop intelligent systems able to make the automatic annotation of protein as reliable as the expert curation. A study⁷⁰ in 2009 showed that even considering protein families for which extensive experimental information is available, SwissProt annotations are very reliable, while derived automatic annotations in three well-known public databases (TrEMBL included) exhibit similar and surprisingly high levels of misannotation. So while homologous protein searching methods⁷¹ are getting better, there is much room for improvement in the identification and the transference of the correct functional features.

1.5 Functional residues

The amino acid sequence of a protein and its structure are intrinsic to its function. The selective pressure on these amino acid residues is not equally distributed. Those amino acids that are more constrained are generally structural determinants (see section 1.4.3) or residues that determine protein functional characteristics⁷².

Functionally important residues can be divided in two sub-groups. A first group is made up of amino acids at the interface in protein-protein interactions or protein-nucleic acids interactions. These interactions are fundamental to many biological processes and may be transient, like those in the signal transduction process, or stable, like those in macro-protein complexes (such as ribosome). Residues at these interfaces are not necessarily conserved, but they have specific features that make them recognizable⁷³.

The second sub-group is made up of residues in regulator and active sites. Regulator sites are important to finely modulate protein activity in response to external signals, while active sites are the places where the molecular function of a protein is carried out. Compared with the first sub-group, these amino acids are usually found in spatially clustered. For these reasons (and others that will be discussed in detail later) they present higher conservation⁷⁴.

This work focuses on regulator and active sites, and in concrete how to collect and use the available experimental information to predict functionally important residues reliably and how to determine possible chemical partners in novel proteins.

1.5.1 Small ligands binding sites

A binding site is defined as the ensemble of amino acids that selectively bind one or more molecules. Binding is mediated by a variety of inter-atomic interactions, principally electrostatic and Van der Waals forces. These forces depend on the amino acid constitution of the site, and this is one of the main reasons for their higher conservation⁷⁴. Other factors such as entropy, hydration, desolvation and flexibility, can play minor but significant roles.

Proteins interact with almost all the different molecular species present in the cells, such as carbohydrates, lipids and hormones among others. Molecules binding at protein active sites can be divided into 2 categories: those needed to carry out the function (cofactors/coenzymes) and those that are the target of the activity (substrates/products). Molecules can actually act as both of them: one can be the cofactor of a protein A and the substrate of other reaction catalyzed by a protein B.

Cofactors and coenzymes are essential to the function of the proteins they interact with. The difference between them is their chemical nature: cofactors are small inorganic ions, while coenzymes are organic molecules⁷⁵. When they are bound tightly, through a covalent bond, they are called prosthetic groups. Union is permanent as long as the native structure of the protein is maintained, such as in the case of the heme group in the hemoglobin. When the interaction is loose, they often act as transient carriers of specific atoms or functional groups. In many cases they can be considered as co-substrates, since they are modified in the chemical reaction. For example the ATP is used to transfer a phosphate group to another entity by the kinases and it is continuously recycled as part of metabolism.

The substrates group is bigger than the group of known co-enzymes. Substrates are all the ligands that are transformed by proteins. Interactions between substrates and proteins are in general transient, with different affinity: for example among the enzymes the turnover number (maximum number of molecules of substrate converted per site per unit of time) varies from 0.5/s for lysozyme to 6×10^5 /s in carbonic anhydrase.

1.5.2 Catalytic sites

Enzymes are sophisticated biological catalysts, able to reduce the activation energy of a reaction and increase its rate under biological conditions. As with other catalysts, they are not consumed during the reaction, nor do they alter the equilibrium. The catalytic site of an enzyme comprises those amino acid residues that participate directly in the reaction catalyzed and can be considered as a subset of the enzyme's binding site.

A profound analysis of the MACiE⁷⁶ database, a publicly available database that gathers detailed information about known enzyme catalytic mechanisms, found that there are seven general classes of catalysis^{77,78}: (de)stabilization of intermediates, steric hindrance, activation of reactive species, covalent catalysis (a bond formed with an intermediate) and proton, hydrogen or electron shuttling. The residues that more frequently appeared in the catalytic sites were (from the most to the least frequent) histidine, cysteine, aspartate, arginine, tyrosine, lysine and glutamate. These are all charged or polar residues at biological pH, which makes sense because catalytic mechanisms often imply movements of charge and/or electrons.

Another important repository of catalytic information is the Catalytic Site Atlas^{79,80} (CSA), a database documenting enzyme active sites and catalytic residues culled from PDB structures. Catalytic residues are those thought to be directly involved in some aspect of the catalyzed reaction. A study published in 2007⁸¹ explored in details catalytic sites combined with structural information extracted from the SCOP⁸² database. The authors found both mechanistic analogues (same catalytic mechanism, related but possibly different reactions) and transformational analogues (same reaction, different mechanisms). The most annotated mechanistic analogue was the catalytic triad, a mechanism in which an amide or ester bond is cleaved by nucleophilic attack. It was found in catalytic sites from 23 enzymatic superfamilies with functions ranging from acyltransferases to peptidases. Unrelated enzymes with different structural organization performing the same chemical transformations were also found. Chloroperoxidases (EC number: 1.11.1.10), acid phosphatases (EC number: 3.1.3.2) and protein Ser/Thr phosphatases (EC number: 3.1.3.16) displayed the greatest biochemical diversity, with three different reaction mechanisms involved in each case. The findings suggested that there are strong evolutionary constraints that guide catalytic site evolution, and some are still not wholly understood.

Due to their fundamental functional role, catalytic residues are highly conserved; a single change would, in most cases, dramatically alter or disrupt the function of the enzyme. In the literature there are numerous examples of disruption of protein function caused by mutations in catalytic^{83,84} or adjacent residues^{85,86}.

1.5.3 Sources of functional residue annotations

There are two fundamental sources of functional residue information: literature mining, searching for experimental evidences of the composition and biochemistry of protein active sites, or data extraction from atomic coordinates of protein structures deposited in the PDB.

The PDB format has a dedicated section where binding information is directly reported⁸⁷, but since there are no standard rules for site composition calculation, it is usually obtained from the atomic coordinates by calculating atomic distances between ligand and residues using the formula:

$$\text{Distance} = \sqrt{(X^p - X^l)^2 + (Y^p - Y^l)^2 + (Z^p - Z^l)^2}$$

Formula 1 *Distance between the atoms of a protein (p) and of a ligand (l) is calculated taking into account the absolute distances respect to the three axis X,Y,Z. Usually two atoms are considered in contact when this distance is lower than the sum of their Van der Waals radii plus 0.5 Angstroms.*

Residues containing atoms below a certain distance threshold from the ligand are considered in contact.

Tools have been developed in order to explore the binding information contained in the PDB and to assess the biological relevance of the. HIC-Up⁸⁸ is a web-based tool to visualize and explore protein-ligand complexes, without providing any additional information. PDBsum⁸⁹ was published in 2001 with the same goal, and authors since then have added new features and information, such as the analyses of ligand binding clusters from different experimental determinations of the same protein⁹⁰.

LigBase⁹¹ was one of the first repositories that used alignments of related sequence and structures for the study of binding sites. In Relibase⁹², published in 2003, functional residue data extracted from the PDB was coupled with protein-ligand interaction features. Binding MOAD⁹³ was released in 2007 with the objective of creating a reference subset of high-quality biologically relevant ligand binding sites. Their selection was based

on the nature of the bound ligands: using pre-established criteria, metals, salts, buffers, solvents and huge compounds, were automatically removed. More or less at the same time Ligasite⁹⁴ was published. It consisted exclusively on the annotation of binding sites in proteins for which at least one apo (ligand free) and one holo (including ligand) structures were available. The biological relevance of binding sites was assessed using literature information and/or ligand size and connectivity. Most recently the Zhang group released a database, BioLip⁹⁵, updated weekly, which principal feature is the identification of biological relevant sites in recently released proteins through a semi-manual protocol. When a new structure is released, the database first sifts out candidate sites based on bound ligand. Selection criteria are the number and type of contacts with the protein, frequency in the PDB and the presence in a pre-generated artifact list. If the ligand satisfies all the requirements, curators manually validate biological relevance, using literature information.

1.5.4 Sources of catalytic residues annotations

Curated databases such as SwissProt⁹⁶ and BRENDA⁹⁷ contain a wealth of information about enzymes and catalyzed reactions. Catalytic site annotation requires a great deal of effort in terms of targeted experiments and expert interpretation. Catalytic site characterization is not possible from atomic coordinates alone, and is not easily standardizable or automatable, so the primary source of information about catalytic sites is the scientific literature, and this information needs expert curation. Consequently, catalytic site annotation databases are less numerous than ligand binding site.

According to Bartlett⁹⁸, a residue can be considered catalytic if it fits one (or more) of the following roles:

- a. It is directly involved in the catalysis as a reactant;
- b. It directly interacts with another residue or water molecule involved in catalysis, helping in the process;
- c. It directly interacts with a substrate or cofactor involved in catalysis;
- d. It stabilizes a transition state.

However, since a widely accepted definition of catalytic residue does not exist yet, the definition and identification of catalytic residues depends on the criteria of each specific annotator.

A number of databases gather enzyme-related information and some of them focus on catalytic residues. The MACiE⁷⁶ and EzCatDB⁹⁹, databases are valuable resources that focus on the amino acidic composition and chemical details of catalytic reactions. The aforementioned Catalytic Site Atlas⁸⁰ is probably the largest reference resource on catalytic sites. In the data gathering phase, manual annotators have extracted molecular details and the amino acids involved in the reaction mechanism from articles associated to almost a thousand PDB structures. This gold-standard dataset was later used for the automatic detection of additional sites in homologous sequences through direct inference, to obtain an extended set.

1.5.5 FireDB

In 2007 was first published FireDB¹⁰⁰, a repository of functional residues developed in our lab. The main objective was to bring together ligands crystallized in PDB structures, the residues in contact with those ligands, and the catalytic sites annotated by hand in the Catalytic Site Atlas. FireDB focused on small compounds; protein-protein and DNA/RNA binding sites were excluded from the database, as previously described. FireDB is more than a simple repository of PDB residue-ligand contacts since it also

attempts to bring order to the interactions. First of all, all sequences are clustered at 97% identity, reducing so the high redundancy of the PDB. If a cluster with more than one sequence is found, a Master Sequence (MS) is generated from the aligned sequences by selecting the most frequent amino acid in each position. In a similar way, binding sites are calculated from the single PDB structures and then are collapsed onto the Master Sequences to generate a Master Sequence binding Site (MSS). The biological relevance of each FireDB MSS is determined from evolutive, structural and empirical characteristics: this approach will be discussed later in the results chapter. The MSS have two important roles: the first is to merge contact information from different proteins, helping to highlight residues that are critical for the binding; the second is to directly transfer ligand contact information to proteins crystallised without a ligand.

1.6 Chemical compounds in the PDB

All the databases presented so far focus their attention on functional residues in proteins, while the analysis of the molecules they bind to is limited to the assessment of the biological relevance of the binding sites. However there also exist databases specialized on ligands and their properties.

The Protein Ligand Database¹⁰¹, published in 2003 provides binding energies and constants and allows comparisons of PDB ligands and geometric similarity searches. LigandDepot¹⁰² (now integrated into the PDB under the name of Ligand Expo) presents ligands with their chemical and structural characteristics as isolated entities. PDB-ligand¹⁰³ is a database in which ligands and their protein environment are structurally aligned and compared. Similar features are offered by the more recent SuperLigands, which also includes a list of drug-like compounds (according to the Lipinski rule of five¹⁰⁴ to evaluate druglikeness) and a search for user determined sub-structures.

All these resources are valuable for rational drug-design or computational screening of drug molecules and this information has been exploited^{105,106} to develop new drugs or investigate characteristics of previously known drugs. However, none of these databases annotate PDB ligands from any point of view other than pharmacological.

The PDB contains a range of small molecules, from biological ligands to inhibitors, analogs, drugs, crystallization additives, and solvents. The main source of information about their nature are the mmCIF¹⁰⁷ dictionaries, elaborated by the wwPDB consortium itself. Here molecules are classified in 11 classes, such as “Saccharides and products”, “Coenzymes” or “Drugs”, but this is a loose classification since diverse compounds can share the same class. When detailed information relating to the nature of the compound is available, it is stored in the related scientific literature, making it difficult to access automatically.

Information about chemical compound properties can be found in dedicated databases. There are several available, each of them with its own functional grouping criteria. Pharmaceutical companies also generate their own repositories that store activity data for tested molecules; the U.S Food and Drug administration stores information about all approved marketed drugs. A number of databases focus their attention on biologically relevant molecules, providing experimental data like bioassays, binding affinities and literature references. The overlap between all these resources is high, since the categorization criteria are not mutually exclusive, but the reciprocal data mapping has been always complicated due to the lack of standardization of the 1D representation of the ligands. Recently the International Union of Pure and Applied Chemistry (IUPAC) created the InChi¹⁰⁸ standard to overcome this problem.

Bringing together all these data with the structural and functional information in the PDB ought to be interesting for a number of reasons, beyond the characterization of a single protein-ligand interaction. For instance, it could also facilitate the discovery of new candidate target-drug associations via drug repositioning. Drug repositioning methods¹⁰⁹

are promising novel approaches for discovering new pharmacological targets for existing drugs.

1.7 Functional site annotation

Several binding site binding prediction algorithms have been developed: a preliminary rough distinction to classify them is based on the source information used to generate predictions: *de novo* or homology transferring. For *de novo* sequence methods, given a problem sequence (target), a multiple sequence alignment (MSA) is generated from the homologous sequences detected in a sequence search against the protein databases. The evolutionary history of the family, and the subfamilies (if any) can be studied from the MSA and the phylogenetic tree: constrained key positions that are conserved within the protein family can be extracted from the alignment. These positions are likely to be related to function or have a structural importance in the protein. Since not all the proteins in the MSA will have exactly the same function, some residues may be conserved in a certain sub-group of sequences rather than completely conserved in the alignment. These positions in the MSA may be important for the detection of subfamilies and the identification of specificity determinants. One of the first approaches to study this information was developed by Casari¹¹⁰ in 1995 and is based on a principal component analysis from the vector representation of an MSA. Another method to identify these positions is evolutionary trace¹¹¹ (ET), that studies the conservation of position across the evolutionary trees, from root to branches. Lately a number of papers¹¹²⁻¹¹⁴ have been published based on this idea, using a range of mathematical models and the increasing availability of protein sequences in the public databases.

Structure-based *de novo* methods also exist¹¹⁵⁻¹¹⁷. Using different metrics, they investigate protein cavities to identify potentially fitting ligands and consequently infer ligand binding site composition. These methods perform quite well with bulky ligand binding sites, but they depend on the availability of a good resolution structure (or model).

From now on we will focus on predictors based on the transfer of functional annotations, since they are more relevant to this work. As previously mentioned, they are mainly based on the conservation of functional and structural characteristics in homologous proteins; in some cases it is also possible to detect convergent evolution. These transfer methods, like the previous examples, can be sequence or structure based, according to the information they are exploiting.

1.7.1 Structure based methods

Since structure is more conserved than sequence and function; structure comparison can detect remote homology that is impossible to recognize from sequence information alone.

In 1997 Wallace and collaborators published an algorithm, TESS¹¹⁸, that was able to scan the PDB and to generate 3D templates of annotated binding sites. These templates were used to analyze a large non-redundant dataset of proteins of known structure in order to uncover function for poorly annotated proteins or to find additional annotations. This promising methodology had one big bottleneck; it was applicable only to resolved structures. Over the years, improvements in the structure prediction field brought new possibilities, and in 2004 a new method called FINDSITE was published. If a structure was not available, models of the target could be generated through structural alignments between the models and a derived database of structures containing co-crystallised ligands. Predicted binding sites were evaluated and ranked. The authors were able to successfully predict 70.9% of the ligand binding sites for a set of 901 proteins.

This prediction strategy has two independent steps and both need to be accurate. If the structure is not available, the model generated for the protein (or more specifically of the binding site) has to be reliable. Then the structural alignment and evaluation criteria have to be good enough to spot binding sites and avoid false positives (like superficial clefts). More recent algorithms have tried to improve one or both of these steps. In 2010 Wass and Sternberg published 3DLigandSite¹¹⁹. In this approach, all candidate sites are grouped based on their 3D position; the cluster with the highest number of ligands is selected as the general area of the binding site. Refinement of the site composition is performed based on the contact occurrence of the single residues and finally a consensus prediction is generated. Residue conservation, calculated from *ad-hoc* generated profiles, is mapped onto the prediction, but it is not discriminating. Other recent methods attempted to improve precision using different solutions for pocket picking and binding site composition selection^{120,121}, or tried to improve every single step, from model generation to the transfer of the annotation¹²².

1.7.2 Sequence based methods

Sequence based approaches search for homologues with experimentally annotated binding residues and use this information to infer the function of the target protein. These methods need an accurate filter process before transferring since, as mentioned before, homology between two proteins means that they have a common evolutionary ancestor, but they may not have the same function.

Methods can be classified according to strategy. Some methods^{123–125} use statistics to rank annotations detected from sequence searches and transfer the best scoring information. However, global sequence comparisons have been shown to be more error prone¹²⁶, so more recent tools^{127,128} go beyond whole sequence comparison to focus on significantly conserved alignment positions. If the target protein has the same conservation pattern as the annotated protein, functional information is transferred. Successive evolutions of this methodology produced functional signatures that are shared by proteins with same function. Inference based on these signatures increases annotation specificity by recognizing functionally inconsistent differences among key residues. Example of these motif-based algorithms are ConFunc¹²⁹, DME¹³⁰, EFICAZ¹³¹.

1.7.3 *firestar*

The first version¹³² of *firestar* was published in 2007 by . It is a sequence-based method that takes advantage of functional signatures to predict functional important residues.

The *firestar* workflow is simple: for a given target, a homology search with PSI-BLAST is performed against the sequences contained in the FireDB. Pairwise alignments are extracted from this analysis and are evaluated by conservation at the residue level (and not by the calculated e-value). Binding information signatures are retrieved from FireDB annotations and compared with the conservation pattern: if they substantially match, a binding site prediction is generated. The key steps in this pipeline are essentially two: the evaluation of the per-residue conservation in the alignment and the assessment of the reliability of the source information.

PSI-BLAST evaluates the reliability its sequence alignments using the statistical significance of the matching positions from the alignments. The reported significance for PSI-BLAST is global: the higher the number of conserved positions, the better the alignment score. PSI-BLAST also produces an evolutionary profile for the target sequence based on the multiple pairwise alignments it generates. These profiles can be transformed into matrices that record scores for each amino acid at each position in the target

sequencen and the matrices can be used to calculate a local amino acid level score for any alignment against the target sequence.

Tress¹³³ showed that the reliability of a heuristic pairwise alignment can be evaluated using profile information from just one of the two proteins. The method evaluates each alignment position using the converted matrix and by considering also the influence of the adjacent positions using a sliding window. To validate the scoring model, structural alignments of the proteins were evaluated: positions with higher reliability values were found in regions where calculated RMSD of alfa carbons was lower. Furthermore it was observed that 80% of ligand binding residues in ligand-binding proteins were located in highly conserved regions of the alignment, and where the structure was also conserved. These regions received systematically a higher score from the method. This work resulted in a tool, SQUARE¹³⁴, which is the core of the local single residue reliability evaluation in *firestar*.

The transference of annotations using functional signatures can be complicated by the presence of multiple predictions from different templates and different ligands. This is especially challenging in the case of bulky yet flexible ligands (e.g. ATP) that in the PDB have a site composition big variability, even in homologous pockets¹³⁵. Automatic biological relevance assessment and size thresholds helped to filter out non-reliable source information: even though the identification of the correct binding site composition can be still challenging.

1.7.4 Evaluation of function prediction methods

Due to the explosion of computational prediction methods in recent years, the scientific community has organized a number of “critical assessments” to independently compare methods. These initiatives are also useful to highlight existing general limitations and bottlenecks and to set new challenges, promoting the progress of the entire field. Two among all these assessments are related function prediction field: CASP and CAFA.

CASP (Critical Assessment of techniques for protein Structure Prediction) is focused mainly on structure prediction. The experiment is essentially a blind test based on structures that are solved but not publicly available and it evaluates how well participant groups are able to predict protein 3D features starting from just the amino acid sequence. To keep pace with emerging needs and lines of research CASP has added a range of specific prediction categories over the years. In the sixth edition of CASP a function prediction category was introduced. At the beginning a number of features were included in the evaluation (for example GO terms and EC numbers prediction). Due to technical difficulties most features were eliminated from the evaluation^{136,137}, and from CASP8 onwards ligand binding site prediction was the only feature evaluated. *firestar* was used as an assessment tool in CASP7 and CASP8^{138,139} and participated as official predictor in the following two^{140,141} CASP experiments. As we will see in the results chapter, CASP played a fundamental role in the improvement of *firestar*.

CAFA (Critical assessment of Function Annotation) is a different community-wide experiment that focuses mainly on function annotation for whole protein sequences. As with CASP, a number of sequences (for which poor or non functional annotations are available) are released to the predictors. Participants have to predict GO terms and/or Human Phenotype Ontologies. The initiative is quite recent; in the first experiment¹⁴² 48,298 sequences were initially released, but just 866 were used as benchmark set for the assessment. The second edition made more than 100,000 sequences available for predictors to predict (although finally only 3,681 of them were used for the assessment). These numbers fit realistically with the emerging needs of massive annotation, as mentioned frequently in this introduction. Even if *firestar* is not a GO predictor from a strict point of view, its predictions have been used and integrated in a more general predictor, SIAM (Angela del Pozo, to be published). SIAM (Statistically Inferred Annotation Method)

is GO terms predictor: given a target sequence, it performs a sequence based homology search among functionally characterized proteins and clusters the results. Once the cluster is defined, SIAM identifies a functional signature, defined as the set of annotations agreed on by the consensus of its member sequences, and transfers them to the target sequence.

firestar enables the identification of binding residues and of bound ligands, and also catalytic residues. This information can shed a light on molecular function, allowing SIAM to add and modify specific associated GO terms.

The third edition of CAFA included the prediction of ligand binding sites for the first time, and *firestar* made a full prediction for almost 41,000 targets. The experiment is now in the evaluation stage.

2 MOTIVATION AND OBJECTIVES

2.1 Motivation

The gap between functionally characterized and unannotated protein sequences is widening due to the explosion of data from Next-Generation Sequencing, and reliable automatic function annotation will be required to close it. The development of accurate and reliable function annotation algorithms is one of the important challenges for bioinformatics, and the number of potential applications for functional information justifies the boom in these methods. Their reliability is crucial since functional information is a starting point for other analyses, and incorrect information could propagate through different levels.

Function can be inferred from homologous proteins, but it has been demonstrated that function relies on a reduced group of residues. This is the reason why even proteins with high sequence similarity can perform different functions. For methods that rely mainly on global sequence comparison, this can be a problem.

Clearly methods that can predict function via local sequence motifs can play a big role here, and FireDB and *firestar* together have been demonstrated to be state-of-the-art resources for ligand binding site annotation and prediction since their publication. For this reason, they are the ideal starting framework to introduce several key improvements, essential for converting both of them in flexible and precise tools, valuable for large scale functional prediction in different contexts.

Improvements in FireDB can be done in biological activity annotation of chemical compounds in complex with resolved protein structure, since database stores little or non information about them. Furthermore we spotted some cases where binding sites automatic biological relevance needs to be polished.

firestar results in the ambit of the CASP experiment show that sensitivity in the search of homologous binding sites is good; nevertheless template sites search could be improved using complementary approaches. On the other hand, the algorithm would require a profound revision in order to pinpoint most informative data extracted from the PDB and to deal with its intrinsic noise, polishing so its specificity.

2.2 Specific Objectives

This work has three lines of research as described below.

Database curation, to provide functional annotation of higher quality:

- Annotation of the chemical compounds in FireDB;
- Revaluation of the biological relevance of all binding *sites* in FireDB;

***firestar* algorithm modification, to increase prediction reliability:**

- Integration of a newly available method to search for remote homologues;
- Introduction of targeted filters to increase the specificity of the predictions;
- Tool assessment in the context of the CASP experiment.

Application of the methods in large-scale analyses:

- Functional coherence analysis of protein families within the Pfam database;
- Integration of *firestar* into a pipeline for the annotation of splice variants;
- Combination of *firestar* predictions with a GO-term based prediction method.

3 MATERIALS AND METHODS

3.1 Sequence analysis

PSI-BLAST¹⁴³ was developed to enhance the sensitivity of BLAST to detect remote protein homologs in a sequence database, generating Position Specific Scoring Matrix (PSSMs) from target sequence. This matrix allows to better weight per-residue evolutionary constraints and to track down remote relationships.

IMPALA¹⁴⁴ is a set of programs that allow to generate and search a database of PSI-BLAST PSSMs. Among other features, it allows you to translate a binary formatted profile to an ASCII formatted array, feature used by SQUARE (see section 3.6).

MUSCLE¹⁴⁵ is a program that generates multiple alignments of nucleotide or protein sequences. MUSCLE alignments combine good quality and low computational cost. The program is freely accessible for download at <http://www.drive5.com/muscle/> web page. It is also available as SOAP web-service at the EBI.

CD- HIT¹⁴⁶ is a program that implements a classification algorithm (clustering) of biological sequences based on the percentage of identity and length to reduce the redundancy of large databases. Its main advantage is the speed, especially important when the selected threshold sequence identity is high. The program is available at: <http://www.bioinformatics.org/cd-hit/>.

HH-suite is an open-source software package for sensitive sequence searching based on the pairwise alignment of hidden Markov models (HMMs). It contains HHsearch⁶⁷ and HHblits¹⁴⁷ among other programs and utilities. HHsearch takes as input a multiple sequence alignment or profile HMM and searches a database of HMMs for homologous proteins. HHblits uses the same HMM-HMM alignment algorithms as HHsearch, but it employs a fast pre-filter that reduces the number of target HMMs from tens of millions to a few thousands.

3.2 Molecular Visualization

PyMOL Molecular Graphics System, Version 1.5.0.4 Schrödinger, LLC is one of the few available open source visualization tools for structural biology. It is suitable for producing high quality 3D images of small molecules and biological macromolecules such as proteins. It can be extended to perform complex analysis of molecular structures using libraries available for Python.

3.3 Compound Matching

InChi Standard¹⁴⁸ The IUPAC International Chemical Identifier (InChi) is a textual identifier for chemical substances, designed to provide a standard and human-readable way to encode molecular information. The identifiers describe chemical substances in terms of layers of information: atoms and their bond connectivity, tautomeric information, isotope information, stereochemistry, and electronic charge information; not all layers have to be provided. It is possible to generate an InChIKey starting from the InChI string, through a compression algorithm that creates a fixed-length string of upper-case characters. The InChIKey has been designed to be easily searched by internet search engines.

UniChem¹⁴⁹ is a freely available compound identifier mapping service. Basically it is a large-scale, non-redundant database of Standard InChIs. Unichem stores pointers between these descriptors and chemical identifiers stored in 34 different chemistry resources. It contains correspondence information from many databases of our interest.

Isomeric SMILES¹⁵⁰ The Simplified Molecular Input Line Entry System is a line notation for entering and representing molecules using short ASCII strings. It is a true language, with a simple vocabulary (atom and bond symbols) and only a few grammar rules. The isomeric SMILES notation allows configuration at tetrahedral centers and double bond geometry specified for any structure, if it is known. This standard is generally considered to be slightly more human-readable than InChI.

3.4 Databases

3.4.1 Primary databases

Protein Data Bank^{21,151} (or PDB) was established in 1971 as an archive for biological macromolecular crystal structures. Over the years it has grown and has now reached almost 130,000 structure depositions, mostly crystal structures, but also many NMR (nuclear magnetic resonance) derived structures and a few models. All the data collected from depositors by the PDB are considered primary data. A part from atomic coordinates, it also includes additional data such as obtained resolution, technical features, experimental details and some more. The PDB is a key resource in the area of structural biology and a number of databases, such as FireDB, are derived from its data.

Catalytic Site Atlas^{79,80} (or CSA) is a freely available catalog of catalytic sites and residues identified in enzymes stored in the PDB. Two types of entries are available:

- A high reliable set, maintained by curators, containing information extracted manually from the primary literature.
- A derived set, containing annotations transferred via homology analysis

Gene Ontology³⁴ (or GO) is a key resource for the function definition and standardization. The Gene Ontology Consortium started in 2000 with the purpose to produce a structured, precisely defined, common, controlled vocabulary for describing the roles of genes and gene products in any organism.

Ontologies are organized as nodes in a network that have a parent-children relationship. There are three main categories of Gene Ontologies.

- **Biological process:** the biological objective to which the gene or gene product contributes.
- **Molecular function:** the biochemical activity (including specific binding to ligands or structures) of a gene product.
- **Cellular component:** the cellular compartment where a gene product is active.

UniProt KnowledgeBase⁸ (UniProtKB) is the central hub for the collection of functional information on proteins and consists of two sections:

- a reviewed section containing manually annotated records with information extracted from literature and curators evaluated computational analysis (UniProtKB/SwissProt).
- an unreviewed section with automatically annotated records (UniProtKB/TrEMBL).

Next to core data (amino acid sequence, protein name or description, taxonomic data and citation references), curation of the protein sequences includes functional sites or regions, as well as variant protein forms produced by natural genetic variation, RNA editing, alternative splicing, proteolytic processing and post-translational modifications (PTMs).

ENZYME³³ database is a repository of information related to the nomenclature of enzymes, organized and elaborated following the recommendations of Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB). Even if every entry has an associated recommended name, EC numbers do not describe enzymes, but enzyme-catalyzed reactions.

3.4.2 Databases used for sequence analysis

PSI-BLAST search needs a reliable profile to have the best possible sensitivity. Since the PDB is too small and biased in terms of sequence representation to generate informative profiles, UniProtKB is the database selected for this task. We do not use the whole official database available, but we pre-filter out known sequence fragments. It has been shown¹⁵² that the inclusion of a high number of incomplete gene sequences, such as those from metagenomic sequencing projects such as the Sargasso Sea¹⁵³ project, in search databases affects the quality of PSI-BLAST profiles. Annotations are retrieved from Swissprot and TrEMBL in order to identify and discard fragments and furthermore the redundancy is then reduced to 70% to improve speed, using the CD-HIT program.

Both **HHsuite** search methods (HHsearch and HHblits) need to generate profiles to search against a profile database. The default database available is called nr20, a clustered database generated from the NCBI data, freely available at <https://github.com/soedinglab/hh-suite>

In order to use these tools against FireDB Master Sequences (MS), a HHsuite formatted database has to be generated after every update. We create a profile for every consensus sequence searching against the nr20 database; and then we combine and index them using the hhblitsdb.pl tool of the package.

3.4.3 Chemical databases

PubChem¹⁵⁴ is an open repository of experimental data relating to the biological activity of small molecules. It is one of the most widely used for deposition and one of the biggest chemical databases. It consists of three primary databases: compounds, substances and bioassays. In May 2017 they contain respectively more than 91 millions, 230 millions and 1.2 millions entries.

Kyoto Encyclopedia of Genes and Genomes¹⁵⁵ (or KEGG) is a long-established project, started in 1995, born from the need for a reference resource that can be used for biological interpretation of genome sequence data. It became a reference pathway database by capturing and organizing experimental knowledge from published literature, first focusing on metabolism but soon followed by other cellular processes. KEGG has been expanded significantly over the years to meet the needs for integrating and interpreting various types of high-throughput data, as well as for supporting translational bioinformatics and now is an integrated database resource consisting of 15 main databases. Besides the systems and genomic information databases, other curated databases bring together chemical and health information.

ChEBI¹⁵⁶ starts as a project of European Bioinformatics Institute (EBI) in 2002, to create a controlled dictionary of Chemical Entities of Biological Interest. The molecular entities in question are either natural products or synthetic products that usually intervene in biological processes. The primary motivation of the database was to provide a high quality, thoroughly annotated vocabulary to promote the correct and consistent use of unambiguous biochemical terminology in molecular biology databases at the EBI.

ChEMBL¹⁵⁷ is an Open database containing binding, functional and ADMET (**A**bsorption, **D**istribution, **M**etabolism, **E**xcretion and **T**oxicity, a form of describing the disposition of a pharmaceutical compound within an organism) information for a large number of drug-like bioactive compounds. These data are manually extracted from the primary published literature on a regular basis, then further curated and standardized. Currently, the database contains 5.4 million bioactivity measurements for more than 1.5 million compounds over almost 10 thousand protein targets.

DrugBank¹⁵⁸ database is a unique bioinformatics and chemoinformatics resource that combines detailed drug (i.e. chemical, pharmacological and pharmaceutical) data along with drug target (i.e. sequence, structure, and pathway) information. It has become a reference database for drug-discovery studies, includes data from the FDA (Food and Drug administration) and since 2008, the year of publication, 3 updates have been published with more than 2,000 citations.

MetaCyc¹⁵⁹ is a curated database of experimentally described metabolic pathways. It contains data that have been experimentally validated and reported in the scientific literature. Even if the main objects of MetaCyc are metabolic pathways and reactions, for us the available collection of more than 13 thousands (in May 2017) metabolites is invaluable to spot possible biologically relevant compounds.

PharmGKB¹⁶⁰ is a publicly available web-based knowledge base whose aim is to aid researchers in pharmacogenomics studies: as a matter of fact genetic variants can be considered the main entities. Born in 2001, all the data provided has been validated through extensive manual and automatic curation; the user can browse not only the biological role of the genes affected by the mutation of interest, but also can retrieve clinical interpretation and even pharmacologically relevant molecules associated with the mutational landscape.

3.5 Statistical Methods

Sensitivity and specificity: these are the principal measures of the quality of a binary classification. Given a classification of a specific data set, there are four basic data sets: correct predictions (TP), incorrect predictions (FP), correct negative predictions (TN) and missed predictions (FN). Sensitivity measures the ability to correctly classify TP while specificity measures the degree of precision with which the TN are identified.

$$sensitivity = \frac{TP}{TP + FN} \qquad specificity = \frac{TN}{TN + FP}$$

Matthews Correlation Coefficient (or MCC) is often used in machine learning as a measure of the quality of binary (two-class) classifications. Unlike sensitivity and specificity, it is generally considered a balanced measure that can be used even if the classes are of very different sizes. The MCC is in essence a correlation coefficient

between the observed and predicted. It returns a value between -1 and +1 where the first indicates total disagreement between prediction and observation and +1 represents a perfect prediction; 0 means no better than random prediction.

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}}$$

This measure was used in the CASP experiments for the first assessment of ligand-binding site prediction category^{139–141}

3.5.1 Programming, databases and web services.

SQL (Structured Query Language) is a special-purpose programming language designed for managing data held in a relational database management system (RDBMS). It consists of a data definition language and a data manipulation language. The scope of SQL includes data insert, query, update and delete, schema creation and modification, and data access control. FireDB has been written in SQL language.

MySQL is a popular open-source relational database (SQL) management system used for the administration of FireDB database.

Perl (or *Practical Extraction and Report Language*) is an interpreted programming language derived from C and awk. Its forte is its flexibility and its ease for parsing text files. Supported by a wide community, it has many useful libraries (eg: DBI, for the database management or CGI for the generation of dynamic content on the web page). Almost all the script behind the firesuite web server has been written in this language.

Apache HTTP Server is an open-source community supported web server software. It's widely used and supports a variety of features, many implemented as compiled modules which extend the core functionality, and from the server-side fully support Perl, the language that most of the firesuite scripts are written in.

PHP (Personal home page Hypertext Pre-processor) is an interpreted programming language, originally designed for creating dynamic web pages. Interpretation mainly occurs on the server side. All the web-forms in the firesuite server are written in PHP.

JavaScript is a dynamic computer programming language. It is most commonly used as part of web browsers, whose implementations allow client-side scripts to interact with the user, control the browser, communicate asynchronously, and alter the displayed document content. JavaScript copies many names and naming conventions from Java, but the two languages are otherwise unrelated and have very different semantics. For the firesuite web-page forms, many pre-submission checks (eg. for FASTA format or PDB ID format) are written in JavaScript.

REST is an *architecture style* for designing networked applications. The name stands for **R**epresentational **S**tate **T**ransfer. The idea is that, rather than using complex mechanisms, such as SOAP, to connect between machines, simple HTTP is used to make calls between machines. So to query a server you can use just a URL. This URL is

sent using a simple GET request, and the HTTP reply is the not embedded raw result data.

3.6 SQUARE. Assessing reliability of pairwise alignments

SQUARE¹³⁴ evaluates the local reliability of pairwise sequence alignments using PSSMs obtained from PSI-BLAST profiles via IMPALA, that have been pre-calculated for every template included in the target database. The method is based on an earlier study¹³³ and it is a keystone in the generation of FireDB since profiles are generated for each Master sequence. However, it is even more important in *firestar*, because identification of the reliable regions in the alignment between two sequences is necessary for transferring functional information from the template sequence to the target sequence.

3.6.1 Reliability derived from template's profile

Unlike other methods that study the evolutionary behavior of a position in a multiple alignment, SQUARE considers the residue window surrounding each position being evaluated.

So:

$$Score = S_a^{res-2} + 2 \cdot S_a^{res-1} + 3 \cdot S_a^{res} + 2 \cdot S_a^{res+1} + S_a^{res+2}$$

where S_a is the score from the matrix of the PSI-BLAST profile for residue *res*. The effect of the inclusion of the residue's environment is shown in figure 1. Calculating the score for a residue window and not just from the evolutive information for each position smooths the variation of the score between adjacent positions.

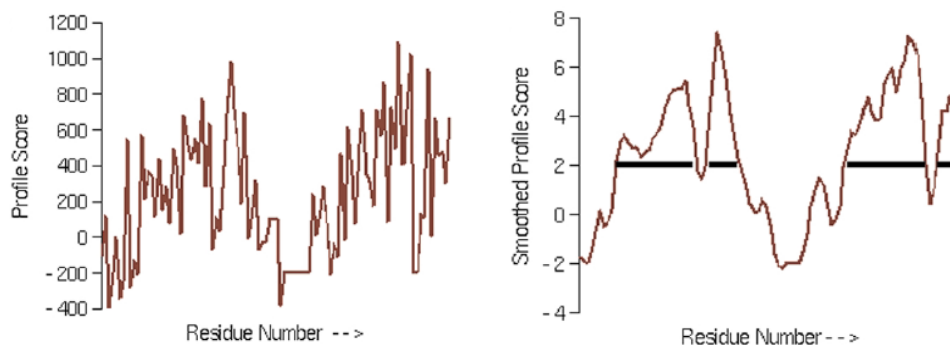


Figure 1 Smoothing of the reliability score for the alignment between the sequences corresponding to PDB codes 1plc and 1aac. On the left a value for each aligned residue was calculated from the template sequence profile for each residue position; on the right the smoothed values considering a window of 5 residues (from Tress et al¹³³)

Calculating a reliability score taking into account the environment also makes biological sense because if a residue conserved between two proteins is to maintain the same role, the surrounding residues must also maintain a certain amount of conservation.

3.6.2 Reliability of functional regions and functional transfer

Given a pairwise alignment, SQUARE assigns a score to every position; this reflects the probability that two aligned residues share the same role in both proteins. In the paper¹³³ the author used protein structural alignments to evaluate the position-based scoring and found that correctly aligned residues had a better mean position-based score. What was most interesting was the fact that correctly aligned binding sites and catalytic residues were located in regions of high reliability and usually had very high scores, so the SQUARE reliability values could be particularly useful to predict functionally important residues. In FireDB, and in the SQUARE and *firestar* web-servers, the position-based scores have been discretized and associated to a shade of blue in order to make them more user-friendly: the higher the score, the darker the blue (table 1).

SQUARE: table of scores	
-	Gapped or non-conserved position
1	45% reliability
2	60% reliability
3	75% reliability
4	85% reliability
5	90% reliability
C	99% reliability

Table 1 Reliability values for alignment position evaluated by the program SQUARE.

As an example of SQUARE results, we show here the *firestar* analysis output of a putative tryptophan synthase from *Aeropyrum camini*, a hyperthermophilic archaea (Uniprot code U3TBS7, figure 2). PSI-BLAST finds as a similar template (PDB ID 1k3u). This is a tryptophan synthase from *Salmonella typhimurium*, the two tryptophan synthase sequences have diverged considerably and have just 20% identity. Despite this the color coded SQUARE alignment reliability scores clearly highlight the regions surrounding the catalytic residues (Glu 2, Asp 13 and Tyr 129) and the residues involved in the binding of the coenzyme pyridoxal-5'-phosphate; in this way SQUARE scores help to detect the most conserved functional regions even in distant homologs.

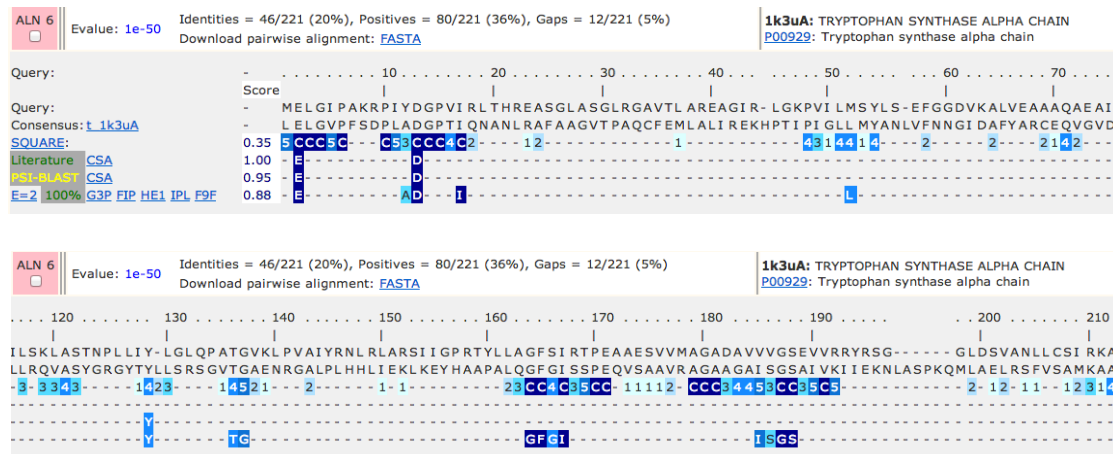


Figure 2 SQUARE evaluation for an alignment between a target sequence, tagged as “Query”, and a template, (PDB ID 1k3u) a Tryptophan synthase alpha chain. SQUARE generates the reliability values, reported as colors in the first line below the alignment. Lines 2-4 below the alignment show the reliability scores for the functional annotations retrieved from FireDB. Every residue is represented with a background that reproduces the same color of the SQUARE score. The darker the blue, the more the position is conserved and the higher the score.

Besides the integration in FireDB and *firestar*, SQUARE is also available at <http://firedb.bioinfo.cnio.es/Php/square.php> as a stand-alone web server.

3.6.3 Profile generation

Since SQUARE scores are derived from profiles built around the template sequences, template profiles need to be pre-calculated. For the version of SQUARE integrated into *firestar* and FireDB, due to the size and biases in PDB sequence space, we generated Master Sequences Profiles against an ad-hoc clustered version of UniProtKB database (see section 3.4.2 for details).

3.7 FireDB

FireDB¹⁰⁰ was first published in 2007. It is a databank of functional information relating to proteins with known structure. FireDB is built around ligand binding data obtained from the 3D structures in the Protein Data Bank (PDB) and the catalytic residues in Catalytic Site Atlas (CSA).

Programmatically, the database was built in SQL and it is managed using MySQL; Perl is used for the access, through DBI library.

The original database schema consisted of 18 tables (see appendix figure 1); Perl scripts carry out the periodical updates, generating all the information from scratch. Here we give an overview of the most relevant tables:

COMPOUND: stores basic information about the small molecules co-crystallised with the PDB structures. This data comes from the mmCIF⁴ library generated by the wwPDB⁵ consortium.

INFOACC: contains sequence information from the PDB, the correspondence between PDB and Uniprot IDs and functional annotation (GO terms and EC numbers).

SITE_{xx}: three tables that store binding site composition, calculated directly from the PDB coordinates using different distance cut-offs (all distances calculated plus Van der Waals radii): 0.5 (SITE₃₅), 1 (SITE₄₀) and 1.5 Å (SITE₄₅). They also store catalytic information extracted from the Catalytic Site Atlas and mapped onto the protein sequence.

CONSENSUS: The PDB is highly redundant. FireDB uses CD-HIT to cluster proteins at 97% identity and generates alignments for each of these clusters with MUSCLE. These alignments generate a new sequence entity, called the Master Sequence (MS), as the cluster representative.

CSITE_{xx}: these tables are built from the **SITE_{xx}** information. All binding sites from templates in the same cluster with an overlap greater than 50% of their size are mapped onto the Master Sequences (MS) to create Master Sequence binding Sites (MSS). The residues that form the MSS are obtained from the collapse of the separate sites. CSA residues undergo the same process, but separately from the MSS. In this case the overlap has to be 100% in order to be clustered into a Master Sequence Catalytic site (MSC).

BINDSITE_{xx}: in some binding sites we can find more than one compound (eg: ATP and MG), but in CSITE_{xx} tables binding information is joined in a unique MSS. BINDSITE_{xx} tables store per compound binding information for all these composite sites

COMPARE_{xx}: In order to identify inter-cluster homologous sites, an all-against-all PSI-BLAST analysis of the Master Sequences is performed. When a hit is found and two MSS overlap for more than 40% of their size, SQUARE evaluates the alignment and this information is stored in these tables.

CCTEVAL_{xx}: these tables bring back together some MSS features to automatically assess their biological relevance. Parameters like the occurrence (the fraction of chains within the cluster that have the MSS occupied), the tendency of the chemical heterogeneity of the bound ligands compounds to occupy conserved sites in the PDB, and the absolute and relative residues conservation are taken into account.

FireDB is publicly accessible and users can browse the available data using a web interface. They can search for a specific PDB chain, Uniprot ID or for a keyword contained in the proteins descriptions. The output of the query is shown in figure 3. The entire MySQL database is also available for download.

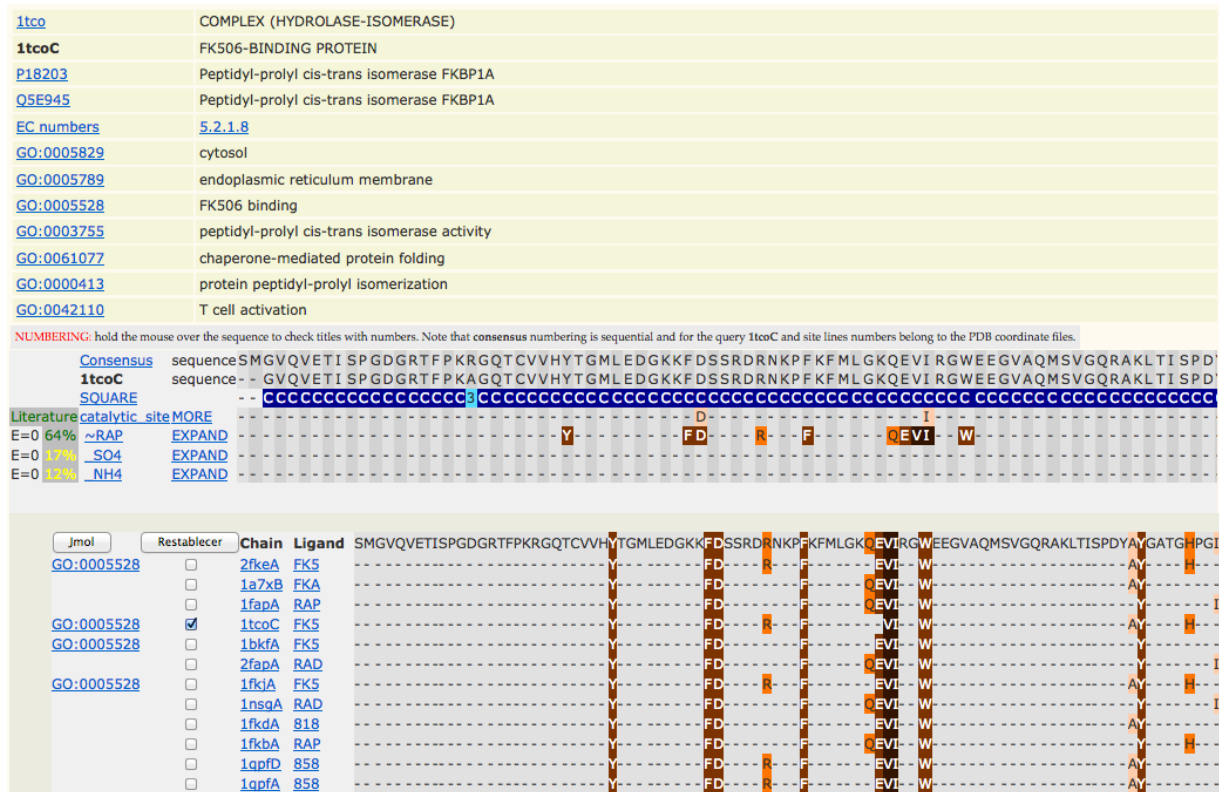


Figure 3 Web interface output for the chain C of the 1tco PDB code (Peptidyl-prolyl cis-trans isomerase FKBP1A). In the header information extracted from Uniprot, EC numbers and GO terms is included and linked to its source. In the middle section the entire protein sequence aligned to the cluster master sequence is shown. SQUARE conservation values are reported as colors (darker blue, higher conservation) and all the cluster master sequence catalytic sites (MSC) and binding sites (MSS) are listed, one per line. For the MSCs, the information support in the Catalytic Site Atlas is shown (PSIBLAST/Literature), while the MSS are accompanied by information on the number of evolutionarily related sites (marked as “E=”), the percentage ligand occupancy of the cluster binding site and a linked list of bound compounds information. Clicking on the EXPAND or MORE links gives access to the single sites collapsed to generate the MSSs or MSCs (shown in the bottom section)

3.8 *firestar*

*firestar*¹³² was developed as an automatic system for the transference of functional information from FireDB. The input is an amino acid sequence; a PSI-BLAST analysis is launched to detect near and remote homology with FireDB's Master Sequences (MS). Pairwise alignments between the target sequence and the FireDB templates are extracted and every aligned amino acid is assigned a conservation score from SQUARE (section 3.6.1). The results are presented to the user as shown in figure 4.

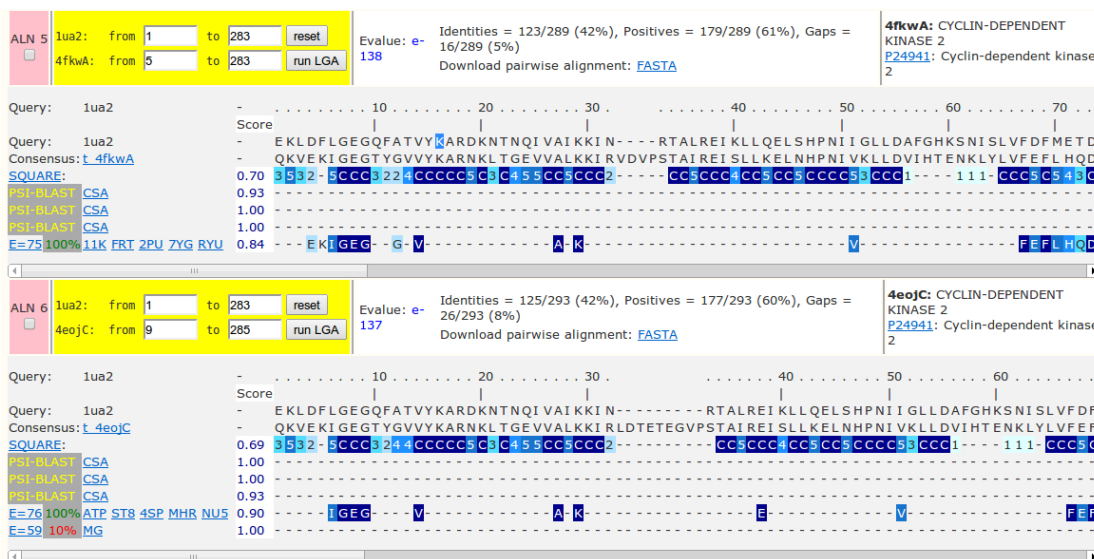


Figure 4 Output of the first version *firestar* for a pyruvate kinase (PDB: 1au2) showing alignments for two FireDB template clusters, 4fkwA and 4eojC. In the header (white background) information from the PSI-BLAST alignment. Below that, the pairwise alignment between the query and consensus MSS sequence. Finally, below the alignment, the SQUARE scores (the color code shows the conservation: the darker the blue, the higher is the conservation). There is a line for every binding site annotated in FireDB. On the left the evolutive related sites (E=), the percentage of the members of the cluster having the site occupied (where the font color represents the site score: red is 1, yellow is 2 and green is 3) and the compounds found. Annotated binding residues are reported with their relative SQUARE score for each binding site line.

The first version of *firestar* was designed to help manual annotators, automating the work of searching for PDB homologs with functional information, generating alignments for each one and evaluating conservation of key residues.

4 RESULTS

FireDB

4.1 Compound Annotation

Ligands are fundamental to the construction of FireDB database, since binding sites composition is determined by the spatial proximity of the nearest amino acids (based on an established cut-off. See section 3.7, Materials and Methods). But the database stored almost no additional information about these molecules, except for the structural and chemical features contained in the mmCIF libraries. Ligand-associated annotation is useful for a number of reasons. For FireDB the most interesting are functional characterization of a protein (misidentification of the natural ligand of the protein can lead to incorrect predictions) and the assessment of biological relevance of a binding site. Further annotation is also interesting for purposes that go beyond the main scope of the database. In this section we describe how FireDB ligands have been classified.

4.1.1 The biological relevance of compound

Biologically relevant ligands are defined here as the natural binding partners of a protein, important to its function. The PDB contains a diverse range of bound ligands, and the vast majority of them are not the natural ligands. As mentioned before, a primary raw classification, in many cases not even sufficient to guess their chemical nature, can be found in the mmCIF libraries (table 2).

Class	Description	Members
ATOMN	Nucleotides and products	548
ATOMP	Amino acids and products	1,086
ATOMS	Saccharides and products	462
HETAC	Co-enzymes	31
HETAD	Drugs	226
HETAI	Ions	117
HETAIN	Substrates, cofactors, inhibitors Non canonical compounds	13,885
HETAS	Solvents	5
HETIC	Ions with coordinated bonds	38
?	Undefined	263
Total		16,661

Table 2 Class name and descriptors for PDB compounds with their absolute frequencies extracted from the FireDB, version of the 22nd of August 2013.

Although some class names, such as co-enzymes, suggest possible biological relevance of their members, it is not possible to establish a direct association. This makes mmCIF classification useless for our purposes.

The first version of FireDB assessed biological relevance indirectly: a ligand is likely to be biologically relevant if it is frequently found occupying conserved binding sites (more details in section 4.2). While this simple rule works well for highly represented natural ligands like ATP or FAD, which often occupy conserved sites (in 81% and 94% of the cases, respectively), it fails for others. One good example is analogs, molecules used to “freeze” the protein in a certain state for crystallographic purposes: they are often found in biologically relevant sites (but not binding all the biological relevant residues) and might be mistaken for biological compounds. A further problem is that certain protein families in the PDB are under-represented and biological ligands that are limited to these families will be less likely to be classified correctly.

In order to deal with these limitations, we decided to carry out an extensive manual re-annotation of the PDB compounds. We first established a strict criterion for a chemical compound to be tagged as biological relevant (here named COGNATE):

In FireDB a ligand is tagged as COGNATE if it appears to be the real natural-binding partner in PDB structures (more than 80% of them) and it is important to the function.

We used as starting point the automatic compound classification as generated by the first version of FireDB, and we carried out a laborious manual validation of the biological importance of each ligand in the list. After this step, we started the re-annotation of compounds originally tagged as NON COGNATE (more than 95%), according to their ascending occurrence in the PDB database, since lower occurrence allows a faster assessment and probably hides novel biologically relevant compounds.

So far 664 of the 16.661 compounds in FireDB (22nd of August 2013 version) have been annotated as COGNATE (approximately of 40% curated ligands). The annotation is obviously a continuous process that goes hand in hand with the growth of the PDB database since new compounds are always added and compounds that may have been crystallised as NON COGNATE can later crystallise with their natural biological partners. We developed a strategy to spot new potential COGNATE ligand candidates to be checked, that will be discussed in section 4.1.6.

The complete set of manually curated COGNATE ligands constitutes, to the best of our knowledge, the largest and most reliable list of annotated biologically relevant molecules in the PDB. The complete collection of FireDB annotated compounds is freely available at firedb.bioinfo.cnio.es/Php/biologicalreference/index.html, where it can be searched via web or downloaded as plain text.

4.1.2 Ambiguous compounds

During the curation process, we also found a number of interesting cases where a molecule was present as the natural partner in some template structures, while it was non-cognate in others. Sucrose (PDB code: SUC) can serve as a representative case (see figure 5).

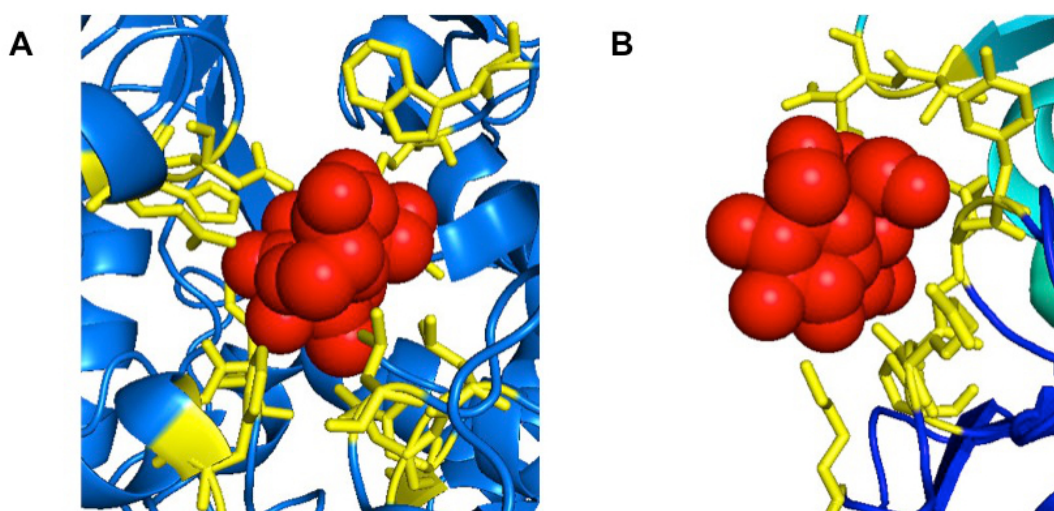


Figure 5 **A)** Sucrose molecule bound to the catalytic site of the A domain of a *Lactobacillus reuteri* glucansucrase (PDB 3HZ3). **B)** Sucrose bound to a superficial cleft of a tRNA-splicing ligase RtcB. In both figures sucrose is represented in red spheres, while the protein chains in cartoon, blue color. Residues within the 4 Angstroms distance from the sucrose are colored in yellow, stick representation.

Sucrose is a disaccharide, composed by glucose and fructose, commonly involved in reactions of the energy metabolism, and it is present in more than 190 PDB structures. For instance in one entry (PDB code 3HZ3) sucrose is complexed with a bacterial glucansucrase. Glucansucrases are large enzymes belonging to glycoside hydrolase family 70, which catalyze the cleavage of sucrose into fructose and glucose¹⁶¹, with the concomitant transfer of the glucose residue to a growing α -glucan polymer. In this protein family, sucrose is the main substrate of these enzymes, and consequently in FireDB the molecule should be labeled as COGNATE. In another structure (PDB code 4DWR) sucrose binds a 3'-phosphate RNA-splicing ligase. These proteins catalyze a GTP/Mn²⁺ dependent reaction that joins 2 RNA strand ends. In this case the PDB entry associated article⁸, explicitly cites sucrose as a buffer and a cryo-protectant and it should be tagged as NON-COGNATE. We estimated that sucrose should be tagged as COGNATE in almost 55% of the cases.

We generated an additional class ("AMBIGUOUS") to mark those natural compounds present in the PDB that often do not have a COGNATE role. This is a warning to FireDB users to further investigate the role of these compounds and the presence of this tag has some implications for the scoring of the predicted binding sites reliability, as illustrated in section 4.2.1. 56 compounds have been tagged as AMBIGUOUS.

4.1.3 Metallic compounds

Metal elements from the periodic table are highly represented in PDB structures, so they constitute a relevant group inside FireDB. As charged ions or in their oxidized form, they often fulfill a biological role. For example zinc ions are fundamental for the stabilization of the fold of the zinc finger structural domain. Members of the DNase I like superfamily use the magnesium in the binding site to catalyze the cleavage of DNA filaments. The same proteins make use of two calcium cations to stabilize the enzyme structure.

Metallic compounds share common features like reduced size, high conservation and a net charge¹⁶². To study the implications of these characteristics into binding site composition, we performed a global comparison of the metal binding sites tagged as biologically relevant in FireDB. A total of 15,542 metal binding sites were extracted, binding to a total of 52,050 amino acids. The global average per-site size is 4 amino acids, with a standard deviation of 1.12. In contrast non-metal binding sites have an average size of 13 residues, and a standard deviation of 7.15.

We looked at the amino acid distribution differences of metal binding sites with the rest (figure 6). While the amino acid composition of non-metal binding sites is similar to the background distribution of the PDB, metal binding sites are clearly enriched in 4 specific amino acids (cysteine, glutamate, histidine and aspartate). In fact, metal binding sites are depleted in all the rest of the amino acids except asparagine.

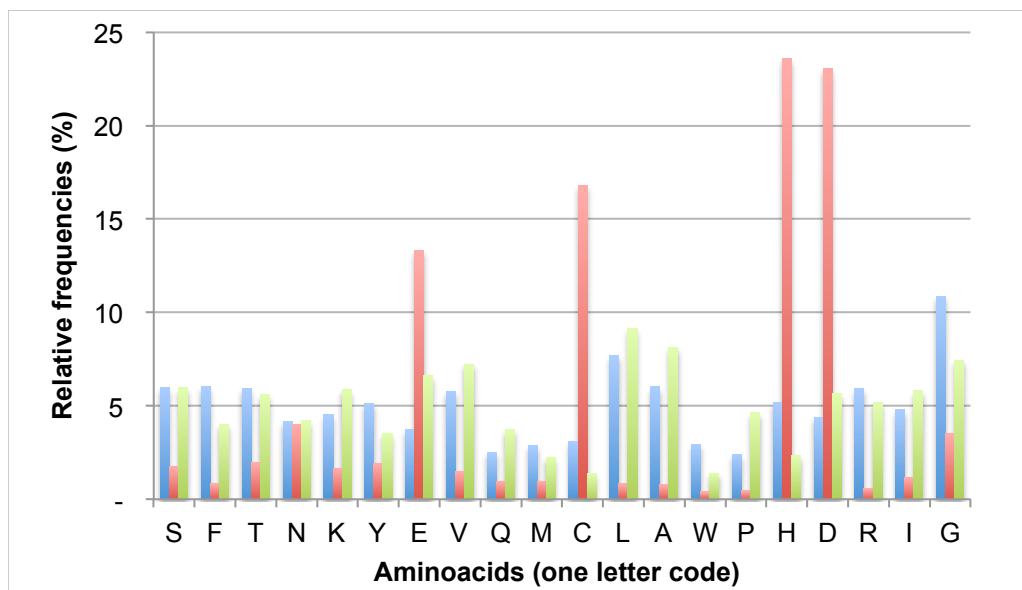


Figure 6 Relative frequencies for all natural amino acids: metal binding sites are in red and non-metal in blue. The green columns represent the overall amino acid distribution in the PDB. Data calculated over the subset of binding sites considered as biologically relevant by FireDB.

4.1.4 Metal binding site conservation

We also investigated the conservation of the amino acid composition in homologous metal-binding sites across FireDB.

As an illustrative example of this evaluation we can look at the binding site homologs comparison for a MSS site annotated from a zinc finger cluster (representative PDB code 1P47, figure 7). The same 4 residues are conserved in 94 of 102 templates (not all shown in the figure), even when the overall identity of the aligned sequence falls to below 25-30%. Among sites with at least two homologs in the PDB, the amino acid composition is conserved in more than 75% metal binding sites.

We created a new tag (“METAL_TAG”) to identify compounds that contain exclusively metal atoms (organometallic compounds have been excluded). We assigned this tag to 31 compounds in FireDB. The complete list is available to search or download at firedb.bioinfo.cnio.es/Php/biologicalreference/index.html.

After considering the results obtained in the comparison, we decided to modify the building schema of FireDB by separating the collapsing of metal binding sites (section 3.7, Materials and Methods) from the rest. In part because overlap between metal and non-metal binding sites is a frequent event (e.g: ATP and Mg^{2+} in ATPases) and in part due to their size: metal sites often “disappear” when they overlap with non-metal binding sites, and this event can mask their aforementioned features. Furthermore this allowed us to integrate a module that exploits the typical size and amino acid composition of metal binding sites into *firestar* to filter out possible false positive candidate sites, as it will be detailed in section 4.5.2.

Cluster Info	% seq Id	Residues
t_1p47A	clst 100	C C H H
t_2jp9A	clst 65	C C H H
t_2ee8A	clst 38	C C H H
t_2ee8A	clst 44	C C H H
t_1x6eA	clst 43	C C H H
t_2cshA	clst 31	C C H H
t_2lceA	clst 40	C C H H
t_2cotA	clst 41	C C H H
t_2m9aA	clst 39	C C H H
t_2ma7A	clst 35	C C H H
t_1un6B	clst 30	C C H H
t_2dlkA	clst 32	C C H H
t_2eozA	clst 34	C C H H
t_2ep3A	clst 37	C C H H
t_2emaA	clst 37	C C H H
t_2rpcA	clst 44	C C H H
t_2yt9A	clst 35	C C H H
t_2dlqA	clst 28	C C H H
t_1a1hA	clst 48	C C H H
t_1a1fA	clst 48	C C H H
t_2jp9A	clst 48	C C H H
t_1llmC	clst 46	C C H H
t_2epaA	clst 44	C C H H
t_2adrA	clst 45	C C H H
t_2rsiA	clst 25	C C H H

Figure 7 Homologous site comparison for one of the zinc binding sites of 1P47 (first line). The first column shows the FireDB cluster names, the second the bound ligand while the third the percentage sequence identities between the cluster and the 1P47 master sequence. The conservation of each amino acid residues of the 1P47 master sequence versus its aligned homologs is represented using a color code: the darker the blue, the higher the conservation.

4.1.5 Compound cross-references

Biologically (or potentially biological) relevant molecules are of high importance for the purposes of the database, though they represent just a small fraction of all compounds contained in the PDB (720 out of 16,661 compounds). The NON COGNATE group of molecules in FireDB is large and heterogeneous: it contains solvents, cryoprotectants and reagent molecules that are part of the crystallization solutions and that have little or no interest for FireDB purposes. But there are other classes such as analogs, antagonists and inhibitors that are nonetheless biologically interesting.

Annotation of these compounds can be useful for two main reasons. Firstly binding information extracted from these sites can still be informative, but the user has to be aware that the native site composition could be different, since these molecules are not the biological partners of the proteins they are in contact with. Second, information about the activity of these non-natural compounds may also be interesting for further studies.

We decided to retrieve information from 8 chemical and biological databases, bearing in mind three basic priorities:

- Availability: the data had to be freely available and the cross matching should be fully automatable;
- Coverage: the goal is to map information to as many PDB compounds as possible;
- Focus: we were interested in information relating to bioactivity but also in the biological relevance in order to expand the annotation of COGNATE ligands.

Three databases (KEGG compound, MetaCyc and ChEBI) were selected because they are focused on biologically relevant compounds. The other five (KEGG drug, DrugBank, ChEMBL, PubChem and PharmGKB) are large databases centered on bioactive molecules. The specific features of each one are illustrated in section 3.4.3, Materials and Methods.

4.1.6 Database mapping

We initially made use of the 1D molecular descriptors (rule-based strings that represent 3D molecules using plain text) generated by the PDB consortium. Isomeric SMILES and InChi codes (section 3.3, Materials and Methods) were directly mapped onto the PubChem database. Despite the huge size difference (PubChem contains more than 91 million compounds entries), we retrieved relatively few PDB hits (~20%). This result clearly showed that there were differences in the representation code used, so we decided to standardize the descriptors using OpenBabel, an open-source chemistry Toolbox. Although the final coverage increased to 30%, this was still a small fraction of the compounds present in the PDB.

Finally we decided to make use of pre-calculated mappings provided by the databases themselves. All databases we chose have at least one correspondence with the others, so we created a database specific vector for every compound. After that, we compared the vectors to find possible disagreements in ID associations and we manually resolved discrepancies.

We found cases where different IDs from the same database were associated to the same PDB ID, as can happen for different stereoisomers of the same molecule. The disambiguation was carried out by manual revision of the stereochemistry. For some databases we found novel PDB associations, using indirect information from a third database. We found 15 new correspondences for KEGG database and 70 for PharmGKB. The total match with PDB ligands and the overlap with the COGNATE/AMBIGUOUS list, as well as overlaps between databases are shown in table 3:

	PubChem	KEGG compound	KEGG drug	ChEMBL	ChEBI	DrugBank	MetaCyc	PharmGKB	COGNATE & AMBIGUOUS
PubChem	15576								
KEGG compound	1961	1962							
KEGG drug	665	443	665						
ChEMBL	6159	1283	621	6200					
ChEBI	2165	1427	461	1330	2172				
DrugBank	5257	1126	507	2658	1195	5305			
MetaCyc	730	477	134	429	459	337	730		
PharmGKB	362	293	311	342	314	359	71	362	
COGNATE & AMBIGUOUS	673	431	106	285	409	395	146	63	720

Table 3 Total number of PDB compounds mapped onto the eight different databases (shown in blue) and the overlap between databases. The overlap with the subset of COGNATE/AMBIGUOUS FireDB ligands is also shown. Concurrence between compounds tagged as biologically relevant in FireDB and compounds in databases of mainly natural occurring molecules are highlighted in green.

In total we were able to map 15,658 PDB compounds to at least one database, almost 94% of the entire dataset. PubChem provided the best coverage (99.5% of the mapped compounds), while we retrieved the lowest number of correspondences from the pharmacological specialized database PharmGKB (2.3%).

This mapping process allowed us to automatically generate a non-redundant list of PDB compounds mapping onto ChEBI, KEGG COMPOUND and MetaCyc. These three databases gather specific information about biologically relevant compounds, so in this way we created a priority list of 2,416 compounds for spotting new possible COGNATE compounds through manual curation. Furthermore this approach provides the basis for automatic selection of the candidates whenever the database is updated.

All this information has been integrated in two new FireDB tables.

4.1.7 Bio-activity annotations

5 of the 8 selected databases provide short descriptions of the biological activity of the compounds. This information is useful to spot protein modulators, agonists, analogs and other different molecules. KEGG, ChEMBL, ChEBI and PubChem annotate structured phrases or tags that define a category. These are associated to multiple compounds. Drugbank provides extensive annotations about the specific pharmacology of the molecule, which is extracted from literature and from drug company indications.

	PubChem	KEGG DRUG	ChEMBL	ChEBI	DrugBank	COGNATE & AMBIGUOUS
PubChem	1001 (6,4%)					
KEGG DRUG	407	608 (91,4%)				
ChEMBL	257	272	296 (4,8%)			
ChEBI	485	382	237	1119 (51,5%)		
DrugBank	348	351	264	323	506 (9,5%)	
COGNATE & AMBIGUOUS	119	91	15	245	75	720

Table 4 Final count of annotated bioactive PDB compounds in drug oriented databases; and their mutual overlap. Blue boxes show the number of annotated bioactive compounds and in round brackets their ratio in comparison with all the chemicals retrieved for that particular database. Concurrence with biologically relevant compounds in FireDB is highlighted in the green boxes.

PubChem unique annotations -> 295	Freq.
Compounds or agents that combine with an enzyme in such a manner as to prevent the normal substrate-enzyme combination and the catalytic reaction	165
Substances that inhibit or prevent the proliferation of NEOPLASMS	81
Substances that reduce the growth or reproduction of BACTERIA	70
Agents used in the prophylaxis or therapy of VIRUS DISEASES. Some of the ways they may act include preventing viral replication by inhibiting viral DNA polymerase; binding to specific cell-surface receptors and inhibiting viral penetration or uncoating; inhibiting viral protein synthesis; or blocking late stages of virus assembly	35
KEGG DRUG unique annotations -> 367	
Antineoplastic	54
Antibacterial	34
Antiviral	25
Amino acid	16
ChEMBL unique annotations -> 198	
DNA inhibitor	18
Bacterial 70S ribosome inhibitor	13
Cyclooxygenase inhibitor	12
Bacterial penicillin-binding protein inhibitor	11
ChEBI unique annotations -> 525	
human metabolite	288
Metabolite	198
antineoplastic agent	104
plant metabolite	84

Table 5 Overview of the four most frequent bioactivity terms for PubChem, ChEMBL, ChEBI and KEGG DRUG. This information is extracted from these 4 databases and stored in FireDB. Drugbank is not included here since every annotation is unique.

Table 4 details the coverage among databases with available information. In total 1,824 unique compounds have been annotated, and the ratio between annotated/un-annotated entries is almost 1:9. We retrieved this information when it was directly associated with our matched compound, but we also extracted and properly tagged

stereoisomer annotations, when available, in order to improve global coverage. ChEBI provides largest set of annotations (1,119 molecules). Table 5 shows the most frequent tags and descriptors retrieved from each consulted database. Drugbank is not included since every descriptor is unique.

Additionally, we started a manual annotation process for PDB molecules with unique or no external database references, mining the structure-associated papers. We chose first compounds with low occurrence in the PDB, to speed up the information retrieval. We extracted a single description phrase for every recorded bioactivity (so one compound can be associated to more than one bioactivity), the organism where the effect was observed and the reference of the source from where we obtained the information (doi or PubMed id). So, for instance, we annotated compound 3SZ with “*pyruvate kinase M2 (PKM2) activator*”, “*Homo sapiens*” and “<http://dx.doi.org/10.1038/nchembio.1060>” respectively. So far, 326 additional molecules have been manually curated, raising the total annotated compound to 2,150. All this information has been integrated into FireDB.

4.2 Binding sites biological relevance in FireDB

In the last years, PDB entries, FireDB master sequences (MS) and the related master sequence (consensus) binding sites (MSS) have been growing at a comparable pace (figure 8). In FireDB version August 22nd 2013, the number of total binding sites extracted from crystal structure coordinates is almost 500,000, these are finally collapsed into 116,514 MSS.

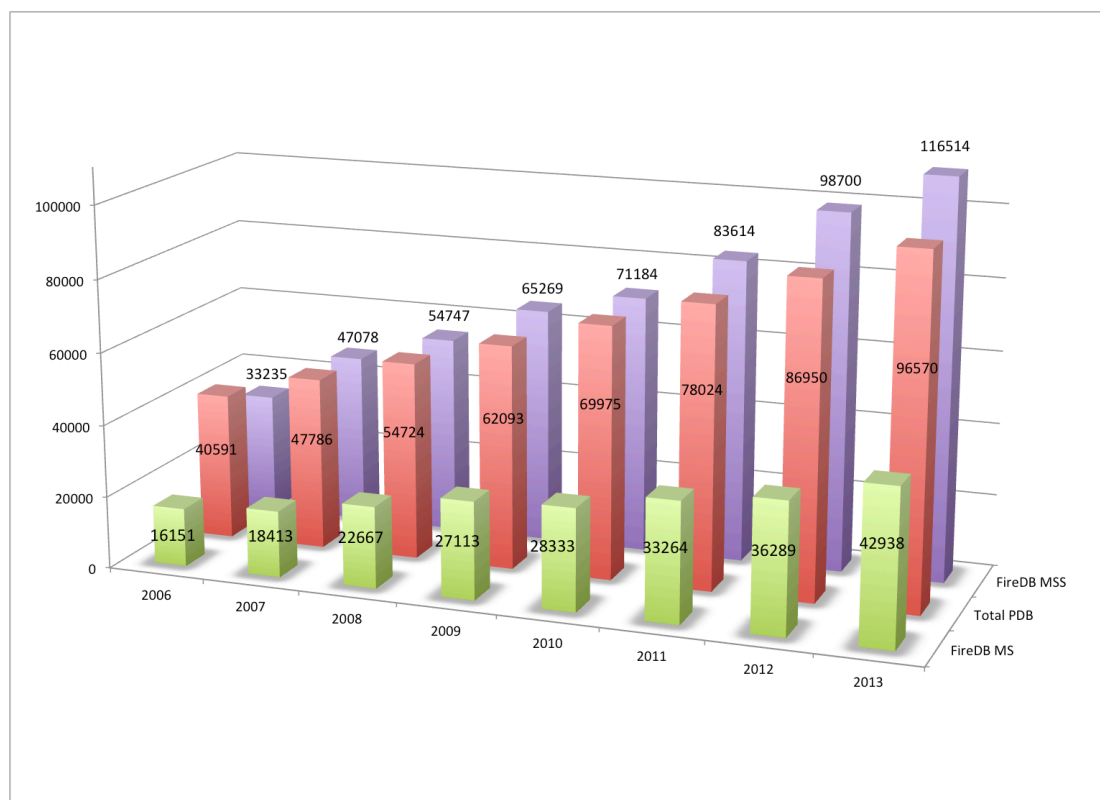


Figure 8 Growth of FireDB Master Sequences (MS, in green) and Master Sequences binding Sites (MSS, in purple) in comparison with the PDB entries growth (in red) since FireDB was first released.

These numbers illustrate how the growing number of structurally solved binding sites requires an automated protocol for relevance. For this reason first version of FireDB

provided an accurate protocol for dealing with this scenario; manual curation is still advisable for small scale analyses, for which FireDB provides a suitable reference annotation. The biological relevance of sites was evaluated automatically using various criteria (see table 6).

Type	Description
Evolutionary	Relative binding site conservation (in comparison with the rest of the sequence)
Evolutionary	Co-occurrence of the binding site in homologous PDB proteins
Redundancy	Percentage of times the site is occupied in the FireDB cluster
Redundancy	Tendency of the bound ligand to occupy conserved sites in the PDB
Structural	Mean size of the binding site for the bound ligand in the whole PDB

Table 6 Criteria used for the evaluation of the biological relevance of binding sites in the first version of FireDB.

One is based on the relative binding residues conservation respect to the rest: binding residues are positively selected and their conservation is usually higher compared with other amino acids in the same protein.

Another one assumes that sites with many homologs in the PDB are more likely to be biologically relevant. The evaluation process looks at the number of hits found in the all-against-all PSI-BLAST analysis of the collapsed Master Sequence binding Sites (MSS).

Another one is based on the hypothesis that unspecific interactions can arise by chance, but the probability of observing them in all the members of a FireDB cluster is low.

The last two criteria focus on the ligand. If the compound usually binds conserved sites and the evaluated pocket size is comparable with the mean size for this specific ligand, the site is more likely to be biological relevant.

The combination of the different parameters allows FireDB to classify binding sites in three levels of relevance:

- Spurious or artifact sites: score 1
- Putative relevant sites: score 2
- Biologically relevant sites: score 3

4.2.1 Improvements in biological relevance assessment

We detected three different situations in which the previous automatic assessment returns incorrect classifications. Firstly when a new binding site is annotated, with few or no homologs, it is usually tagged as an artifact since the first and the second criteria cannot be considered. To understand the extent of this scenario, we have to bear in mind that in FireDB version August 22nd 2013 there were 51,618 binding sites with no homologs (44%). We faced the same situation when a novel cognate compound is crystallised with the protein. In FireDB 11,885 compounds (71%) appear in 2 or fewer sites, so misleading or no information is available for the last 2 criteria of table 6. Finally overrepresented clusters or small sites can be erroneously tagged as biologically relevant due to their overlap with many different sites.

PDB binding site information (figure 8) is growing and in principle it should be possible to find one or more homologs for the majority of the biologically relevant annotated sites. Simple sequence overlap does not imply homology, since the function relies on binding residues. So the conservation of these positions over the different aligned homologs should be interpreted as a proof of biological relevance.

We overhauled the previous classification and established new criteria to assess the sites, based on:

- Site size
- Ligand biological relevance
- Amino acid composition
- Amino acid conservation in homologous sites.

Following this protocol, we discarded *a priori* sites with a size smaller than 3, since they can give non-specific conservation signals. Then for every site we evaluated the information from aligned homologs stored in FireDB. Whereas before we were taking into account the raw total number, now we are selecting features considered relevant for the assessment: only alignments binding the same type of ligand (metal or non metal) and containing a COGNATE or AMBIGUOUS compound are valid. Alignment percentage identity is also evaluated: close homologs (more than 80% identity) are excluded and remote homologs have a higher weight. Finally we evaluate the single residue conservation, using SQUARE reliability scores. Those that pass a conserved residue cut-off (site type and size dependent) are considered supporting. Finally a site is tagged as “RELEVANT” if it has more than 50% of supporting sites.

When no homologs from the PDB are detected, we only take into account the biological relevance of the ligand: if a site contains a COGNATE compound(s) it is tagged as “NOVEL”. All the remaining sites are tagged as “NOT SUPPORTED”.

For some binding sites (for instance metals) we have assigned specific constraints that need to be satisfied, since their binding sites features have been already characterized. So far we introduced ligand-based rules for zinc and for some calcium binding sites (EF-hand), after a literature review of binding site architecture for both metals^{163,164}. Zinc binding sites have to contain cysteine, histidine or aspartic acid, while calcium sites have to contain asparagine, serine, threonine, aspartic or glutamic acid. In both cases these amino acids must represent more than 75% of the site.

As an example of how these two different strategies perform, we present here the homologous site comparison for the sodium binding MSS of the cluster **1b57A** (figure 9).

t_1b57A	clst 100	V H G G G S	6	1.00	Expand NA
t_3eklA	clst 38	V H G G G S	7	1.00	Expand NA
t_4delA	clst 38	V H G G G S	6	1.00	Expand NA
t_3q94A	clst 23	V H G G G T	7	0.98	Expand NA
t_1qvfB	clst 22	A H G G A S	7	0.95	Expand NA
t_3eklA	clst 38	- H G G G S	15	0.83	Expand SO4 13P TD4 2FP
t_4delA	clst 38	- H G G G S	14	0.83	Expand 13P PGH
t_3q94A	clst 23	- H G G G T	14	0.82	Expand 13P
t_2fjkA	clst 21	- H G G A S	14	0.81	Expand 13P
t_1rv8D	clst 21	- H G G A S	10	0.81	Expand SO4
t_3n9sA	clst 21	- H G G A S	14	0.81	Expand TD4
t_3c56A	clst 21	- H G G A S	16	0.81	Expand PGH PH4 TD4
t_3gayA	clst 20	- H G G S S	17	0.81	Expand HDX P6T P6F
t_2isvB	clst 19	- H G G S S	13	0.81	Expand PGH SO4
t_3qm3A	clst 61	- - - G G S	7	0.50	Expand SO4
t_4a21A	clst 34	- - - G G S	7	0.50	Expand SO4
t_3gakA	clst 18	- - - G S S	6	0.47	Expand SO4

Figure 9 Representation of NA (sodium) binding MSS site homologues -FireDB cluster **1b57A**-. In the third column the percentage of identity between the two MS is reported, and after this the aligned sites. Conservation is represented by the background color: the darker the blue, the more conserved. The fifth column shows the original size of the sites, the sixth the mean SQUARE score of the overlapped residues and the final column the compounds bound in the cluster. Note that not all homologous sites bind sodium.

This cluster contains 3 chains, 2 from PDB 1B57 and one from PDB 1ZEN. Both are class II Fructose-Biphosphate aldolases, an enzyme that catalyzes the aldol condensation of a ketose and an aldose to form fructose 1,6-bisphosphate. The enzyme is also able to catalyze the reverse reaction. In their work⁹, the authors explain that two ions (present in the crystal) are required in the binding site, a catalytic zinc and a sodium; both of them are needed to activate the enzyme. The Na²⁺ coordinates the substrate inside the binding site, stabilizing the intermediate states.

The sodium binding site was classified as an artifact in FireDB according to the previous classification algorithm (score number 1). Homologous sites analysis shows that it overlaps with other sodium binding pockets. Looking at the sequence conservation, we observe how the amino acid positions that bind sodium are almost entirely conserved even in remote homologues (22% and 23% identity). New classification schema tags this site as "RELEVANT".

All sites have been re-evaluated with our new protocol and the overlap between the previous and new biologically relevant dataset is presented in figure 10:

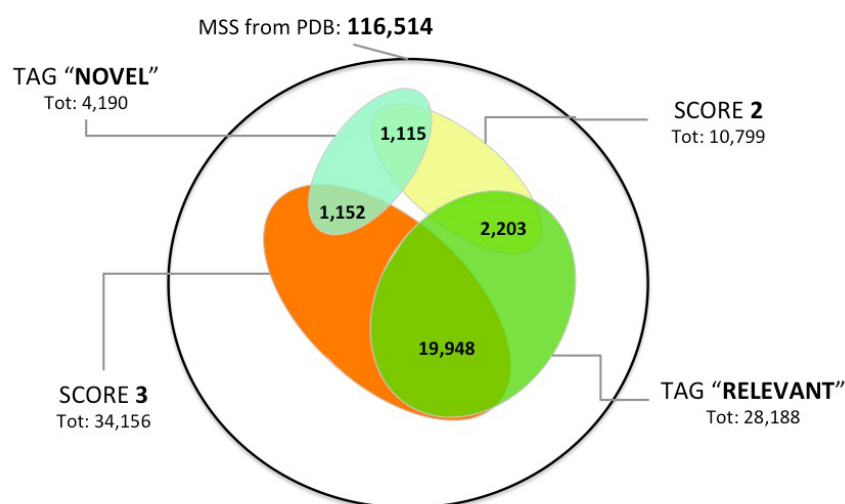


Figure 10 Total number of the initial and new biological relevance evaluation of whole FireDB binding site dataset. Size of the overlaps is also reported.

RELEVANT and NOVEL tagged sites overlap as expected with score 2 and 3 (respectively putative biological and biologically relevant sites). But they also rescue other sites, previously tagged as artifact: 6,037 new sites are now classified as RELEVANT and 1,923 as NOVEL. On the other hand, there are 13,056 and 7,481 sites tagged with scores 3 or 2 respectively that are now classified as NOT SUPPORTED.

We manually inspected some of those cases where the two protocols do not agree, in order to understand the reasons behind discrepancy. For those sites newly classified as "RELEVANT" with a previous score of 1, the new method was able to use evolutionary information to correctly detect biological relevance in many sites, as we previously illustrated (figure 9). Additionally we spotted complex cases that are still problematic: for example small sites that overlap with larger biologically relevant binding sites, and that would require manual validation to understand whether they really are biological significant.

We also looked at sites that had obtained a score of 3 under the old method but that the new method classified as "NOT SUPPORTED" (orange in the figure 10). The new protocol discards more than 55% of these MSSs because of their size (smaller than 3 residues). Among the rest, 9% contains NON COGNATE tagged compound(s) and have

no homologues; the remaining 41% have few supporting alignments, but may be biologically interesting.

To illustrate the differences in results between these two protocols we present here two particular cases. The first is a case where initial classification system works better, the heme binding site of leghaemoglobin of yellow lupin (PDB code: 1GDI). Figure 11 shows how the conservation of the site is detectable, but it fades in homologues below 40% identity (probably due to the size of the site) even though every MSS is binding exactly the same COGNATE compound. This site does not pass the filter due to the residue conservation and coverage cut-offs, new algorithm classifies it as NOT SUPPORTED.

t_1gdIA	clst 100	L F S F H K V L L H K V H F V Y L I	18	1.00	Expand	HEM
t_3qqrB	clst 43	L F S Y H T V M I H T V H F T Y L I	21	0.92	Expand	HEM
t_1binA	clst 57	L F S - H K L L L H - V Q F V Y L I	21	0.85	Expand	HEM
t_2r50C	clst 44	M F - F H - V M L H Y V H F T Y L I	20	0.83	Expand	HEM
t_3qqqA	clst 44	L F - Y H T V M L H N V H F T Y L -	19	0.82	Expand	HEM
t_3zhwA	clst 45	M F - - H S V M L H Y V H F A Y L I	21	0.80	Expand	HEM
t_2oifG	clst 41	M F P F H - V M L H Y V H F T Y L I	19	0.79	Expand	HEM
t_1d8uA	clst 45	M F - - H S V M L H Y V H F V Y L -	20	0.77	Expand	HEM
t_3wctF	clst 24	M F - R H R V G L H R V Q Y V - - -	18	0.68	Expand	HEM
t_4b4yA	clst 25	I F P W H S I V - H - V L F L - I -	18	0.66	Expand	HEM
t_4o35A	clst 20	L F - Y H K V V L H V V S F V Y V M	23	0.62	Expand	HEM
t_1x9fL	clst 18	P F - R H R V - L H R L F F F - - -	18	0.59	Expand	HEM
t_1oj6B	clst 21	L F - Y H K V V L H V V S F V - - -	19	0.56	Expand	HEM
t_4u8uL	clst 23	L F - - H R - - - H R I F Y F F I I	17	0.53	Expand	HEM
t_2bk9A	clst 21	K F P - H R I V I H R V S Y L - - -	19	0.52	Expand	HEM
t_2d2mB	clst 25	L F - R H R V G L H R I G Y F - - -	20	0.510	Expand	HEM
t_3s1IA	clst 25	L F - - Q K L V L H Y V H Y V Y A M	20	0.49	Expand	HEM
t_1h1bA	clst 21	K F - - H R - - L H N V H - F - - -	14	0.45	Expand	HEM
t_4hrrG	clst 16	Y F - R H T L A F H R I A F I - - -	19	0.45	Expand	HEM
t_4q1bB	clst 23	I F N - Q A L T I H L I H Y V Y I F	24	0.44	Expand	HEM
t_2r80B	clst 18	F F - - H K V S L H L V N F L - - L	19	0.44	Expand	HEM
t_3a59F	clst 16	F F - - H K V S L H L V N F L - - L	19	0.44	Expand	HEM
t_1x9fG	clst 21	L F - - H R I G L H - V H F F - - -	20	0.43	Expand	HEM
t_1h1mA	clst 18	K F P - - V - L H N V N Y F - - -	13	0.42	Expand	HEM

Figure 11 Representation of HEM (protoporphyrin IX containing iron) binding MSS site homologs -FireDB cluster 1gdIA-. Conservation is represented by the background color: the darker the blue, the more conserved.

The second shows a site where new classification works better, the NAG binding site of receptor that regulates, in response to brassinosteroid binding, a signaling cascade involved in plant development (PDB code 3RGZ). NAG (N-acetyl-D-glucosamine) is a monosaccharide, highly represented in the PDB (more than 5,600 entries) since it is often included in crystallization mix, such as in this case. In figure 12 we can observe how the conservation of the site is poor through different homologues (all NAG-binding): for this reason new protocol correctly classifies it as NOT SUPPORTED. Redundancy of sites and high NAG ligand representation in PDB causes the assignment of score 3, here and in another 2,191 NAG-binding MSSs. The new algorithm assigns the “NOT SUPPORTED” tag to 2,134 of them (more than 97%).

t_3rqzA	clst	100	S	S	R	F	N	S	V	D	H	9	1.00	Expand	NAG
t_3t6qB	clst	20	K	S	H	-	N	S	L	D	-	8	0.65	Expand	NAG
t_4j0mB	clst	38	-	S	S	-	N	S	-	-	-	6	0.42	Expand	NAG
t_3ojaB	clst	22	-	S	-	-	N	S	-	D	-	4	0.40	Expand	NAG
t_3b2dA	clst	24	S	S	H	L	D	-	-	-	-	8	0.40	Expand	NAG
t_3b2dA	clst	24	-	-	-	-	D	A	V	N	H	5	0.40	Expand	NAG
t_3v47B	clst	25	-	A	-	K	N	-	-	D	-	5	0.34	Expand	NAG
t_4tzhA	clst	25	-	S	-	-	H	R	-	D	-	4	0.33	Expand	CL
t_4hq1A	clst	19	-	H	-	-	-	R	V	N	-	5	0.28	Expand	NAG
t_2z7xB	clst	21	-	Q	-	-	-	S	-	N	V	7	0.28	Expand	NAG

Figure 12 Representation of NAG (*N*-acetyl-*D*-glucosamine) binding MSS site homologues - FireDB cluster 3rqzA-. Conservation is represented by the background color: the darker the blue, the more conserved.

The results obtained suggest that new method complements, more than replaces, the previous algorithm definition of biological relevance, since both are able to cover specific cases where the other fails. The combination of both of them allows to reach a total number of 52,915 sites assessed (non redundant sum of all the sites with score higher than 1 or non tagged as “NOT SUPPORTED”), a 7% more compared with the previous classification alone.

4.3 Final database schema and public accessibility

Along with the addition of new information and tables, we made several changes to simplify the old schema (see appendix figure 1 and 2). All tables related to site calculation cut-offs other than 0.5 Å (1 and 1.5 Å + Van der Waals radii, represented respectively with the suffix 40 and 45) have been removed, since in our experience the more restrictive distance cut-off is the most informative. BINDSITE tables have also been removed, because overlapping site compositions (two molecules in the same pocket) are no longer allowed to merge in the same MSS. The new compounds chemical information we have generated has also been made available through a tab alongside the FireDB and *firestar* web pages, at <http://firedb.bioinfo.cnio.es/Php/ligand/index.html>, making it easily accessible to a range of users. A snapshot of the new ligand information page is shown in figure 13:

Ligand Info

Home FireDB firestar SQUARE Ligand Info Help

Structural Biology and Biocomputing Programme CNIO

A

LIGAND 3 letter code (mmCIF format)

Enter keyword for ligand searches

B

String Search

1

ID: GFT

Common_name: (2S)-2-azanyl-3-[cyclohexyloxy(methyl)phosphoryl]oxypropanoic acid

Synonym: Cyclosarin bound Serine

FireDB_hits: 0

biological_tag: NON

metal_tag: NO

2

ID: MEL

Common_name: [((1R)-2-((2S)-2-((4-[AMINO(IMINO)METHYL]BENZYL)AMINO)CARBONYL)AZETIDINYL)-1-CYCLOHEXYL-2-OXOETHYL)AMINO]ACETIC ACID

Synonym: MELAGATRAN (ASTRA-ZENECA)

FireDB_hits: 3

biological_tag: NON

metal_tag: NO

3

ID: 024

Common_name: 4-BROMO-3-(CARBOXYMETHOXY)-5-[3-(CYCLOHEXYLAMINO)PHENYL]THIOPHENE-2-CARBOXYLIC ACID

C

PDB Data

Pdb Id: MEL

Chemistry type: NON-POLYMER

PDB Class: HETAIN

Mol. weight: 429.513

Synonym: MELAGATRAN (ASTRA-ZENECA)

Formula: C22 H31 N5 O4

Common name: [((1R)-2-((2S)-2-((4-[AMINO(IMINO)METHYL]BENZYL)AMINO)CARBO...
CYCLOHEXYL-2-OXOETHYL)AMINO]ACETIC ACID

ISO SMILE: NC(=N)c1ccc(CNC(=O)[C@@H]2CCN2C(=O)[C@H](NCC(=O)O)c2CCCCC2)cc1

FireDB Data

PDB structures: 3

biological tag: NON COGNATE

metal tag: NO

External Refs

Manual_Refs

BINDING_SITES (3)

Figure 13 Web page output to browse the chemical information stored for the Melagatran (MEL) ligand. A) Users can access the data using the PDB three letter code for the ligand or searching for a string present in the common name or synonym. B) Results of the string search generate a pop-up window to help users select relevant compounds. C) Visualization of the information: pull-down menus are generated that contain general information, external references, manual annotations (if any) and a binding site summary.

FireDB is highly accessed through its web page at <http://firedb.bioinfo.cnio.es/Php/FireDB.php> and according to CNIO official statistics in the last six months of 2013 was visited daily almost 50 times. In order to make FireDB easier to access, we implemented a RESTful service with which it is possible to connect directly to the database and to retrieve the relevant information in XML format. A model connection script is provided on the web page and all the options available are clearly explained in the code as well in the help pages. The current MySQL database and all the previous stable releases are also available at <http://firedb.bioinfo.cnio.es/repository/>.

firestar

4.4 Consensus Predictions

firestar is a method for the automatic transference of the functionally important residues in FireDB: both tools together provide a useful framework for assisting functional annotation by human experts.

The output of the original *firestar* web server was a single web page (figure 4, Materials and Methods) where all the aligned candidate binding sites (and related ligands) were presented along with the local conservation score, as calculated by SQUARE. However the increasing number of binding sites in the PDB and the high proportion of non-biologically relevant sites made this representation difficult to use for all but a handful of experts. For this reason we developed a new version of *firestar*, both for the web server and for the stand-alone algorithm that was able to evaluate and merge results from many templates into a single readable output.

To generate consensus predictions from the many homologous binding sites detected by *firestar*, a rule-based module was introduced. This was applied directly to the results of the first version of *firestar*, with the aim of automatizing the annotation protocol that would be followed by a human expert annotator. This algorithm:

- Evaluates the source information to discard non biologically relevant information;
- Evaluates every template and its binding information, selecting the conserved residues that are more likely to be functionally important in the target protein;
- Merges predictions from the separate MSS into a final consensus prediction(s);

4.4.1 Candidate search and filtering

In the initial version of *firestar* the first step to generate a prediction was a PSI-BLAST-based sequence similarity search of the target protein sequence against an *ad-hoc* created database containing Master Sequences (MS) stored in FireDB. *firestar* then selected just those containing a Master Sequence Site (MSS in section 3.7, Material an Methods) from among the aligned MS templates found in the first step. In the first output version, at this point the alignments were evaluated with SQUARE and all the results were presented in a unique page (figure 4, Materials and Methods).

The new version of *firestar* generates alignments with both PSI-BLAST and HHsearch (see section 4.5.1) and produces a consensus prediction starting from the MSS. *firestar* evaluate the candidate MSS detected by both search algorithms and discards non-relevant information: in figure 14 we illustrate the filtering process step by step.

For this, *firestar* relies on the assessment of biological relevance performed by FireDB (see section 4.2). In particular, *firestar* considers reliable only the information that comes from MSS tagged as biologically relevant, putative or novel (~45% of all the MSSs stored in FireDB version of 22nd August 2013). For catalytic site annotations, only human curated sites of CSA are considered.

Selected alignments are analyzed by SQUARE to assess the individual binding residue conservation. Using residue conservation scores, we:

- Exclude poorly conserved candidate sites: mean conservation score for the site has to be higher than the mean conservation score of the whole alignment.
- Exclude poorly conserved individual binding residues within each candidate site, discarding those below a pre-established SQUARE cut-off.

Finally, in order to assess its integrity, we compare the resulting candidate site size with its original size, as stored in FireDB. In this way we discard degenerated binding sites or isolated conserved residues, whose higher conservation might have arisen for reasons other than ligand-binding (e.g: structurally important residues). Candidate sites that are smaller than a manually established threshold (variable depending on the size of the original FireDB site) are discarded.

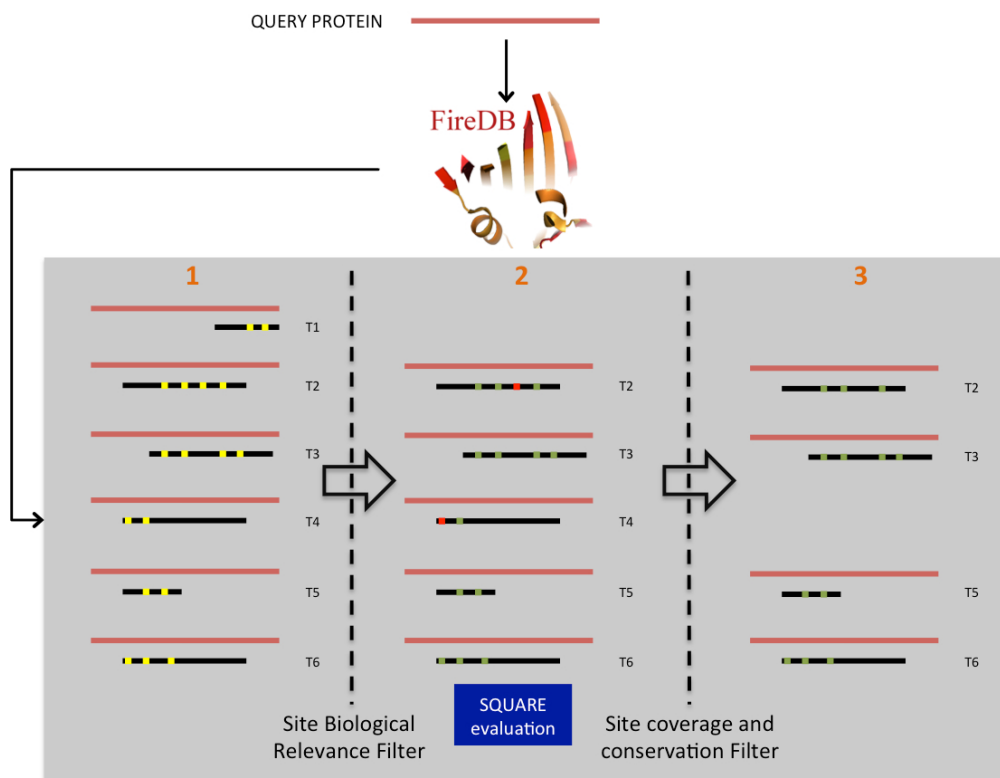


Figure 14 *firestar* pipeline to filter information extracted from FireDB. A PSI-BLAST sequence analysis with the query protein (represented by red line) is launched against Master Sequences (MSs) stored in FireDB. In panel 1 on the left we can see alignments (T1 to T6) containing candidate binding residues (yellow boxes) from the homologous Master Sequence Sites (MSSs, black lines). The first filter step is based on database biological relevance score: only predictions coming from sites tagged as reliable pass. After that SQUARE calculates conservation at the residue level for binding sites and other aligned residues (panel 2, here green squares correspond to conserved positions, red to poorly conserved) and sites that have low conservation scores compared with the rest of the alignment are filtered out in a second filter step. Finally those predictions where the conserved candidate site is much smaller than the original (\leq than 50%, for example T4 in the figure) are also excluded.

4.4.2 Candidate merge and generation of a consensus prediction

Once selected, candidate sites are merged to generate the final prediction (see figure 15).

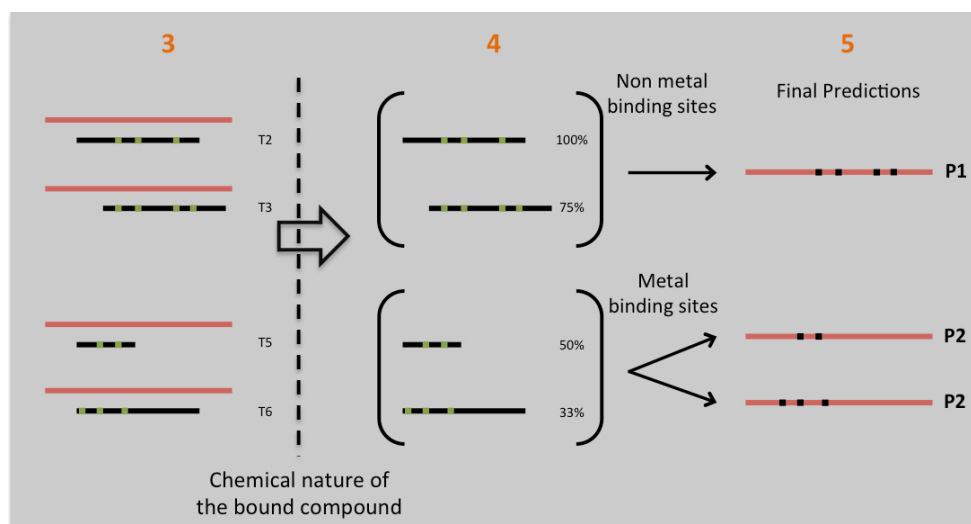


Figure 15 Panel 3 corresponds to panel 3 in the previous figure. All the information evaluated as reliable through the different filters is now collapsed to consensus prediction(s). The different candidate sites (black lines are the templates, green square are conserved candidate residues) are divided according to the chemical nature of the annotated bound compound. Then relative overlap is calculated (panel 4, percentages in brackets). In Panel 5 sites that present an overlap of 60% or higher are merged, and the predicted residues mapped onto the query protein (red lines). The different predicted candidate MSS are also grouped as pockets (P1, P2) if their overlap suggests that they could share the same 3D position in the protein (33% or higher).

In the first step of the merge *firestar* groups all the candidate sites depending on the type or the chemical nature of the bound compounds. So metal binding, non-metal binding or catalytic sites can only be merged with other sites of the same type.

Then the program evaluates overlap between individual predictions, marks those that will be merged and generates one (or more) final consensus predictions. The final amino acid composition of each predicted site is obtained from the combined composition of all sites that overlap over at least 60% of their respective size; the predicted ligand is the most frequent COGNATE (if any) ligand found in the templates. The output of our pipeline is now a more simple and informative representation of the predicted binding sites and this drastically improves the usability of the method.

4.4.3 New output web page

A new summary output web page has been generated to present the consensus predictions (see figure 16).

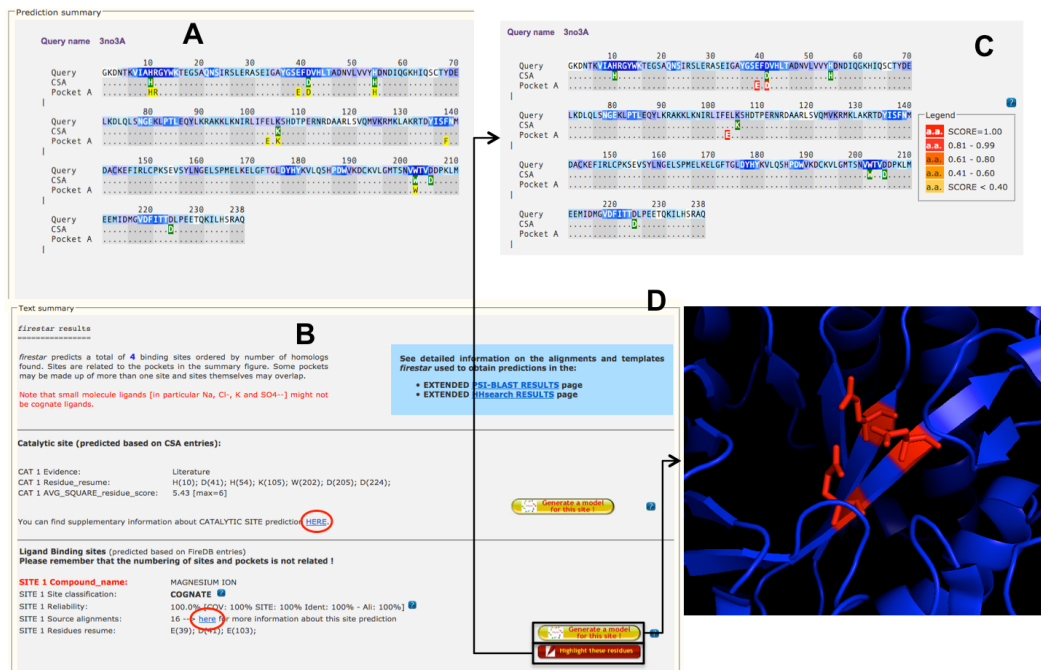


Figure 16 Summary output for 3NO3, chain A. **A)** Top: target sequence is reported with per position conservation scores (the darker the blue, the higher the conservation); catalytic and binding residues are highlighted respectively in green and yellow in separate lines. **B)** The details for each prediction are shown in text format. Additional information is accessible by clicking on related links (highlighted here by red circles). Through the links in the light-blue box it is possible to visualize the alignments generated in detail (original single-page firestar output). **C)** Relative frequencies of the predicted residues across the collapsed candidate sites can be highlighted. **D)** The “generate a model” yellow button allow the user to generate an automatic model of the site in PDB format.

The entire target sequence and its per-residue SQUARE conservation score is presented at the top of the page. Below the sequence, on different lines, predicted ligand binding and catalytic site residues are reported with colored residues. To make the visualization clearer (especially for proteins with multiple predictions) predictions that overlap over more than 40% of the alignment are represented on the same line, since they should be located in the same pocket on the protein. Predictions are reported separately at the bottom of the page, ranked by a reliability score. The prediction list contains fundamental summary information: amino acid composition, predicted ligand and site biological relevance and pharmacological annotations (if any) for the ligands used. All these data can be browsed in details through available links; it is possible to check the original detailed output (figure 4, Material and Methods) that includes all the binding site alignments, and also to retrieve extended information for every prediction and ligand, as shown in figure 17. Finally it is possible to highlight individual predictions next to the sequence and to generate and download a PDB file with predicted residues highlighted. As a whole, the output web site provides an intuitive representation of the new consensus predictions generated by *firestar*.

QUERY: 1tcoC

GVQVETSPDGGRTFPKAGQTCVHYHTGALDEGKFKDSSRDNRNPFKFLMGKQGEVR
GWEQVAGHSVGRKALTSPOYATGATGHPGIPPHATLVYDVELLKLKLE

CATALYTIC SITE PREDICTION

Here cross-references between EC codes and GO terms and TEMPLATES are listed for every CATALYTIC SITE prediction

CAT 1 EC:5.2.1.8 > GO:peptidyl-prolyl cis-trans isomerase activity ; GO:0003755

CAT 1 TEMPLATES FROM

LITERATURE [J&S6](#)

BINDING SITE PREDICTION

SITE 1

All ligands_ID and their absolute frequency: PK3(20) RAR(14) D4N(6) 163(4) 861(4) DMS(4) RUQ(3) GPF(2) TST(2) SB1(2) D15(2) SB7(2) L3Q(2) SUB(2) D55(2) JPK(2) JZF(2) B58(2) 001(2) B1R(2) ARD(2) AF1(2) SB3(2) 4G1(2) 1KH(2) H9R(2) 0M2(2) 3KT(2) 854(2) 0DS(2) SBK(2) 3KY(2) 1KZ(2) RAD(2) 1RE(2) PKA(2) 0M0(2)

Associated GO terms: -

Compounds with pharmacological annotation [PK3](#) [RAF](#) [DMS](#) [001](#)

COMPOUNDS WITH PHARMACOLOGICAL ANNOTATION

PDB COMPOUND [PK3](#) [PDB web page](#)

	DrugBank	DB02864	Tacrolimus is a macrolide antibiotic. It acts by reducing peptidyl-prolyl isomerase activity by binding to the immunophilin FKBP-12 (FK506 binding protein) creating a new complex. This inhibits both T lymphocyte signal transduction and IL2 transcription. Although the activity is similar to cyclosporine studies have shown that the incidence of acute rejection is reduced by tacrolimus use over cyclosporine. Tacrolimus has also been shown to be effective in the topical treatment of eczema, particularly atopic eczema. It suppresses inflammation in a similar way to steroids, but is not as powerful. An important dermatological advantage of tacrolimus is that it can be used directly on the face; topical steroids cannot be used on the face, as they thin the skin dramatically there. On other parts of the body, topical steroid are generally a better treatment.
	CHEBI	42029	immunosuppressive agent
	PubChem	442643	Agents that suppress immune function by use of diverse mechanisms of action. Classical cytotoxic immunosuppressants act by inhibiting DNA synthesis. Others may act through activation of T-CELLS or by inhibiting the activation of HELPER CELLS. While immunosuppression has been brought about in the past primarily to prevent rejection of transplanted organs, new applications involving modulation of the effects of INTERLEUKINS and other CYTOKINES are emerging.
	KEGG DRUG	D08356	Immunosuppressant
	CHEMBL	ChEMBL_202733	FK506-binding protein 12 inhibitor

Figure 17 Extended ligand information page for the *firestar* prediction for PDB structure 1TCO. For every predicted catalytic site the templates from which the catalytic site was sourced are available through external links. For predicted binding sites the complete list of template binding ligands from templates in the clustered PDBs, together with available annotations from FireDB, is shown. Whenever pharmacological information exists for a compound, this is also retrieved from FireDB.

4.4.4 Site reliability score

Initially, predictions were ranked by the mean SQUARE conservation score of the binding site residues, since higher conservation correlates with higher reliability. This is a reasonable criterion but it is site-size dependent, since smaller sites (e.g. metal-binding sites) usually have better mean conservation as a result of higher pre-residue selective pressure.

In order to get a more refined evaluation of the reliability of the predictions, we created a new composite score taking into account 4 different parameters that from our experience are important to evaluate a predicted binding site:

1. Overall site conservation, in terms of mean SQUARE score of the binding residues;
2. Coverage of predicted amino acids in comparison with the annotated size of the ligand binding site in FireDB;
3. The percentage of identity of the closest template homologue in which the binding site is found;
4. The fraction of the total aligned templates included in the prediction.

In all criteria higher values correlate with higher reliability. The final score is the mean of these 4 normalized parameters. A higher weight is given to the conservation score since it is our main discriminating criteria; it counts double in the calculation of the mean.

As an illustrative example of how this composite score works we can look at *firestar* predictions for a putative glycerophosphodiester phosphodiesterase with a known structure (PDB code 3NO3) from *Parabacteroides distasonis* (figure 18). This enzyme catalyzes the hydrolysis of the 3'-5' phosphodiester bond of glycerol-3-phosphoethanolamine into glycerol-3-phosphate (G3P) and ethanolamine. A metal cation

is required at the binding site to coordinate two residues involved in the binding and catalysis of the reaction, histidine 74 and aspartate 61. *firestar* reports 3 binding sites: for the catalytic cation (MG), for the substrate binding site and for a second metal binding site partially overlapping the first one. The conservation score is high for both metal binding site predictions. However, the other three parameters in the composite score distinguish between the two predictions. The catalytic magnesium binding site has perfect coverage, the source templates include close homologs and the site is present in the majority of the extracted alignments. The second metal binding site is found only in distant homologues and has few supporting alignments, so it has a poor overall reliability score.

SITE 1 Compound_name:	MAGNESIUM ION
SITE 1 Site classification:	COGNATE 
SITE 1 Reliability:	97.2% [COV: 100% SITE: 100% Ident: 100% - Ali: 86%] 
SITE 1 Source alignments:	12 --> here for more information about this site prediction
SITE 1 Residues resume:	E(59); D(61); H(74); E(123);
SITE 2 Compound_name:	SN-GLYCEROL-3-PHOSPHATE
SITE 2 Site classification:	COGNATE 
SITE 2 Reliability:	85.2% [COV: 100% SITE: 100% Ident: 26% - Ali: 100%] 
SITE 2 Source alignments:	14 --> here for more information about this site prediction
SITE 2 Residues resume:	H(30); R(31); E(59); D(61); H(74); E(123); K(125); F(158); W(222);
SITE 3 Compound_name:	MAGNESIUM ION
SITE 3 Site classification:	COGNATE 
SITE 3 Reliability:	57.3% [COV: 80% SITE: 91% Ident: 16% - Ali: 7%] 
SITE 3 Source alignments:	1 --> here for more information about this site prediction
SITE 3 Residues resume:	E(59); D(61); K(125); W(222);

Figure 18 snapshot of *firestar* prediction section output for glycerophosphodiester phosphodiesterase (PDB code 3NO3). The three sites are reported with the predicted ligands. The reliability scores are shown in green boxes all the components of the score are listed between square brackets: COV = coverage of predicted amino acids compared to annotated binding site size; SITE = mean site SQUARE score; Ident = the percentage of identity of the closest template homologue with site; Ali = percentage of aligned templates with homologous site. Red boxes highlight the conservation score, which was used in earlier versions of *firestar* as the reliability score

4.5 Improvements in *firestar* algorithm

The availability of an experimentally validated test dataset is fundamental for the development of prediction method, in order to get a real evaluation of the performance. Since our source information is the PDB, ideally we should test our algorithm against ligand binding structures not included in the PDB. The Critical Assessment of techniques for protein Structure Prediction (CASP) provides structures not yet in the PDB and has proved to be the perfect testing ground for *firestar*. CASP is a biennial experiment supported by the structural and computational biology community with the goal of blindly testing the performances of state-of-the-art 3D structure prediction methods. Over approximately three months, participants are sent new protein sequence targets every day and have to submit a predicted structure within a given timeframe. The chosen proteins have already been resolved but not yet released publically. CASP started in 1994 and it is still active. Initially it concentrated solely on protein 3D structure prediction, but over these years it has included other protein feature categories. For example, “*protein-protein interactions*” was introduced in the second edition: this category finally leads to the

creation of a new independent experiment called CAPRI (Critical Assessment of Prediction of Interactions), now at its 39th round.

Function prediction was introduced for the first time in 2004 (CASP6) in response to the increasing interest in the scientific community. Since function is a wide concept, a number of different features were accepted and evaluated in this edition, from GO annotations to 1D prediction for post-translational modifications. However exhaustive target evaluation was not possible for most proteins, since reliable information was not available for most of the targets, even several months after the end of the experiment.

Our group participated as the evaluator of two editions of CASP, the seventh and the eighth. While the CASP7 edition focused mainly on GO terms and EC code predictions, the CASP8 evaluation was based entirely on the prediction of ligand binding sites. The chosen metric was the MCC score (see section 3.5, Materials and Methods), since it is able to deal with binary classifications when there is a large imbalance between the numbers of true positives and true negatives, as is the case in the ligand-binding category (binding residues are highly outnumbered by non-binding residues). The experience of the evaluation and the data collected during the assessment (see appendix tables 1 and 2) laid the foundations for the development of *firestar*.

Over the CASP7 and CASP8 experiments, *firestar* was able to predict ligand-binding sites for 46 out of 49 targets that were crystallised with biologically relevant ligands. During CASP8, even though *firestar* was not officially participating in the experiment, it generated predictions for all but one target and achieved a mean MCC score of 0.761 over the 26 predicted targets. It predicted a total of 353 residues: 237 were correctly predicted (TP, 72%), 93 were false positive (FP, 28%), and 53 confirmed binding residues were missed (FN). There were 23 additional predicted residues that did not enter in any of these groups since the assessors had tagged them as neutral. These are residues that should bind a biological ligand, but the ligand was not present in the crystallographic structure. These residues did not count as TP or FP (if predicted) or FN or TN (if not predicted). Sensitivity and specificity were 80% and 67% respectively, suggesting that at the time *firestar* had a certain tendency to over-predict. Indeed, *firestar* was set to make predictions at a distance of 1.5 Å while the official distance used to define a ligand-binding residue was 0.5 Å. Most of false positive predictions were residues close to the binding site: of 108 false positives, 74 were within 2.5 Å of the bound ligand. *firestar* still had a better mean MCC score than all officially participating groups in CASP8, human or automatic predictors.

In CASP9 *firestar* participated officially for the first time as server predictor. In the official assessment¹⁴⁰, it ranked 2nd as automatic server and 3rd in the global ranking. 29 targets were included in the evaluation, and this number should have provided statistical robustness to the results. Unfortunately, the selection of biologically relevant compounds to calculate the binding sites was controversial: in many targets (13) the ligand or ligands used were all non-biological. While in some cases at least the size of the crystallised and predicted compound was more or less comparable (eg: target T0609, where tartaric acid was assumed to be a good replacement for galactose), in other cases the difference was huge (eg: target T0622, where sulfate was taken to represent NAD). In addition there were a further 4 targets where a biological ligand was crystallised with a non-biological one and both were considered in the definition of the binding site and two targets that bound COGNATE/AMBIGUOUS metals to biologically relevant sites, but these metal ligands were present in the crystallization conditions and were not the biological binding partner (T0518 and T0635).

Data and results collected from these experiments (with the exception of CASP9) have been fundamental for the development and the improvement of *firestar* presented in this work.

4.5.1 Introduction of HHsearch sequence search method

The main factors affecting *firestar*'s capability to detect a binding site depends on:

- The presence of a homologous template in the source database, FireDB.
- The ability to find this sequence homolog among the others.
- The ability to generate a good sequence alignment between template and target.

While the first point basically depends by the growth of the PDB database, the other two rely on the chosen search method.

firestar since the beginning integrated PSI-BLAST as the search algorithm: PSI-BLAST in *firestar* generates non-biased profiles for target sequences from a 70% non-redundant version of the UniProtKB database (see section 3.4.2, Materials and Methods) and this is used to explore the FireDB sequence space. The PSI-BLAST output provides a ranked list of candidates and their respective alignments with the target.

New generation sequence search methods, based on profile-profile comparisons, have a better sensitivity for distant homologues since they generate better alignments^{165,166}. For this reason we decided to test a sequence search method based on profiles, HHsearch, since it was the best performing server for template-based predictions in the official CASP9 assessment⁶⁸. HHsearch generates profiles from a sequence database, and these are used to search against a database of *ad-hoc* generated FireDB profiles. As test dataset we used the genes annotated by GENCODE (version 3C¹⁶⁷) for chromosomes 21 and 22, a total of 798 genes. *firestar* using PSI-BLAST predicted 12.657 ligand-binding residues, while using HHsearch it predicted 15.078 residues; the union of the two methods produced 17.027 non-redundant ligand binding residues, a 34% improvement on coverage over PSI-BLAST alone. The results showed that the HHsearch is able to predict 19% more residues than PSI-BLAST but also demonstrates that the two methods are complementary rather than overlapping. So we decided to integrate HHsearch as search method in addition to PSI-BLAST.

In order to further evaluate the integration of both sequence search methods, we used the CASP8 dataset, where *firestar* results were obtained with PSI-BLAST as the only integrated search method. We ran the algorithm with three different combinations of search methods: PSI-BLAST only, HHsearch only and both methods combined. Since the FireDB database structure changed over the years and also because the version used during the CASP8 experiment was no longer available, we decided to use the most recent version. To exclude the use of information coming from target structures themselves and to limit the contribution of newly available close homologs, we put a maximum percentage of identities limitation to the templates, 35%. It has to be said that even using this restriction, new templates with a sequence identity lower than the established cut-off are present and *firestar* used them to generate predictions but, at the same time, we also discarded any previously existent template with higher sequence identity.

We compared our results with those of the CASP8 experiment and observed that although the new *firestar* using PSI-BLAST and an updated database had a comparable number of true positive predictions to CASP8, it also had a noticeable increment in false positives (see figure 19). These results confirm that the introduction of HHsearch generated an additional set of templates, not completely overlapping with the ones generated by PSI-BLAST, and the two of them are complementary methods.

These extra templates were important to increase *firestar* sensitivity, but also came with new false positives. The absolute number of false positive is higher if we compare each method with the CASP8 results, and even more when they are combined. We found that the greatest increase of predicted false positive residues was for sites where an adenine phosphate nucleotide compound was bound ("ADP only" column). For

other types of binding site, the numbers of false positives actually went down (see figure 19).

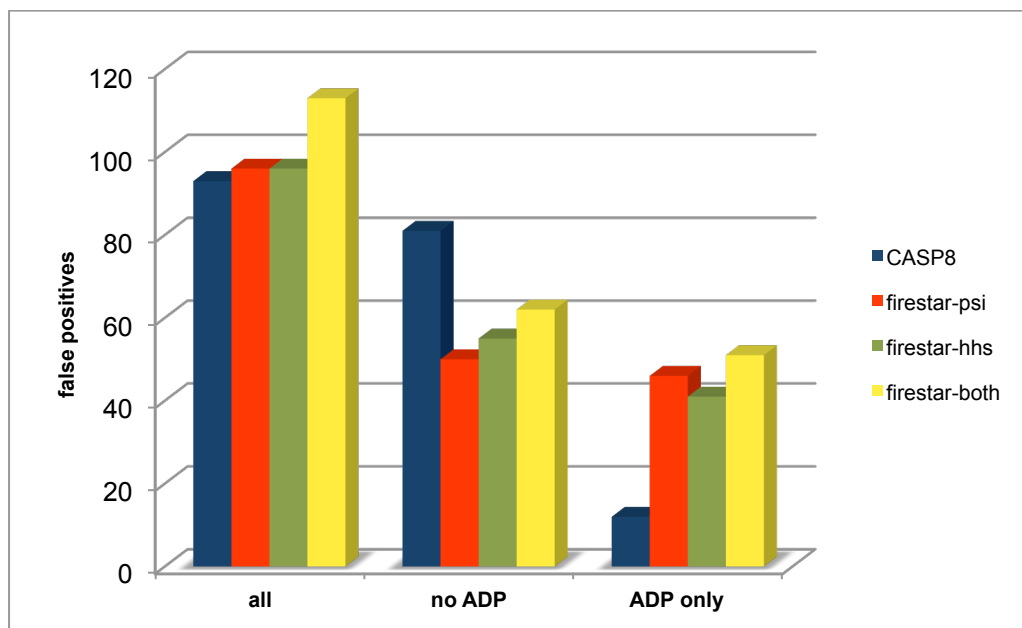


Figure 19 False positive count of firestar predictions for the CASP8 experiment dataset. Blue bars refer to the original CASP8 predictions, while the other colors refer to more recent predictions using updated search databases and different sequence search methods. Predictions are considered all together (26 targets) and by separating adenine phosphate nucleotides sites (4) from the others (22).

4.5.2 Metal Binding Sites

The number of annotated metal binding sites inside FireDB is far higher than any other type: zinc (ZN) and magnesium (MG) are the most common compounds in FireDB, with respectively 23,403 and 18,343 annotated binding sites. In contrast the most common cofactors, the protoporphyrin IX containing iron (HEM) or flavin-adenin dinucleotide (FAD), total 7,733 and 4,033 binding sites respectively. This high frequency comes with a high level of noise: looking at the automatic biological relevance assessment, 56% of them are tagged as biologically relevant. Reduced size, net charge and the recurrent presence in crystallization mixes are the main causes of this high rate of non-specific associations. Bearing in mind the results obtained from the analysis composition of metal binding sites (see section 4.1.3), we went forward and searched the literature in order to establish metal-specific rules able to determine unequivocally whether each candidate prediction has higher probabilities to be biologically important.

The most frequent metal in PDB is zinc; it appears in many structures, frequently as a divalent cation. In some proteins it holds a catalytic role¹⁶⁸, stabilizing negative charges (carboxypeptidases or dehydrogenases), activating the reactive species (alkaline phosphatases) or coordinating active groups (cytidine deaminase). In other proteins zinc has a structural role¹⁶⁹, such as in the zinc finger domains in transcription factors or in RING finger domains, which have a key role in the interaction of the protein with other macromolecules in the cell. The different architecture of the zinc binding sites has been studied for specific proteins and for specific families¹⁶³. We extrapolated from all this information a pattern in terms of size and amino acid composition that distinguishes biologically relevant zinc binding sites from other zinc binding sites:

- Amino acid signature: cysteine, histidine, aspartic and glutamic acid;
- Constraint: at least two of the previous amino acids has to be present;
- Size: from 2 to 5 residues

These features have been transformed in a filter applied in *firestar* pipeline, after the exclusion of poorly conserved residues step (shown in section 4.4.1): if a candidate zinc binding site does not fulfill these characteristics, it is automatically filtered out.

Apart from zinc, we also extrapolated a biological relevance pattern for the second most frequent metal, magnesium, from the literature and FireDB:

- Amino acid signature: aspartic and glutamic acid;
- Constraint: at least one of the previous amino acids has to be present;
- Size: minimum 3 residues;

Furthermore, we started the same with calcium, based on a published study¹⁶⁴:

- Amino acid signature: aspartic and glutamic acid, asparagine and serine;
- Size: minimum 2 residues;

It is important to keep in mind that this pattern applies to EF-hand domains, since non EF-hands calcium binding proteins are more variable; further studies and different patterns will be probably required to establish a complete prediction filter for calcium. These filters could be extended in the future to the others metals in FireDB based on available information on the binding architecture of these compounds.

4.5.3 Non Metal Binding Sites

The sites grouped under the tag of “non-metal binding” are much more variable than metal binding sites. The number of different molecules bound by PDB structures in non-metal binding sites in FireDB is high and the size of these ligands runs from molecules with few atoms (for example, Oxygen - O₂) to molecules like LHI (C₉₃H₁₅₅N₇O₂₃P₂S), a lipid II analogue made up of over 200 atoms. This heterogeneity and the consequent variability of the binding sites make it difficult to elaborate ligand specific rules.

An in depth analysis of our CASP8 centered experiment showed us that not all the candidate residues in binding sites had the same importance. There is a core of important residues that are usually well conserved (for example, the P-loop for ATP or GTP binding sites), while other binding residues may be much more variable.

The master Sequence binding site (MSS) of the aligned ATP binding site residues from FireDB cluster 1u5qA can serve as an example (figure 20). This cluster groups TAO2 serine-threonine mitogen-activated protein 3 kinase domains that phosphorylate the MAP 2 kinases *MEK3* and *MEK6* twice, activating them. The phosphate binding loop, the purine base binding loop and the third phosphate group binding loop are very well conserved across the different sites in the MSS. The non-conserved amino acids in the MSS lie outside these three loops and may be specific to this kinase subfamily. Another possibility is that they are required for the binding of Staurosporine (STU), an unselective inhibitor of protein kinases that binds competitively in the ATP binding pocket, since its binding residues are collapsed with the ATP binding residues to form the 1u5qA MSS.

t_1u5qA	clst	100	I	G	H	G	S	V	A	K	I	M	E	Y	C	L	G	S	G	N	L	D	K	21	1.00	Expand	STU	ATP									
t_3a7qB	clst	43	I	G	K	G	S	V	A	K	-	M	E	Y	L	-	S	A	N	L	D	-	19	0.78	Expand	ATP	ADP	STU									
t_2clqA	clst	31	L	G	K	G	-	V	A	K	V	M	E	Q	V	G	S	D	N	L	D	-	22	0.72	Expand	STU	IE6	IM6	IE8	BH9	IE0	IE4					
t_4eheA	clst	30	I	G	S	G	-	V	A	K	L	T	Q	W	C	E	G	S	N	N	F	D	41	0.72	Expand	BAX	B1E	FP4									
t_4eqcA	clst	38	I	G	Q	G	-	V	A	R	V	M	E	Y	L	-	G	S	D	N	L	D	31	0.67	Expand	FLL	2QL	X4Z	0H2	XR1	DW1	ATP					
t_3w8qA	clst	29	L	G	A	G	N	V	A	K	V	M	E	H	M	-	G	S	N	L	D	-	32	0.67	Expand	ATP	ADP	AGS	ANP	YSO	KSA	5EZ	22T	MT8	92P		
t_1s9lA	clst	29	L	G	A	G	N	V	A	K	-	M	E	-	M	-	S	S	N	L	D	-	19	0.66	Expand	ATP											
t_3alnA	clst	26	I	G	R	G	-	V	A	K	-	M	E	L	M	-	S	S	N	L	D	-	20	0.66	Expand	ANP											
t_2xnmA	clst	28	I	G	-	-	C	V	K	V	M	E	Y	C	E	-	G	A	N	F	D	-	26	0.66	Expand	WCX	WGZ	BX1	T3M	XK3	5Z5	430	5R1	GGY	OL2	JUP	ED8
t_4fieB	clst	35	I	G	E	G	S	V	A	K	V	M	E	F	L	-	G	A	D	-	L	D	32	0.64	Expand	ANP	2OQ	X4Z	2QL	23D	N53	2OO	NJD				
t_4nstC	clst	26	I	-	E	G	T	V	A	K	-	F	E	Y	M	-	D	S	N	L	D	-	17	0.64	Expand	ADP											
t_4mneE	clst	30	L	G	A	G	N	V	A	K	-	M	E	-	M	-	S	N	L	D	-	-	20	0.63	Expand	ACP	ANP	ATP									
t_3slsB	clst	30	L	G	A	G	N	V	A	K	-	M	E	-	M	-	S	N	L	D	-	-	20	0.63	Expand	ANP											
t_4an2A	clst	30	L	G	A	G	N	V	A	K	-	M	E	-	M	-	S	N	L	D	-	-	19	0.63	Expand	ACP	ATP										
t_4aguC	clst	25	I	G	E	G	-	V	A	K	V	F	E	Y	C	D	-	T	-	N	L	D	20	0.62	Expand	D15											
t_4lqdB	clst	39	L	-	E	G	S	V	A	K	-	M	E	Y	C	-	-	-	-	-	L	D	15	0.61	Expand	ANP											
t_4eojC	clst	27	I	G	E	G	T	V	A	K	V	F	E	F	L	H	Q	D	Q	N	L	D	24	0.61	Expand	ATP	MFR	MHR	NU5	4SP	ST8	VAR					
t_3d0eB	clst	25	L	G	K	G	-	V	A	K	T	M	E	Y	A	-	-	E	N	M	D	-	29	0.61	Expand	G96	ANP	G98	G95	G93	L20	X39	ISS	GVP	X37		
t_3hkoA	clst	25	-	G	Q	G	S	V	A	K	-	M	E	L	C	-	H	-	-	L	D	-	14	0.60	Expand	ANP											
t_4qclA	clst	26	I	G	E	G	T	V	A	K	V	F	E	F	L	H	Q	D	Q	N	L	D	36	0.57	Expand	ATP											
t_2bmcB	clst	27	L	G	K	G	K	V	A	K	L	L	E	Y	L	G	T	-	E	N	L	D	45	0.57	Expand	ATP											
t_4fzaB	clst	42	I	G	-	G	-	-	A	K	-	M	E	Y	L	-	-	-	A	N	L	-	16	0.55	Expand	GVD	DKI										
t_4fr4B	clst	25	I	G	K	-	-	V	A	K	V	V	D	L	L	-	-	G	D	N	L	D	20	0.55	Expand	STU											
t_3d2kA	clst	26	L	G	K	G	K	V	A	K	L	L	E	Y	-	T	-	-	-	L	D	-	25	0.55	Expand	AK4	FXG										
t_4fkwA	clst	26	I	G	E	G	-	V	A	K	V	F	E	F	L	H	Q	D	Q	N	L	D	26	0.54	Expand	56Z	316	LZ2	HMD	62K	FRT	03Z	BRY	FCP	LZ7	LZ4	
t_2j7lA	clst	36	L	-	-	-	V	A	K	V	I	E	F	C	P	-	G	-	-	L	D	-	25	0.53	Expand	XZN	G61	274									
t_3aloA	clst	30	I	G	-	G	-	V	A	K	-	M	E	-	M	-	S	N	L	D	-	-	16	0.53	Expand	ANP											
t_1zxcC	clst	29	L	G	Q	-	-	V	A	-	V	-	E	Y	C	-	-	-	-	N	F	D	14	0.53	Expand	ATP	ANP										
t_1v0bA	clst	22	I	G	E	G	-	V	A	K	V	F	E	H	L	D	-	D	Q	-	L	D	19	0.53	Expand	INR	PVB										
t_2c30A	clst	36	I	-	-	G	-	V	A	K	V	M	E	F	L	-	G	-	D	-	L	D	17	0.52	Expand	X4Z	B49										
t_2wtkB	clst	30	I	G	K	G	-	V	T	R	-	M	-	-	-	S	S	H	L	-	-	-	22	0.52	Expand	ANP											

Figure 20 Aligned residues from sites collapsed into the FireDB MSS for the ATP binding site of *t_1u5q*. These residues in these sites are evolutionarily related and conservation at the different positions (calculated using SQUARE) is represented using a color code: the darker the blue, the higher the conservation. In the alignments three conserved blocks of residues (the phosphate binding loop, the purine base binding loop and the third phosphate group binding loop) are clearly recognizable (inside the green boxes) even in remote homologues, while conservation among residues outside the highlighted boxes is harder to detect.

We concluded that residues identified in the candidate alignment, should be evaluated with different conservation threshold. Residues aligning with the core part should be evaluated with a more permissive cut-off, since they've a higher probability to be in the binding site. Residues coming from the variable part, on the contrary, should be evaluated with a more restrictive cut-off, especially if coming from remote homologs.

We integrated an adaptive filter based on conserved patches. *firestar* uses the homologous site analysis information (when available) stored in the FireDB table COMPARE35 (see section 3.7, Materials and Methods), to evaluate aligned candidate sites. Core residues are given a higher weight during the *firestar* filtering, while the variable residues are maintained only if their conservation, in terms of their SQUARE score, is high. Furthermore COGNATE containing sites are tagged for a separate merging.

4.5.4 Merging per-residue frequency filter

We designed a per-residue frequency filter to apply at the merging step to penalize isolated predicted residues that are not well conserved and at the same time to favor information coming from close homologs. The final composition of the predicted site is now not merely combination of all residues from a group of overlapping candidate alignments. Instead every predicted residue is evaluated by its frequency and origin (figure 21).

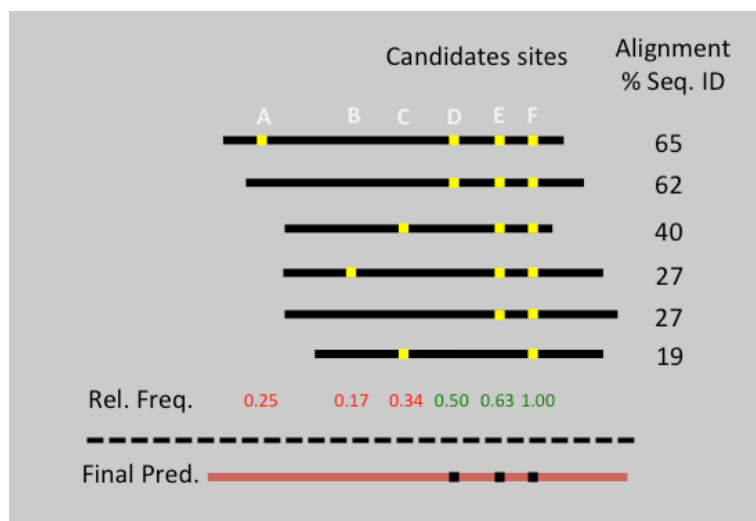


Figure 21 Illustrative example of consensus prediction merging schema in *firestar*. For residues coming from close homologous alignments (>60% Seq. ID), frequency is calculated as follows: (relative frequency in close homologs + frequency in the rest)/2. For positions A, D, E, F is calculated respectively as $(0.5+0)/2 = 0.25$, $(1+0)/2 = 0.5$, $(1+0.75)/2 = 0.63$ and $(1+1)/2 = 1$. For all the residues coming from more distant homologs, the relative frequency is calculated as number of hits / total number of candidate sites. So for positions B and C is respectively $1/6=0.17$, $2/6=0.34$. Please note that position B and C, in spite of having the same global frequency of positions A and D, have a lower relative frequency and are discarded. Positions D, E and F are transferred to target sequence (red line and black squares), since their relative frequency is higher than 0.35 (green values).

The relative frequency of every residue in a consensus prediction is calculated taking into account the percentage identity of their source alignments. Predicted positions extracted from close sequence homologs (we chose > 60% identities, based on our experience) are considered more reliable and are given a higher weight. Their relative frequency is calculated as the mean of their frequency in close homologs (the number of supporting alignments divided by the total number of alignments) and the frequency in more distant homologs. For all the others residues, frequency is calculated over the total number of alignments. The presence of close homologs also raises the relative frequency cut-off, from 0.25 to 0.35. This strategy has been shown to be effective to spot positions specific to a particular protein sub-family.

4.5.5 The effect of the filters and new CASP8 dataset assessment

The effects of the different improvements have been studied step by step, running the algorithm against the CASP8 dataset multiple times. It would be impossible to report all the tests in this thesis, because of their number and even more for their non-linear evolution. In some cases filters have been introduced, discarded and later re-introduced in a different combination. As results of these analyses we obtained the final set of improvements that we have presented here. Below we present two example cases to illustrate in detail how the final improvements affected the predictions as well as the results for CASP8 targets.

The first case is the prediction of a zinc binding site for the target T0480 (PDB code 2K4X), the ribosomal protein S27A from *Thermoplasma acidophilum*. It forms part of the 30S subunit of the bacterial ribosome, but its specific role inside this macromolecular complex has yet to be unveiled. This peptide contains Pfam domain PF01599, which is

probably involved in the recognition and capture of mature mRNA to be translated and is proposed to form a zinc finger. In the CASP 8 experiment *firestar* predicted 6 binding residues, the 4 real zinc binding cysteines and two false positives, an arginine and an histidine. Now prediction includes only the four true positive cysteines (figure 22, B).

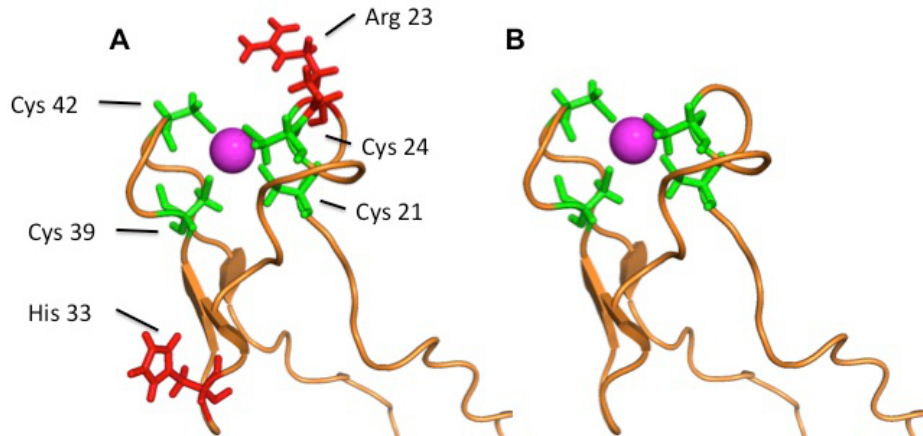


Figure 22 **A)** Original prediction of *firestar* during the CASP8 experiment for target T0480 mapped (amino acid represented with sticks) onto the real structure (PDB code 2K4X). In green the true positive predictions and in red the false positives, arginine 23 and histidine 33. **B)** New *firestar* prediction after the introduction of filters: both false positives have been filtered out without true positive loss.

Arginine 23 comes from one MSS containing an iron-sulfur cluster (SF4, figure 23, A), with an original size of 11 amino acids. The candidate site has been filtered out because its reduced size compared with the original (only 3 over 11 residues pass the conservation filter). Histidine 33 comes from one MSS binding to zinc (figure 23, B): it is aligned to another histidine after a 9 residue deletion. It is clear that there is a problem in the alignment, caused by a well-conserved block of residues around position 40. This candidate residue is now filtered out in the final merging step because it only appears one time over more than 110 candidate sites. It is worth mentioning that PSI-BLAST provided only 7 templates (6%) for this target.

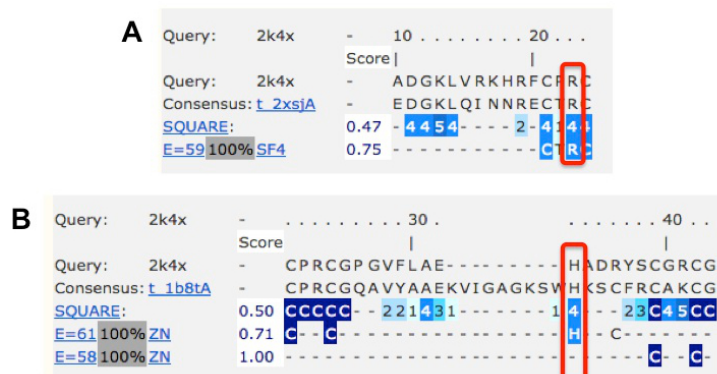


Figure 23 Alignments extracted from PSI-BLAST analysis results for T0480. In CASP8 all residues with SQUARE score > 3 passed the filter. **A)** Source template MSS for the FP arginine 23 (red frame) **B)** Source template MSS for the FP histidine 33 (red frame)

A representative example for a non-metal binding site is the case target T0483 (PDB code 3DLS), a Per-Arnt-Sim kinase (PASK) ATP binding domain. It is part of an

evolutionary conserved protein kinase that acts as a sensor involved in energy homeostasis; the deposited structure contains an ADP molecule with two biologically relevant magnesium divalent cations.

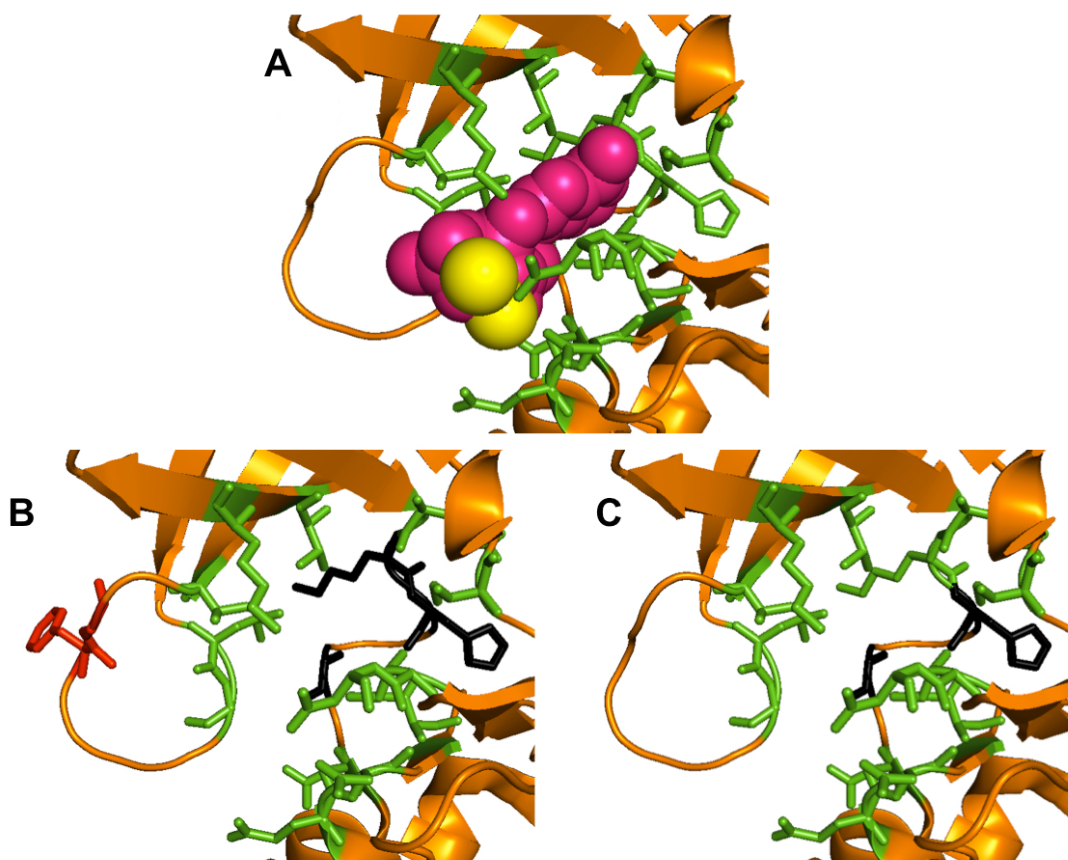


Figure 24 A) Graphical representation of T0483 binding site (PDB code 3DLS). Contacting residues are represented in sticks and colored in green. The ligands are represented as spheres, pink for the ATP and yellow for magnesium B) *firestar* predictions in the CASP8 experiment are mapped onto the binding sites and the ligands have been removed in order to facilitate the view: in red are represented the false positives, while in black are colored the non predicted residues (FN) C) prediction for the same target after the introduction of new filters, same color code for residues.

firestar performed very well in CASP8, predicting correctly 13 of 16 binding residues with two false positives (MCC=0.83), but after the introduction of filters we were able to identify an extra TP and to filter out 2 FP (figure 24, C). Phenylalanine 37 and glycine 38 have a good SQUARE score (higher than 3, figure 25); looking at the homologous site information (figure 26), it is clear that these two residues are in the P-loop but they are less conserved compared with the rest. The penalization discards glycine 38, but phenylalanine still passes the final merging step, where it is finally discarded for low relative frequency. The use of different weights for core and specific residues also allows us to detect a 14th true positive in the loop that binds to the purine base.

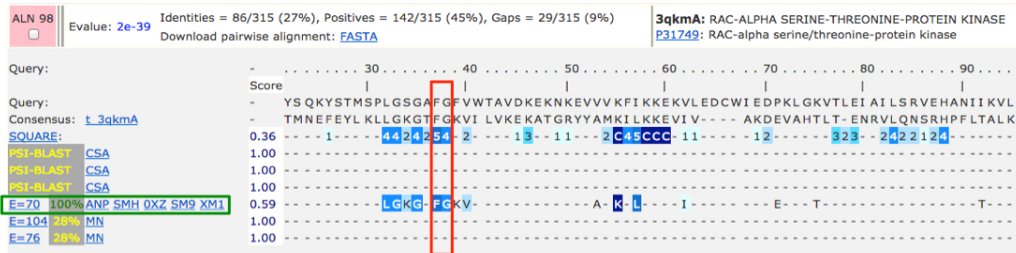


Figure 25 Alignment extracted from PSI-BLAST analysis of target T0483. The homologous binding site is highlighted in a red box. Every annotated residue is mapped with its SQUARE score (the darker the blue, the higher the score). Phenylalanine 37 and glycine 38 are highlighted in the green box, and have SQUARE scores of 5 and 4 (filter threshold = 3).

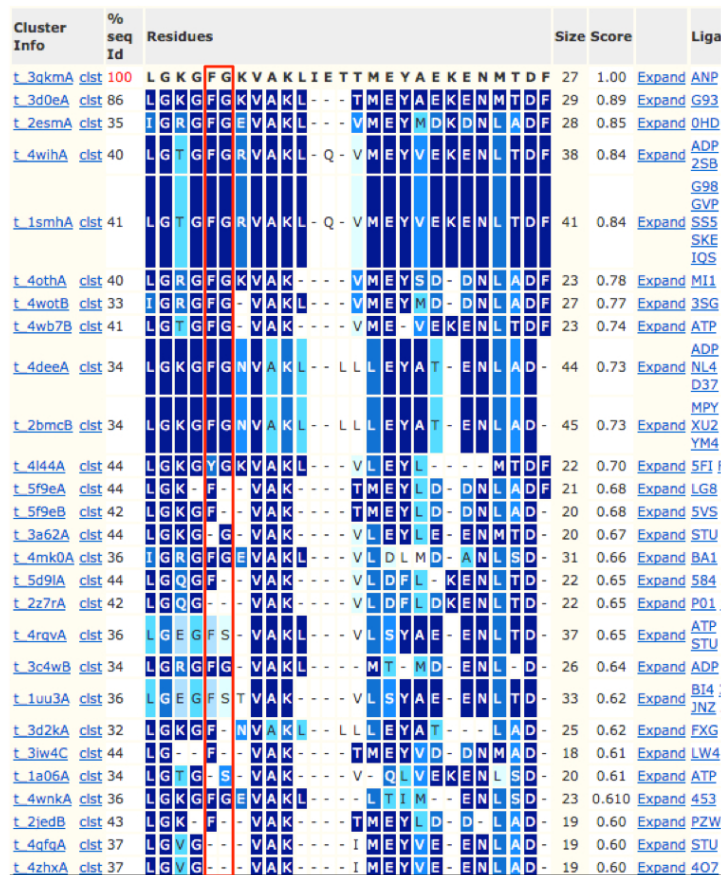


Figure 26 Representation of homologous binding MSS site for the ANP (a non-hydrolyzable analog of ATP) binding site of cluster 3qkmA. Conservation is represented by the background color: the darker the blue, the more conserved. Phenylalanine 37 and glycine 38 corresponding position are highlighted in the red box, and they have worse conservation than the surrounding residues.

To evaluate the performance of the new algorithm, we decided to run the CASP8 dataset again. The introduction of new parameters reduced global number of predicted false positives by 53% (from 93 to 43) and increased sensitivity, since the number of the true positives increased from 237 to 257. Most targets where the filters predicted more true binding residues and at the same time reduced false positives were non-metal binding sites (for example T0396, T0483, T0485, T0490). For a number of targets in the original CASP8 predictions, *firestar* identified the correct whole binding site, but generated

a large number of false positives (T0453, T0457, T0461, T0470, T0476, T0480). Here the new filters maintained the high sensitivity and drastically increased specificity.

Target	Compounds ID	Site size	CASP8 experiment			New <i>firestar</i>		
			FP	TP	MCC	FP	TP	MCC
T0391	FES	9	0	7	0.876	1	9	0.945
T0394	PO4	8	1	7	0.871	0	7	0.934
T0396	FAD	23	4	16	0.681	3	20	0.833
T0406	NI	3	2	3	0.770	0	3	1.000
T0407	ZN(3)	9	0	9	1.000	0	8	0.941
T0410	FE	3	2	3	0.773	1	3	0.865
T0422	ADP	15	2	13	0.861	1	12	0.854
T0425	ZN	3	0	3	1.000	0	3	1.000
T0426	ZN	3	1	3	0.864	0	3	1.000
T0430	AMP-MG	19	7	13	0.648	5	14	0.722
T0431	HEM	19	11	15	0.660	8	15	0.705
T0440	FE(2)-ZN	9	0	9	1.000	0	9	1.000
T0444	FE	4	5	4	0.661	2	4	0.814
T0450	FAD	32	11	32	0.854	8	29	0.833
T0453	CA(3)	4	5	4	0.647	0	4	1.000
T0457	MG	4	6	4	0.626	0	4	1.000
T0461	ZN	3	1	3	0.864	0	3	1.000
T0470	MG	4	3	4	0.751	0	4	1.000
T0476	ZN	4	4	4	0.693	0	4	1.000
T0477	ADP	10	3	10	0.871	2	10	0.909
T0478	MG-FE	7	0	0	0.000	0	5	0.842
T0480	ZN	4	2	4	0.800	0	4	1.000
T0483	ADP-MG(2)	16	2	13	0.831	0	14	0.932
T0485	SAM	19	6	11	0.577	4	15	0.769
T0487	MG	4	0	3	0.865	0	3	0.865
T0490	FAD	33	12	23	0.644	8	30	0.831
T0508	SAM	19	3	17	0.858	0	18	0.971

Table 7 Comparison of the performance of the old and new *firestar* algorithm over all the targets included in the function prediction category of the CASP8 experiment. The table shows the size of the binding site defined by the assessors, false and true positives predicted in CASP8 and the Matthews Correlation Coefficient (MCC). In the last column the cell is colored in green if the MCC is better than the original, in red if it is worse or background color if there is no variation.

The new *firestar* algorithm improved the predictions for 20 targets (table 7), predicted the same residues (correctly) for three targets and performed worse than in CASP8 for three targets. For these targets the new filters discarded false positives, but at the same time penalized a true binding residues. For target T0450 *firestar* filters discarded 3 templates and two of them were providing useful information. The final prediction lost three false positives, but it also lost three true positives.

Although T0478 is listed in the table, we have not taken it into account in the comparison. In CASP8 *firestar* did not predict binding residues for this target because there were no valid templates. The more recent version of FireDB used in the assessment

included a recently released remote template that allowed the algorithm to predict the site. Remarkably, the algorithm was able to predict 5 of out of 7 true binding residues. Over the 26 remaining CASP8 targets the new version of *firestar* obtained a mean MCC of 0.912, a considerable improvement on the previous mean MCC of 0.790.

4.5.6 CASP10 experiment

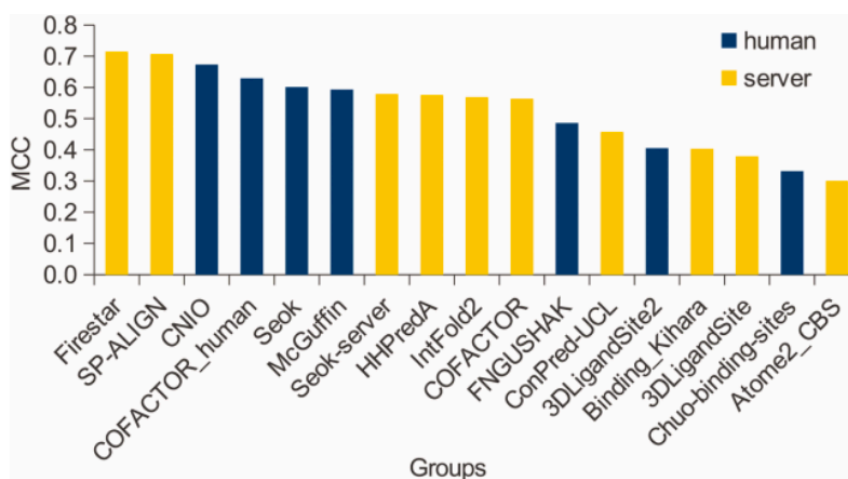
All *firestar* improvements were developed and tested on the variety of ligands and binding sites in the CASP8 dataset. To test the changes made to the algorithm independently and to understand if we overfitted the filters to the CASP8 data, we participated in the tenth edition of CASP experiment, celebrated in 2012.

19 groups sent predictions for the function prediction category, and *firestar* was the only sequence-based method. Only 13 of the 114 released targets were used for the official binding site prediction (FN) category assessment¹⁴¹. Looking at the results (table 8), even though *firestar* once again had the highest MCC score in the experiment, the mean MCC was only 0.715. MCC scores for three targets (T0657, T0659 and T0720) were remarkably low. Without these three targets, *firestar* MCC would have been 0.838.

Target ID	Compounds ID	Site size	FP	TP	MCC
T0652	AMP	13	4	12	0.821
T0657	ZN	4	16	4	0.423
T0659	ZN(2)	3	-	-	0
T0675	ZN(2)	8	0	7	0.929
T0686	MG	3	2	3	0.772
T0696	NA	3	0	3	1
T0697	LLP(2)	14	3	12	0.823
T0706	MG(2)	6	1	4	0.723
T0720	MN(10)-SF4(10)	14	0	4	0.495
T0721	FAD(2)	31	7	23	0.726
T0726	ZN	3	0	3	1
T0737	FAD	22	3	16	0.764
T0744	FNR	19	2	15	0.825

Table 8 *firestar* official CASP10 results for the thirteen targets evaluated in the function prediction category. For target T0659 no prediction was generated. Metal binding targets highlighted in blue.

firestar ranked as the best function predictor method in the experiment, performing better than the other automatic servers, but also than all human predictor groups (figure 27). Although it was the best performing method in CASP, the authors of the CASP paper stated that the size and the imbalance (in terms of size and type of binding sites) of the data set resulted in no statistical significance of the overall ranking between the first ten methods.



		FN119	FN326	FN475	FN208	FN473	FN285	FN261	FN430	FN273	FN227	FN082
Firestar	FN119	NA	0.59	0.56	0.31	0.22	0.17	0.07	0.33	0.07	0.07	0.01
SP-ALIGN	FN326	0.59	NA	0.79	0.54	0.54	0.50	0.27	0.74	0.50	0.08	0.05
CNIO	FN475	0.56	0.79	NA	0.45	0.31	0.33	0.22	0.67	0.27	0.09	0.03
COFACTOR_human	FN208	0.31	0.54	0.45	NA	0.76	0.91	0.31	0.55	0.62	0.36	0.13
Seok	FN473	0.22	0.54	0.31	0.76	NA	0.84	0.91	0.95	0.48	0.74	0.24
McGuffin	FN285	0.17	0.50	0.33	0.91	0.84	NA	0.93	0.74	0.29	0.59	0.27
Seok-server	FN261	0.07	0.27	0.22	0.31	0.91	0.93	NA	0.95	0.67	0.82	0.45
HHPredA	FN430	0.33	0.74	0.67	0.55	0.95	0.74	0.95	NA	0.81	0.95	0.31
IntFold2	FN273	0.07	0.50	0.27	0.62	0.48	0.29	0.67	0.81	NA	0.64	0.31
COFACTOR	FN227	0.07	0.08	0.09	0.36	0.74	0.59	0.82	0.95	0.64	NA	0.64
FNGUSHAK	FN082	0.01	0.05	0.03	0.13	0.24	0.27	0.45	0.31	0.31	0.64	NA

Figure 27 Top: Best function prediction servers participating in the CASP10 experiment, ranked by the mean MCC obtained over the 13 selected targets. **Bottom:** Wilcoxon signed rank test results among the first ranked 11 groups. Differences in accuracy of binding residue predictions were not significant (both images are extracted from the assessment official paper¹⁴¹).

4.5.6.1 Prediction re-assessment

After a critical evaluation of assessment data, based on the experience accumulated over the previous 3 editions of the experiment, we carried out a parallel assessment. Firstly, we wanted to fill an important gap left by the official assessors: neutral residues. True positives for these 13 targets were obtained only from crystallographic closeness of cognate ligands. There are some cases where additional information (from literature, crystallised analogs, conservation) supports the probable implication of other residues in binding the assessed ligands or ligands that are not crystallised with the protein structure. Neutral amino acids do not count as true or false positives, or true or false negatives. In addition a second assessment allowed us the possibility of enlarging the data set, thanks to the deposition in the PDB of new structures in complex with cognate ligands.

We gathered 22 total targets in complex with biological ligands (table 9; in May 2016 there were still 5 undeposited 3D structures and 8 non-crystallised proteins). Two of them (T0745 and T0754) were canceled during the experiment because one group found the crystal structures before the deadline and used this information to make perfect binding site predictions. At the end we were able to include 7 additional targets that correspond to an expansion by more than 50% of the assessment data set. We compared our calculated contacts with those obtained by official assessors (appendix table 3) and we found some differences, despite using the same distance cut-off of 0.5 Å. Probably we

used slightly different values of Van der Waals radii, since there is no complete agreement about the values and their calculation^{170,171}.

Target	PDB ID	Ligand ID	Contact residues	Neutrals
T0652	4HG0	AMP	74,78,79,80,100,101,102,103,104,165,178,180,183	
T0657	2LUL	<u>ZN</u>	121,131,132,133,143	13,15,16,17,18,19,21,22,23,25,27,29,40,42,51
T0659	4ESN	<u>ZN</u>	43,48,53	
T0661	5UWB	PEF	29,30,33,34,37,59,60,63,64,67,72,97,98,101,105,108,109,112,114,134,136,138,139,144,146,148,159,162,164,165,169	
T0675	2LV2	<u>ZN</u>	21,24,37,42,49,52,65,70	
T0682	4JQ6	RET	80,83,84,87,116,120,134,137,138,141,180,183,184,187,209,213	
T0686*	4HQ0	<u>MG</u>	28,30,103	26,134
T0687*	4HQF	<u>MG</u>	31,33,106	29,137
T0694	5JH8	HIS	5,74,109,111,172,174,176,295	177,180,181,217,219,229
T0696	4RT5	<u>NI</u>	18,69,104	
T0697	4RIT	LLP	91,150,151,152,190,243,245,247,272,274,301,303,304	
T0706*	4RCK	<u>MG</u>	25,27,129	23,101
T0715	4C3S	NAD	133,134,135,136,137,138,161,162,163,198,201,205,216,217,218,221,224,235,236,237,269,357,359,387,432,433	
T0720	4IC1	SF4, <u>MN</u>	32,34,35,62,99,113,114,115,187,188,191,194,197	
T0721	4FK1	FAD, <u>MG</u>	10,12,13,14,33,34,35,36,37,38,39,41,42,44,45,46,60,78,79,80,109,110,111,114,126,127,136,137,235,237,268,269,277,278,281	
T0726	4FGM	<u>ZN</u>	273,277,307	39,40,41,95,236,237,238,362,363
T0732	4PEG	5GP	45,90,91,180,201,202,205,230,231,232	50,95,99,146
T0737	3TD7	FAD	37,38,40,41,42,44,45,49,78,83,87,114,117,118,120,121,123,124,128,130,135,174,237	80,82
T0738	4IS3	NAD	13,15,16,17,18,38,42,66,92,93,94,95,115,142,143,144,157,161,187,188,189,190,192,194,195	
T0744	2YMV	FNR	22,23,24,26,58,61,120,121,122,124,196,214,216,255,270,271,272,273,314,316	
T0745	4FMW	SAH	96,97,98,115,116,118,120,124,127,128,141,142,144,154,155,156,158,161	121
T0754	2LV9	<u>ZN</u>	13,15,27,30,35,38,52,55,69	

Table 9 Complete list of the targets binding biological ligands in CASP10 used for the re-assessment. Contacting amino acids are those below a distance cut-off of 0.5 Angstroms + VdW radii. Neutrals are established using literature information available, homology with other proteins and conservation. Green targets are new, released after (and so not previously evaluated by) the official assessment. Red targets are those cancelled by organizers. The targets marked with an asterisk were close homologues that bound the same ligand. The neutral residues in these targets are those that are within the van der Waals radius cut-off in one or two targets, but not in the others.

Using this new data we performed a new assessment and new MCC scores have been calculated for all the groups (appendix table 4). Detailed results for *firestar* predictions are presented in table 10. Based on our binding site definition, we clearly obtained a better mean MCC, (0.817 vs 0.715). Differences with the previous MCC results

can be explained by the change of the amino acid composition of the binding sites and/or by the introduction of the neutral residues. Several targets fall into the first group where the substitution, introduction or exclusion of some residues made the score change slightly: T0652, T0697, T0720, T0721, T0737 and T0744. Among these, the biggest *firestar* variation is a 0.076 increase for target T0652.

Target ID	Ligands ID	Site size	FP	TP	Neutrals	MCC
T0652	AMP	13	3	13	-	0.897
T0657	<u>ZN</u>	5(15)	1	4	15	0.793
T0659	<u>ZN</u>	3	-	-	-	-
T0661	PEF	31	4	15	-	0.572
T0675	<u>ZN</u>	8	0	7	-	0.929
T0682	RET	16	7	16	-	0.821
T0686	<u>MG</u>	3(2)	0	3	2	1.000
T0687	<u>MG</u>	3(2)	0	3	2	1.000
T0694	HIS	8(6)	2	5	3	0.660
T0696	<u>NI</u>	3	0	3	-	1.000
T0697	LLP	13	3	12	-	0,855
T0706	<u>MG</u>	3(2)	0	3	2	1.000
T0715	NAD	26	4	10	-	0.505
T0720	SF4, <u>MN</u>	13	0	4	-	0.542
T0721	FAD, <u>MG</u>	35	5	25	-	0.744
T0726	<u>ZN</u>	3(9)	0	3	-	1.000
T0732	5GP	10(4)	2	6	0	0.662
T0737	FAD	23(2)	1	17	1	0.824
T0738	NAD	25	2	23	-	0.911
T0744	FNR	20	2	15	-	0.803
Mean MCC						0.817
Std. Deviation						0.164

Table 10 Results of the new assessment over 20 CASP10 targets for *firestar*, using the new definition of the binding sites shown in table 9. Bold IDs correspond to targets used in the official ligand-binding category assessment. The table shows the crystallised ligand (metals are underlined), the size of the binding site with additional neutral residues (if any), true and neutral (TP and neutrals) residues identified and wrongly predicted residues (FP). The MCC scores are highlighted in green if they improved with respect to official assessment, blue if they are the same and red if they worsen.

The introduction of neutral residues had a major impact over the MCC scores of all groups, and we are going to present here 2 cases. Target T0657, the pleckstrin homology (PH) domain from human a tyrosine-protein kinase (*TEC*), is one interesting example: all groups sent a prediction, two groups predicted just a zinc binding site and the other 17 predicted a binding site for inositol-tetrakisphosphate, either on its own in addition to a zinc residue. The resolved structure (PDB code 2LUL, an NMR ensemble) contained only the structural ZN, required to maintain a loop. A functional pleckstrin domain is located at the beginning of the sequence, and most PH homology domains bind phosphatidylinositol lipids from membranes as part of a process to recruit proteins to the membrane. A simple BLAST search shows that the inositol polyphosphate binding residues in the PH domain are conserved and in fact the Swiss-Prot entry for *TEC* (P42680) annotates the PH domain of this protein as mediating binding to inositol polyphosphate in the plasma

membrane. *firestar* and other predictions could be correct, but the final NMR ensemble does not include an inositol polyphosphate ligand (in fact NMR structures are rarely resolved with any ligands), and so far we can not confirm the prediction experimentally. For this reason we introduced 15 neutral residues to represent the inositol polyphosphate binding site. Example 2, target T0726, is an aminopeptidase of the N family. It contains 2 domains, PDZ and Peptidase_M61 catalytic domain. While the first is involved in the binding of specific peptide sequences, the second, the MEROPS peptidase family M61 usually has a divalent cation, often zinc but also cobalt, manganese or copper, activates the water molecule to hydrolyze a peptide bond. Again, all groups sent a prediction: many made predictions for an amino acid binding pocket as well as a zinc ion (17 of 19, but not *firestar*), and once again only a ZN cation was present in the crystal (the amino acid is a reagent). For this reason, using the information available, we selected and added 9 neutral residues. We also performed a Wilcoxon signed rank test to evaluate the statistical significance of the new results (appendix table 5), and results are very different from the official assessment. The majority of p-Values obtained are significant (because of the growth in the test set), showing that there are clear differences in accuracy between groups.

4.5.6.2 Evaluating problematic targets

According to our results, *firestar* performed poorly (MCC below 0.7) for six targets: T0659, T0661, T0694, T0715, T0720 and T0732. We revised these predictions in order to identify possible improvements and limitations of our algorithm. In table 11 we show how many servers predicted the sites and which was the best MCC obtained.

Target ID	Ligands ID	Site size	<i>firestar</i> MCC	Server Predictions	Best MCC
T0659	<u>ZN</u>	3	-	4/11	0.455
T0661	PEF	31	0.572	6/11	0.623
T0694	6KY-HIS	8(6)	0.660	11/11	0.797
T0715	NAD	26	0.505	10/11	0.866
T0720	SF4, <u>MN</u>	13	0.542	8/11	0.608
T0732	5GP	10(4)	0.662	10/11	0.665

Table 11 List of targets where *firestar* algorithm obtained a MCC < 0.7. Next to *firestar* MCC, the total number of server groups that sent a prediction and the overall best MCC are reported.

We present here the analysis of two target predictions, T0659 and T0720. They are particularly interesting for us because they are metal binding sites. As we detailed previously, this type of sites present higher conservation due to their size and their specific features. *firestar* usually performs very well: over the other 7 metal binding targets it had a 0.960 mean MCC.

T0659 (PDB code 4ESN) is a hypothetical protein from *Ruminococcus gnavus* (PDB code 4ESN) that contains a novel ZN binding site, constituted by three cysteines. While no sites pass the reliability threshold we established for CASP, actually *firestar* detected the conserved cysteines. The ZN binding site has either diverged from a FE2/S2 inorganic cluster (FES) binding site or converged on the same residues as figure 28 shows. Almost all templates found by *firestar* bound FES using the same ligand binding residues that formed the novel ZN in T0659. Two of the three cysteines in the ZN binding site were very conserved while the third cysteine only appeared in a few alignments. However, *firestar* did not predict residues since the binding site was treated as an FES

binding site. The average size of the FES binding sites is 9 residues, and *firestar* discarded the binding residues detected in those alignments (figure 28) because the coverage of the FES binding residues was 33% or lower. A human predictor would have noticed the FES conservation footprint using the information generated by *firestar* possibly made a prediction, but target T0659 was a server-only target.

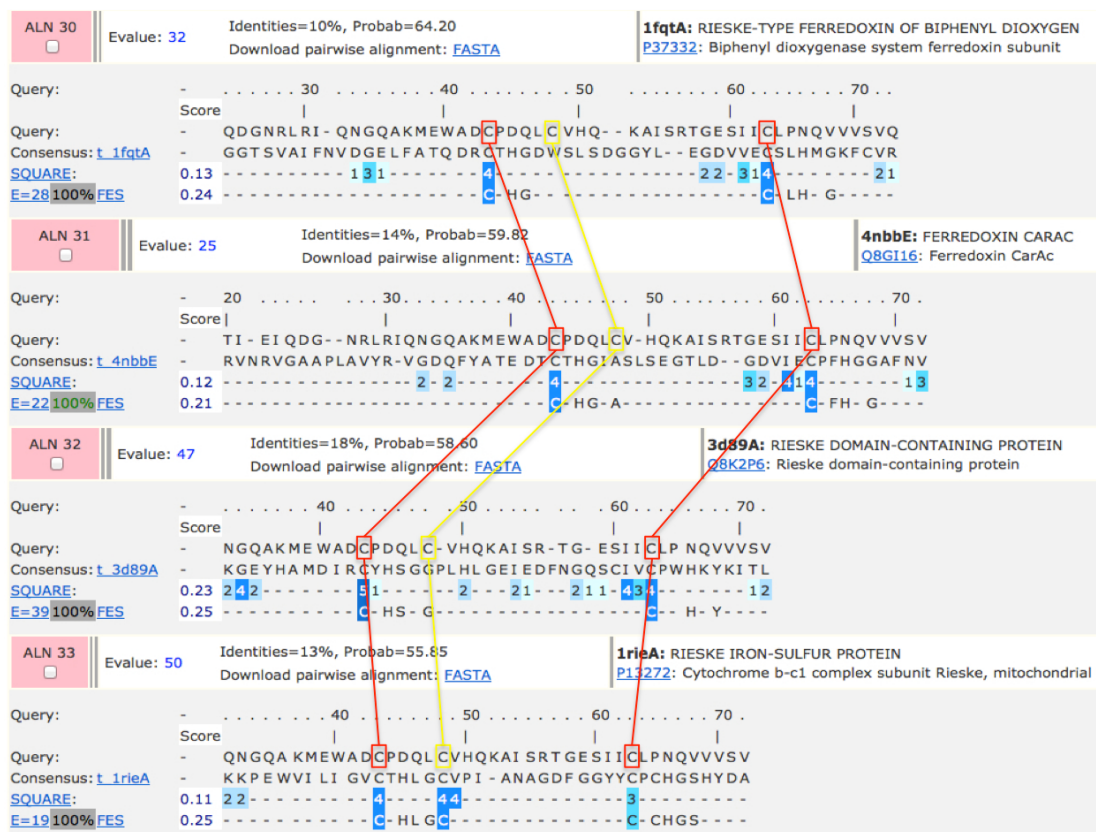


Figure 28 Extended output from the analysis of target T0659 from the CASP10 experiment. These are the templates found by for the target sequence, first line of each alignment. The three cysteines that constitute the ZN binding site in T0659 are shown in the red and yellow boxes. All the aligned templates bind FES (FE2/S2 cluster). Cysteines 43 and 63 (red highlights) are conserved in all the alignments, while the position corresponding to cysteine 48 (yellow highlights) binds FES in the majority of templates, but it is usually a different amino acid.

These results suggest that this could be a case of divergent evolution: the binding site specificity drifted and lost the ability to bind iron-sulfur, while maintaining the capability to bind a metal. Or it could be exactly the opposite: it was originally a ZN binding site that gained the ability to bind iron-sulfur. Thanks to this target we decided to work on a possible improvement of the algorithm. When it is not able to generate a canonical consensus prediction but detects conservation for binding site residues in complex with biologically relevant compounds, *firestar* should now report these residues without predicting the possible ligand.

The other interesting case is T0720, a CRISPR-associated exonuclease Cas4 from *Sulfolobus solfataricus*. The protein was co-crystallised with manganese and an iron-sulfur cluster (SF4), and it had very few remote homologs in the PDB (only 9 unique hits from the HHsearch/PSI-BLAST analyses had functional information). While the conserved manganese binding site was predicted, missing only histidine 62, information for the other site was insufficient and noisy and the filters discarded all the templates. Furthermore the

site was split, meaning that 3 of the 10 binding residues are around sequence position 35 while the rest are located around position 182: no alignment spanned all 202 residues of the target. This was a difficult prediction; in fact *firestar* had the 3rd best server result. In our human prediction, using *firestar* server output, we included four cysteines of the SF4 binding site based on the conservation detected in the extended *firestar* output pages, as shown in figure 29.



Figure 29 Extended output from the analysis of target T0720. These are 2 template fragments found only in the HHsearch analysis that support functional information for three of four cysteines included in our human prediction. Even though both bind SF4 (FE4/S4 cluster), only part of the site is conserved.

The other 4 problematic cases are similar: these proteins contain large binding sites, and the information coming from FireDB is insufficient or noisy. One good example is T0715: even though there are many templates containing NAD in the PDB, no close homolog was found. The occurrence and conservation filters discarded all but the core binding residues, so *firestar* lost more than 50% of the site.

All servers had the same problem with T0732: few remote templates were rescued, the majority bound adenosine monophosphate (AMP). 5GP binding proteins were also present, but insufficient to allow *firestar* to focus on the specific functional residues.

For T0694 and T0661 no templates containing the bound ligands were found. In fact *firestar* was not able to predict the compound in the binding site and prediction was based only on residue conservation.

4.5.6.3 False positives

One of the problems we detected for *firestar* in the CASP8 experiment and even more after the introduction of HHsearch as search method was the increase of false positive predicted residues. In table 12 we listed for all participant groups the number of predicted targets over our overall dataset (20 targets) and the total number of false and true positive residues (FP and TP). Only one human group predicted fewer than 30 FP, but the group missed 4 targets. After that, 3 servers accumulated 36-37 FP each, and among these *firestar* was the server that predicted most targets.

These simple statistics, although realized over a small dataset, suggest that the filters introduced are able to effectively discriminate true negative residues, without losing sensitivity, since the algorithm is still able to identify a good number of true positive residues.

Group ID	Target predicted	TOT FP	TOT TP
McGuffin	16	27	151
<i>firestar</i>	19	36	187
IntFOLD2	17	36	155
3DLigandSite	12	37	126
zhang	20	48	196
CNIO	19	48	199
Binding_Kihara	19	53	82
Seok-server	19	58	151
I-TASSER_FUNCTION	20	59	188
Seok	18	59	163
COFACTOR_human	19	72	174
FNGUSHAK	18	75	166
COFACTOR	20	78	170
HHpredA	18	81	173
SP-ALIGN	20	89	206
Atome2_CBS	11	90	151
3DLigandSite2	14	91	136
ConPred-UCL	18	122	160
chuo-binding-sites	19	913	217

Table 12 The list of the Groups that participated in the CASP10 experiment, with the number of predicted targets over the assessment group (20). In the 2 columns the cumulative number of false and true positive predicted residues is shown. Automatic server groups row are highlighted in light blue. No fill rows represent human groups.

There is an additional aspect that was not considered in the CASP assessment and that is incorrectly predicted targets. Predictions are sent for the whole target dataset (114 proteins in CASP10) during the assessment, and “No predicted binding site” is actually a prediction. This prediction is not the same as not sending any prediction. It is true that the assessment of predictions for targets with no ligand would be complicated, but where possible, it would be interesting to estimate the tendency of methods to over-predict or to predict a non biologically relevant site. One example we found in our assessment is the case of T0710 (PDB codes: 5CEA), Bd3460 immunity protein from the predatory antimicrobial organisms *Bdellovibrio bacteriovorus*. This structure is actually one among 5 depositions, and in one of them a sulfate (SO₄) anion is included. It is clearly an artifact due to the presence of the molecule in the crystallization mix. Among the 18 predictions sent for this target, 15 predicted a binding site, and 11 of them wrongly predicted a binding site for a sulfate, a metal or a solvent. *firestar* did not predict any binding site.

At present the CASP assessment penalizes under-prediction and incorrect predictions, but there is no penalty for over-prediction. This means that over-predicting (either by predicting non-cognate ligands or by predicting binding clefts) is good strategy where it appears that there is no bound ligand. *firestar* does not over-predict: in fact *firestar* predicts ligand binding sites for just 44 of the 114 targets because it does not detect sufficient information to make a reliable prediction for the remaining targets. This is a strategy that the CASP assessors refer as cherry picking. By way of contrast the two next highest scoring server methods, SP-ALIGN and ITASSER make predictions for all targets regardless of ligand type, including making predictions for residues that bind

glycerol and ethylene glycol, both well known solvents. Obviously over-prediction should also be taken into account in future assessments.

4.6 Applications in large-scale collaborative projects

One of the objectives after the improvements introduced in both tools and the validation of the *firestar* method was their use in large-scale annotation projects.

4.6.1 Human proteome sites annotation and selection of gene principal isoform

As part of the ENCODE^{172,173} project, GENCODE¹⁶⁷ provides high-accuracy manual annotations of protein-coding loci and alternative variants in the human genome. Studies have revealed that virtually all multi-exon human genes¹⁷⁴ are capable of producing multiple RNA transcripts by alternative splicing. Alternative splicing events that occur within coding regions will produce alternative transcripts that potentially will be translated into distinct gene products. While genome annotation projects are producing rapidly a huge amount of information, this data presents serious challenges for functional annotation, as we explained in detail in this work. If alternative splicing does have the potential to expand the cellular functional repertoire in eukaryotic species, it would seem to be important to assign roles to these splicing variants.

Another important task is to identify the representative (or principal) isoform of the gene, the isoform against which all others should be compared, from among these possible alternative transcripts. The selection of the principal isoform is not straightforward, since there is no agreement about its specific characteristics. Over the years, databases such as Ensembl¹⁷⁵ and SwissProt have got round this problem simply selecting the longest isoform as the main variant. Although this is a convenient choice, Tress and collaborators demonstrated in a work¹⁷⁶ that longest isoform is not the best choice for up to ~25% of the genes and proposed a methodology to pinpoint principal functional isoforms, based on conservation and the characteristics of known proteins, principally structural and functional features. Using an experimental set of 215 human proteins, they determined a principal variant for 179 of them, 83% of genes with multiple alternative variants. *firestar* and FireDB were included in these initial investigations to provide reliable functional residues predictions.

This work provided the basis of the creation of APPRIS¹⁷⁷, a database that houses annotations of splice isoforms for different organisms, human included. It was designed to provide value to manual annotations of the human genome by adding reliable protein structural and functional data and information from cross-species conservation. Developed alongside the GENCODE annotation process, it flags isoforms with likely altered structure, function or localization, and exons that are evolving unusually. The information from APPRIS is fed back to the manual annotators and has led to the annotation of new isoforms. As additional feature, it selects (whenever it is possible) a principal isoform for each gene based on the available annotations.

The database flowchart is presented in figure 30.

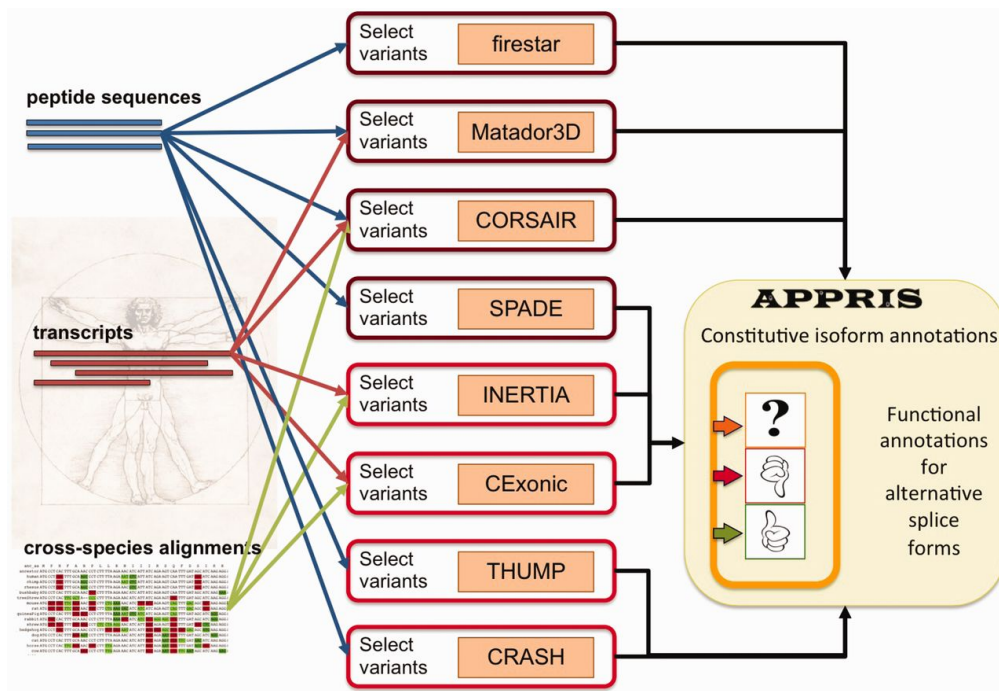


Figure 30 On the left the three types of input information are shown. They are used individually or in combination to feed the 8 methods (center) that produce annotations for whole GENCODE dataset. The information is used for the selection of the principal isoform. Information coming from the four core methods (brown boxes) is fundamental to principal isoform selection; the other methods (red boxes) is more important where these four methods are not able to make a decision (image extracted from the APPRIS paper¹⁷⁷)

The APPRIS system is composed of eight separate annotation modules, chosen for their ability to highlight important features to spot principal isoforms. Determination of a principal isoform is based on two principles. The first is that there is often one isoform that performs the main cellular function or that is expressed in the majority tissues or in most stages of development, and that the rest of the annotated isoforms are alternatively spliced isoforms that may perform distinct roles. The second principle is that the principal isoform should have more evolutionary history, so it ought to be the variant that is most conserved across related species. Selection is based on a jury of these 8 methods. The isoform selected as principal will be either the variant that has the most conserved protein features (since it is much more likely that alternative isoforms have lost rather than gained protein features such as 3D structure and function) or that has more evidence of cross-species conservation, or, most frequently, both. Four methods (SPADE, CORSAIR, Matador3D and *firestar*) make up the core of the jury system, with the other methods becoming more important in cases where these four methods are not able to make a decision. APPRIS is updated with each new stable GENCODE release.


4.6.1.1 *firestar* annotations within the APPRIS database

The most recent version of the APPRIS database covers the Ensembl_88/GENCODE_26 release of the human genome (assembly hg38). GENCODE annotates 20,263 protein-coding genes that are predicted to generate a total of 95,139 coding transcripts through alternative splicing. Among these genes, 22.72% (4,603) are annotated with a unique transcript/isoform, meaning that they do not have alternatively spliced transcripts. The rest (77.28%, 15,660 genes) are predicted to produce at least one alternative isoform.

firestar predicted a biologically relevant binding site for 7,935 genes (39.16%). Among the APPRIS jury core methods, *firestar* has the lowest coverage (matador3D: 72.15%, SPADE: 83.94%, CORSAIR: 87.47%). In the case of *firestar* the “winning” isoform is the one with the highest number of functional residues. It is also usually selected by APPRIS to be the principal isoform. In fact there are very few cases where *firestar* predictions do not agree with the final APPRIS principal isoform selection. We found 70 genes, 0.3% over the whole GENCODE dataset (no prediction is a prediction for APPRIS purposes). Among the jury core methods, it has the lowest disagreement rate (matador3D: 0.56%, SPADE: 0.85%, CORSAIR: 0.83%).

Looking in more detail at these 70 cases, there are 43 where *firestar* disagrees with all others core methods. The reasons for this can be different; this is not only related to problems with predictions: for instance some annotated transcripts are larger due to a read-through event (defects in transcription termination). These read-through transcripts are not allowed to be APPRIS principal isoforms because they are generally not translated¹⁷⁸ but if the read-through exons are from the neighbouring coding gene, *firestar* may predict more predicted functional residues for these transcripts than for the principal variant.

Other cases are also interesting, since there is a general agreement between the methods in APPRIS. These are potential candidates for manual annotators to focus on, and we present one example here.

Id	Name	Biotype	Species	Assembly	Location
ENSG00000273154	 RP4-583P15.15	protein_coding	Homo sapiens	GRCh38	20:63708864-63739103

Seq. id	Seq. name	Length (aa)	Biotype	CCDS	Flags	Principal Isoform
ENST00000490623	RP4-583P15.15-001	417	nonsense_mediated_decay	-	TSL5_start_codon_NF	MINOR
ENST00000496820	RP4-583P15.15-003	169	protein_coding	-	TSL5	MINOR
ENST00000632538	RP4-583P15.15-004	290	protein_coding	-	TSL3_start_codon_NF	PRINCIPAL:1

<input checked="" type="checkbox"/> All / <input type="checkbox"/> None	Seq. id	Seq. name	Length (aa)	No. Functional Residues	3D Structure Score	Domains score	Conservation score	No. Transmembrane Helices	Signal Sequence	No. Mapping Peptides
<input checked="" type="checkbox"/>	ENST00000490623	RP4-583P15.15-001	417	4	3	43.2	0	0	-	-
<input checked="" type="checkbox"/>	ENST00000496820	RP4-583P15.15-003	169	0	0	0	0	0	-	-
<input checked="" type="checkbox"/>	ENST00000632538	RP4-583P15.15-004	290	0	2	0	0	0	-	-

Figure 31 Screenshot of APPRIS web page showing results for gene RP4-583P15.15 General information from Ensembl is shown for the gene; different annotated transcripts are shown in the Principal Isoforms table, where the principal isoform is highlighted in green. In the APPRIS annotations table the per-transcripts results for all integrated methods are summed up. In the web viewer it is possible to browse individual annotations.

RP4-583P15.15 is a human gene that codes for a so far functionally uncharacterized protein. There are three annotated transcripts in GENCODE version 26, and almost no annotations are retrieved by APPRIS for any of them, except for transcript ENST00000490623. This transcript has three supporting annotations (see figure 31): a predicted ZN binding site (4 residues), three segments that match to deposited structures and a Pfam correspondence. However, the transcript has been tagged as nonsense_mediated_decay, meaning that it should be involved in regulation of gene expression, and it should not be translated. So the selected transcript is ENST00000632538, which presents only support from structural information. The nonsense_mediated_decay tag is probably erroneous.

4.6.2 Pfam domain analysis

Pfam is a long established database that gathers information about protein domains. Proteins can be considered as a combination of one or more functional regions, commonly termed domains. The identification of domains that occur within proteins can therefore provide insights into their function. Pfam's central entity is the family, a set of protein regions that share a significant degree of sequence similarity, thereby suggesting homology. A high-quality manually curated alignment, called a seed alignment, is created and a profile hidden Markov model (HMM) is constructed from the seed alignment. At a higher level of hierarchy, clans are collections of related families.

A clan can contain two or more Pfam families that are considered to have arisen from a single evolutionary origin. Four pieces of evidence are used to assess whether families are related: structural, functional, significant matching of the same sequence to HMMs from different families and profile-profile comparisons; the presence of related structures and significant profile-profile comparison scores are primary indicators of a relationship. So in principle it is possible for clans to contain families with heterogeneous functions, but discrepancies have to be evaluated since they can be suggestive of a non-homologous origin.

In order to evaluate the functional uniformity of the clans from a binding site point of view, we decided to use *firestar* and FireDB to annotate all 14,831 families contained in the Pfam-A database (version 27.0). After different trials, we decided to use three different approaches:

- Using the pre-calculated mappings between Pfam families and PDB structures, we extracted all the MSS binding sites from FireDB that overlap Pfam families.
- From the downloaded seed alignments, we selected a sequence randomly and used it as input in *firestar*
- Again, from the downloaded seed alignments, we generated a profile to feed a modified version of *firestar*

After merging FireDB and *firestar* predictions, we obtained a COGNATE binding site prediction for 4,190 non-redundant Pfam domains, 28% of the entire Pfam-A data set. In terms of concordance, the two approaches do not always overlap: *firestar* detects sites that the FireDB-Pfam mapping does not, and this is something to be expected because *firestar* can extend FireDB binding sites to homologous sequences. Surprisingly, the FireDB-Pfam mapping detects sites that *firestar* did not. One reason was heterogeneous conservation of the residues through all the members of the seed that affected the profile. Another reason was the random selection of the single seed alignment sequence to launch *firestar*, due again to heterogeneity, if the site was not present, we did not detect it. Another reason for the difference was the *firestar* filters, which discarded sites because they were considered non-biologically relevant. Most Pfams that did not have a *firestar* prediction are DUFs (domains of unknown function), protein-protein binding, repeats or trans-membrane domains. Among non-predicted domains, the Sugar_tr family (PF00083.19) represents an interesting example of to present a false negative caused by lack of source information. This family gathers sugar transporters, which are responsible for the binding and transport of various carbohydrates, organic alcohols, and acids in a wide range of prokaryotic and eukaryotic organisms. Unfortunately PDB contains almost no transporters structures in complex with their biological ligand, and for this reason *firestar* is not able to generate a prediction. When we took into account all predictions, the number of Pfam domains with binding sites rises to 5,392 (36%).

We then evaluated the consistency of the binding site (and especially the ligand) predictions among families belonging to the same clan. In some clans we found a complete homogeneity: we predicted an ANP (Phosphoaminophosphonic Acid-Adenylate Ester), ADP or ATP binding site for all 19 members of the ATP- grasp clan (CL0179). We

detected an iron-sulfur cluster binding site for all 22 members of the 4Fe-4S clan (CL0344), and a copper binding site for all 9 members of the Multicopper oxidase-like domain clan (CL0026).

For other clans we found discrepancies in our predictions: the methods predicted a coenzyme binding site for 163 of the 178 members of the NADP_Rossmann (CL0063) clan. They detected a calcium binding site for 12 of the 14 domains in the EF_hand (CL0220) clan. The other two domains are tagged as “efhand-like” and “Ca²⁺ insensitive EF hand”. We also validated cognate binding sites for 174 of the 178 members of the Peptidase_MH (CL0035) clan.

As part of a collaboration with the Pfam group, we presented them our preliminary results and based on the data, they systematically reanalyzed non-consistent clans: here we present the case of the P-loop_NTPase (CL0023) clan. This is a large clan that gathers AAA+ family proteins. These NTPases contain chaperone-like modules that appear to function as molecular matchmakers in the assembly, operation, and the disassembly of diverse cellular proteic machineries. We predicted a cognate ligand binding site for 178 of the 195 members, so they decided to focus on the 17 families with no prediction, and established a workflow to systematically evaluate the families:

- Search of Walker A and B motif (two characteristic motifs of the clan)
- Search for overlaps with other families within and outside the clan
- When a family structure was available, structural-pairwise alignment comparison, visual assessment against known structures in clan and comparison with SCOP entries
- Literature mining

Finally they decided to maintain 10 out of 17 families in the clan because structural information and/or overall sequence similarity suggested homology and fulfilled their aggregation criteria. Families without structures, overlaps, Walker A or B motifs (degraded or otherwise) were considered on a case-by-case basis: 3 were tagged as uncertain, 3 were definitely removed from the clan and the last one (PF04326) was removed and added to AlbA clan and renamed as AlbA_2.

These preliminary results suggest that binding site information can be really informative in the definition of protein function and can be used as an additional feature in the definition of Pfam families and clans.

4.6.3 GO terms prediction for large scale annotation projects

The huge amount of new sequences obtained yearly has caused the exponential growth of sequence databases, but functional annotation of genes and their products has so far not been able to keep up this incredible pace. Using an already established approach, a group of researchers from the function prediction community joined and organized the first CAFA experiment. The Critical Assessment of protein Function Annotation algorithms is designed to provide a large-scale assessment of computational methods dedicated to predicting protein function. The general set-up can be easily visualized in figure 32, extracted from the assessment paper¹⁴² published after the first edition.

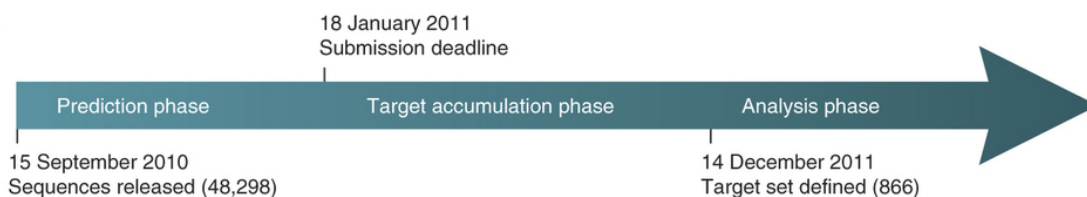


Figure 32 Timeline for the first edition of the CAFA experiment extracted from the Nature paper¹⁴² published in 2013.

Briefly, CAFA organizers provide a large number of almost unannotated protein sequences and give a deadline date. During this time (the prediction phase) participating groups predict the function of these proteins by associating them with Gene Ontology terms. In the following phase, target accumulation, the assessors gather experimental functional evidence for the target dataset. The prolonged duration of this phase (almost a year in the first experiment) is meant to give time for the scientific community to generate as many annotations as possible. Even so, it was possible to retrieve information for only the 0,01% of the initial set for the first experiment. Finally, in the analysis phase, methods are tested against the established benchmark set.

The second edition of the experiment started on the 29th of August 2013. The whole protein dataset was constituted by 102,117 sequences. A small part of them (1,301) came from a large-scale collaborative project called the Enzyme Function Initiative, or EFI, whose goal is to develop integrated strategies that will enable focused experimental enzymology, genetics, and metabolomics and was constituted by putative enzymes. The organizers gathered the rest, more than 100 thousands proteins picking them from 27 different organisms.

We decided to participate with a modified version of *firestar*. In principle the method is not able to predict GO terms directly; FireDB stores functional annotation associated to PDB entries coming from UniProt, so *firestar* could transfer this information from the different templates used to generate the prediction to the target. But while it is true that functionally important residues can be found in very diverse proteins, it is not true that all of them are equally relevant to determine function. To overcome these limitations we used two different approaches.

In the first one we extracted from the Molecular function GO domain terms related to ligand binding and we associated them to the correspondent PDB molecules. For example Beta-lactose (PDB code LAT) is associated to the general GO:0005529 term (carbohydrate binding) and to the more specific GO:0030395 term (lactose binding) and in total we were able to annotate 112 compounds. Using this information, whenever *firestar* predicts a binding site and the correspondent interacting ligand, it automatically transfers the associated GO terms to the target protein.

In the second approach we used the mapping generated from the Gene Ontology consortium itself (<http://www.geneontology.org/external2go/ec2go>) between Enzyme Codes and GO terms. Basically using the Catalytic Site Atlas information stored in FireDB, *firestar* is able to predict catalytic sites and at the same time, through the mapping, it can also transfer the correspondent GO terms. To generate GO predictions we decided to use information coming from manually annotated CSA entries and we set a more restrictive conservation and coverage filters. If a catalytic site is fully conserved and the SQUARE scores of the single residues are higher than an established cut-off, the GO term(s) is directly transferred. If a third of the site is poorly conserved while the rest is highly conserved, the parental GO term(s) is transferred, while if more than a third is not conserved at all, it is directly discarded.

Our lab also participated to this edition of the experiment with the Statistically Inferred Annotation Method, or SIAM, developed by Angela del Pozo (manuscript in preparation). Briefly the algorithm searches for sequence homologs of the target protein in the Swiss-Prot database. Using a non-parametric statistical coefficient of concordance,

the set of the functional annotations (GO terms) that better fit the pool of homologs found are transferred to the target.

In principle the two methods do not overlap; *firestar* generates specific annotations, related to binding or/and to the catalytic activity and all the terms basically come from the “Molecular function” gene ontology domain. Since the source of information for SIAM is SwissProt, terms can come from the three domains and in general they are expected to be less specific. For these reasons, a third set of predictions was submitted, coming from the integration of the previous two.

In figure 33 general statistics for *firestar* and SIAM results in the CAFA2 experiment dataset are presented.

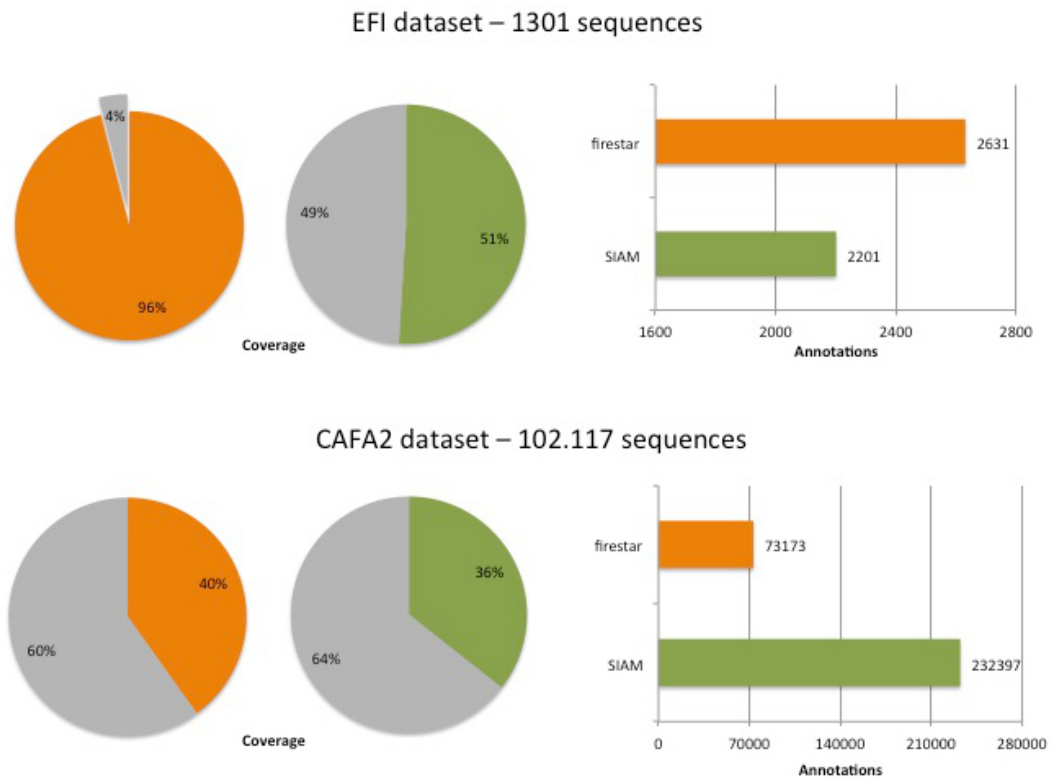


Figure 33 Overview of the predictions for the CAFA2 experiment. The pie charts refer to the coverage of the methods, orange for *firestar* and green for SIAM; grey portion refers to unannotated sequences. Bar charts compare the total number of annotations generated.

Considering only the limited EFI dataset, *firestar* was able to generate a prediction for all but 52 sequences, with an average of two annotations per sequence. SIAM coverage is worse, but globally the method was able to transfer more GO terms per sequence. Looking at the entire dataset, the coverage of *firestar* is again slightly better, but the number of annotations transferred by SIAM was more than three times greater.

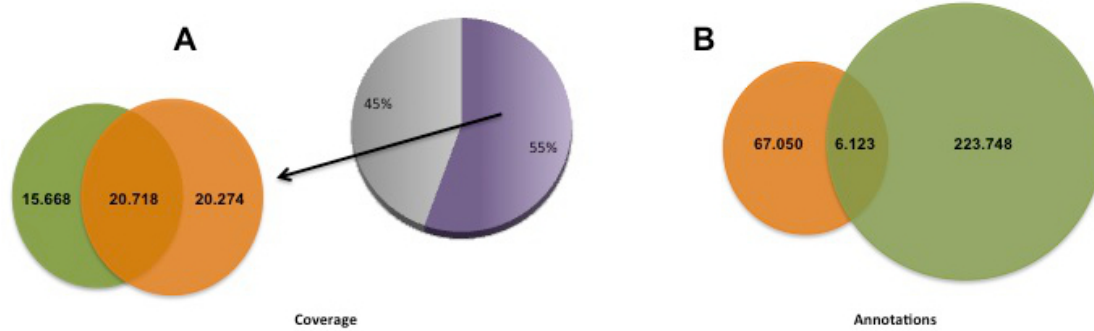


Figure 34 Overview of the integration of *firestar* and SIAM predictions for the CAFA2 experiment. **A)** The global coverage of the combination is shown on the right, while the overlap between the two methods (orange for *firestar*, green for SIAM) is presented on the left side **B)** The summary of the global number of annotations: in the intersection the number of the sequences of SIAM (green circle) refined by *firestar* is shown.

The two methods combined were able to generate predictions for 55% of the targets; among these, almost 37% have at least one GO term assigned by both algorithms (figure 34). Looking at the annotations the reduced intersection (compared with the predicted targets overlap) indicates clearly that SIAM and *firestar* are strongly complementary. In order to merge overlapping predictions, a set of rules has been set up, based on the hierarchy of the GO term database structure:

1. When *firestar* predicts a more specific term than SIAM, it is directly added.
2. When *firestar* predict a less specific term, and it comes from an incomplete catalytic site prediction, the more specific SIAM term is eliminated.
3. When *firestar* discards a GO term due to not detected conservation, SIAM term is discarded.

This strategy resulted in a scenario where, among 6,123 annotations generated for the same target (figure 34 B, intersection), 95% of them involve *firestar* predicting a more specific term, so *firestar* was able to improve the specificity of SIAM predictions. In the remaining cases, *firestar* refined the SIAM prediction. There were also 2,526 SIAM predicted and 50 *firestar* predicted GO terms that were discarded on the basis of the rules presented (not shown in the figure). So preliminary results presented here suggest that the integration is useful.

5 DISCUSSION

In this work we presented improvements of two tools previously developed in the CNIO for the prediction of functional residues, FireDB and *firestar*, and three practical applications in large-scale annotation projects.

FireDB is a curated inventory of catalytic and biologically relevant small ligand-binding residues culled from the protein structures in the Protein Data Bank (PDB). A great deal of manual curation has gone into the annotation of the small-molecules contained in the database¹⁷⁹. Ligands in the most recent version of FireDB are classified according to a range of criteria. The introduction of biological relevance and the metal/non-metal categorization has changed the construction of the database, and the bio-activity annotation and external mapping of the entire ligand molecule set has expanded the ambit of FireDB.

The automatic biological relevance assessment of the binding sites in FireDB is fundamental to the database. We introduced a new classification protocol based on the local conservation of the individual sites and on ligand type that works alongside the original biological relevance classification. In depth analysis of differences between the two sets of annotations showed that the new protocol complements the original approach.

Finally a number of changes have been made to improve database performance, usability and accessibility.

firestar is a server for predicting catalytic and ligand-binding residues in protein sequences. It is based on information from the FireDB catalog. The earliest versions of *firestar* required expert human interpretation of the results; we implemented a new schema to achieve the full automatization of the server and to improve its accessibility to non-expert users¹⁸⁰.

The sensitivity of *firestar* has also been improved with the integration of an additional state of the art homology search method, HHsearch, to go along with PSI-BLAST.

Data collected during the different rounds of the CASP experiments showed a slight tendency to over-prediction, magnified by the inclusion of HHsearch. As a result new filters have been added to improve *firestar* specificity, ranging from compound specific rules to per-residue conservation within candidate templates. The filters have been tested against a CASP dataset, and show a good improvement in true positive detection along with a drastic reduction of false positives.

The improved algorithm was tested in the CASP10 experiment, and *firestar* was the highest rated functional residue prediction method among those tested. We have reassessed the CASP10 results, including new data made available after the official CASP assessment, to pinpoint limitations and strong points of the algorithm.

The two tools have been used in three different large-scale annotation projects. They have been permanently included in the APPRIS database¹⁷⁷, a database of splice isoform annotations developed as part of the GENCODE project. FireDB/*firestar* provide high accuracy annotations for APPRIS and this functional information is an important part of the selection process of principal isoforms.

The results of annotating Pfam domains with *firestar* and FireDB functional residue information shows that this information can be used to disambiguate the evolutionary relationships of protein families.

Some FireDB information has been matched with GO terms via annotation derived from the Catalytic Site Atlas and by associating compounds to specific binding related GO terms. Starting from this limited dataset, *firestar* has been used to refine predictions from SIAM, a homology based GO term predictor. This approach was tested in the context of the CAFA2 experiment.

5.1 Ligand annotation

We described the analysis and annotation of PDB ligands in section 4.1. There are many small molecules that co-crystallise with proteins but those that are biologically relevant are the most important for the purposes of FireDB. The definition of what is considered biologically important is quite straightforward (molecules that participate as cofactors, substrates or products in biological reactions). In FireDB these biologically relevant compounds are termed COGNATE. The fact that a ligand is COGNATE does not automatically mean that the site in which it binds is biologically relevant. Although these ligands are likely to bind biologically relevant binding sites, they can also bind outside of biological sites (metals in particular), and there are many examples of biologically relevant sites occupied by non-COGNATE compounds. In the original FireDB, ligand classification was realized indirectly, by assuming that ligands binding frequently to PDB relevant sites were probably COGNATE. However, there were various examples where this was not always true, as in the case of sucrose (figure 6, section 4.1.2). Manual curation of biological relevance improves the reliability of the database. Here the automatic mapping of all compounds to biology-centric databases will at least speed up the prioritization step, facilitating the curator work.

The distinction between metals and non-metals (section 4.1.3) is important because it opened the door to further refinements in FireDB. Ligands with specific binding site characteristics (in terms of architecture, preferential amino acid composition and conservation) could be grouped and merged independently. This process has two advantages. First of all binding sites with different characteristics are no longer collapsed together to generate MSS. This makes the MSS characteristics less noisy and improves predictions at the *firestar* level. Separation also makes it possible to create type-specific rules that can be used for biological relevance evaluation in FireDB and for filtering of truncated or degenerated predicated sites in *firestar*. Ideally, this work should be expanded to cover more COGNATE compound in the future, although the present PDB content does not have sufficient coverage for many ligands.

NON COGNATE compounds are in the majority in the PDB. The external cross-linking in FireDB follows the general tendency of diverse primary databases (such as the 4 data centers of wwPDB¹⁵¹, Uniprot¹⁸¹ and many more) to unify related data from different scientific repositories. At the same time, cross-linking highlights a common hurdle in scientific data: standardization. It is striking how difficult it was to retrieve perfect matches between PDB and PubChem, despite the existence of established formats like SMILES and InChi that should ensure easy computational compound matching between different sources (section 4.1.6). To overcome this problem, we decided to make use of pre-calculated cross-matching directly generated from the selected databases and from Unichem¹⁴⁹, a non-redundant database of pointers between chemical structure repositories. When discrepancies were found, we manually solved them to improve matching consistency (section 4.1.6).

Further uses for the external matches have not been explored in this work; a possible secondary refinement step for binding site annotation could be realized analyzing structural differences of antagonists and analogs and their correspondent COGNATE compounds. This would allow the identification of specific NON-COGNATE binding residues that could be eliminated or tagged.

In addition, annotation of biological activity of NON-COGNATE ligands could widen FireDB usage. The construction of the database itself allows rapid associations of COGNATE and pharmacological chemicals that bind at the same site since this is automatically done in the definition of MSS clusters (section 3.7). Going a step further, it would be possible to compare all MSS containing the pharmacological molecule of

interest in FireDB. Finally, information from homologous site comparisons could provide new interesting insights. Overlaps between COGNATE and NON-COGNATE ligands could be studied case by case to identify new candidates sites for drug repositioning studies. While the real capability for an existing drug to bind alternative sites has to be evaluated with specialised tools (chemoinformatics, *in-silico* simulations among others), the approaches to propose and discover new associations are varied, from tridimensional site arrangement comparison¹⁸², expression profiles analysis¹⁸³ to literature derived gene-drug associations¹⁸⁴. Sequence homology has been used before as a scoring parameter¹⁸⁵, but as far as we know conservation at functional residues level has yet to be explored.

5.2 Biologically relevant binding sites

The increasing deposition of protein structures brings a consequent increase in the availability of binding site information. Most of these binding sites will be artifacts of the crystallization process, so the continuing assessment of biological relevance is important, not least for the repercussions on function prediction algorithms such as *firestar*.

The first protocol for the biological assessment of binding sites in FireDB was designed to take into account the biased functional content of the PDB. The assessment counted on characteristics based on the frequency of appearance of each ligand and structural criteria, along with evolutive criteria, for this reason (section 4.2, table 6). We based our classification criteria principally on the observation that a binding site with no homologs has a higher probability of being a crystal artifact. Of course this may be not true for all sites, so the manual curation of biologically relevant compounds helps to recover NOVEL sites: we were able to rescue 4,190 sites with the tag "NOVEL" from the 51,618 binding sites with no homologs in FireDB, more than 8%.

Manual evaluation suggests that this protocol is reliable (section 4.2.1). Additional improvements in coverage and specificity can come from the completion of the manual curation of biologically relevant compounds and the growth of PDB. The study of complicated cases such as the heme binding site of leghaemoglobin (figure 12, section 4.2.1) can also be instructive. Although there is a clear conservation pattern, FireDB tags the site as NOT SUPPORTED, due to the cut-off for conserved residue coverage. Although we tried a range of combinations in order to overcome this problem (common to bulky ligands), we obtained more incorrectly classified sites as a side effect of loosening our thresholds (results not shown). Specific ligand rules or ligand group rules, similar to the ones developed for metallic ligands, could be implemented to deal with this type of case.

Other available repositories provide biological relevance assessment of PDB extracted binding sites^{94,95,186,187}. These resources share similar goals, but they all have peculiarities that make them unique and interesting resources. BioLip⁹⁵ is the most similar to FireDB; it was published in 2012 and proposed a semi-automatic assessment of sites. New PDB entries are evaluated weekly and an automatic protocol identifies candidate biological relevant sites that are reviewed by a human annotator. Despite this there are important differences in the definition of biologically relevant ligands (in BioLip all but crystallization mix components are biologically important), site definition (in FireDB sequences are clustered into MSSs while in BioLip every protein is an entity *per se*) and type (in FireDB protein-protein and DNA-protein interaction are beyond the remit of the database).

5.3 Availability and future database developments

For scientific annotation resources such as FireDB, one of the most important characteristics is the free and easy availability of the data for the community. This is important for the evolution of the field and also because external feedback helps to identify needs and to spot undetected problems. For this reason we worked to improve FireDB documentation and to make the entire resource available as a downloadable mysql database. Accessibility through REST services allows programmatic access to the data via scripts without creation of a local database. Furthermore this ensures access to the most recent data.

The main information sources for FireDB are the PDB and the CSA, but any source of experimentally validated functional residues could easily be integrated. For instance, annotations for post-translational modifications, such as phosphorylation¹⁸⁸, or functional binding sites in intrinsically disordered proteins¹⁸⁹, particularly challenging from the structural point of view. Whether these features could then be used for *firestar* predictions would depend on their evolutive pattern, since SQUARE detects conservation hotspots.

5.4 Functional residues prediction

firestar is a template-based method for the prediction of small ligands binding residues. But these are not the only residues involved in protein function determination: protein-protein interface binding residues are other important categories. Due to the characteristics of these sites, *firestar* is actually not able to predict these residues: protein-protein binding sites can be very large and variable in their composition¹⁹⁰, and SQUARE analysis at single residue level may not be very informative. A number of different approaches, based on functional determinants^{113,191} or more recently on coevolution¹⁹² have demonstrated to be better suited for this kind of predictions.

Apart from *firestar* and structural methods (we mentioned some of them in CASP10 results analysis), there are other functional residue methods that use only sequence information^{111,191,193} and so generate *de novo* predictions. Based on the information generated from MSAs, they can classify residues as functionally important or not, and often generate a reliability score for their prediction. The most important advantage of these approaches is that in principle they are designed to obtain predictions for any structure or sequence. Template-based method are limited by the available information, but if a homolog is detected, they can provide further resolution of the binding site, such as the ligand involved and whether a residue is catalytic and not just conserved.

5.5 *firestar* performance analysis

firestar's ability to predict ligand binding sites relies on 3 aspects: the source information, the homologous search method and the evaluation of the conservation of individual positions.

5.5.1 Source information

FireDB content directly affects the sensitivity of *firestar*, meaning that if a homologue binding site for a certain protein is not present, *firestar* cannot make a prediction: one good example is the metal binding site of T0478 in CASP8 experiment. As result of the work of the crystallographic groups and of the structural genomics

consortia¹⁹⁴, PDB content is not only increasing, but also structure resolution has been prioritized with the aim of expanding coverage of structural protein space. Many of these proteins have been crystallised with one or more ligand, and so the general growth has come with an increase in FireDB binding site information, as shown in figure 8, section 4.2., and this in turn allows *firestar* to make more predictions.

Furthermore the structural spaces and the functional sites are not perfectly coupled since certain functional sites may be present in different foldings by evolutionary convergence, and this phenomenon has been already studied⁸¹. This should reduce to some extent the limitation that the presence of functional sites in PDB structures is subject to the same bottlenecks that affect the resolution of protein structures. Therefore, if this trend continues, we should be able to find in the FireDB database all different protein-ligand complexes, or at least one close homolog, existing in nature but the reality is that there are some limitations.

Even though the expansion of the PDB may provide representative templates for many families of biological sites, there is evidence to suggest that not all families will be equally represented. Furthermore many ligands may not be crystallised because their interactions are less stable, or more transient. For example, of the 51 compounds crystallised in the 22 CASP8 targets and 20 CASP10 targets, 20 were cofactors or nucleotides and 29 were metal ions with structural function or involved in catalysis. Only in two cases did we find differences: for target T0694 (a chitinase) histidine seems to be a product, while for target T0661 (a lipid transporter) Palmitoyl-phosphatidylethanolamine (PEF) is the substrate of its action. This suggests that substrate and products are under-represented in the PDB and this will limit the ability to predict substrates and products with *firestar*. Information on these problematic functional sites may have to be collected from experiments of biochemical characterization of proteins whose structure is not necessarily known.

Another aspect to take into account is the effect of the expansion of FireDB on the false positive detection rate of *firestar*. As we shown in figure 19 in section 4.5.1, *firestar* predictions were affected in terms of false positive detection by the growth of the database (data for PSI-BLAST only analysis) for bulkier ligands such as ADP. We corrected this by improving the assessment of biological relevance, and so avoiding the introduction of noise from non-specific sites. Furthermore we set up conservation and coverage filters (sections 4.5.2, 4.5.3 and 4.5.4) to discard information from degenerated or truncated binding sites.

5.5.2 Homologous search methods and alignment quality

The ability of *firestar* to obtain predictions depends mainly on its sensitivity to homologous templates in the FireDB database. The first search algorithm implemented, PSI-BLAST, is a widely used method to detect remote homologs. In order to widen the candidate search strategy and to improve *firestar* sensitivity, we decided to integrate a more recent algorithm, HHsearch. It has two main differences from PSI-BLAST:

1. It is a profile-profile searching method, so in principle it should be more sensitive and it should generate better alignments in comparison with PSI-BLAST, which is just a profile-sequence searching method⁶⁷;
2. Secondary structure information can be included in the search parameters;

We observed several targets where HHsearch was able to detect many more templates than PSI-BLAST (for example the case of zinc binding site of CASP8 target T0480, section 4.5.5). However, results of the analysis (section 4.5.1) suggest that PSI-BLAST alignments provide different information, and for this reason we decided to maintain both search engines. Variability in alignments is useful, especially for challenging targets, since the predictions depend on the SQUARE evaluations.

The failed prediction for T0720 showed the limits of the search methods. *firestar* cannot generate an automatic prediction because the two parts of the candidate site are split and remote in sequence, and PSI-BLAST and HHsearch are not able to generate an alignment that spans the two parts. Both sites are filtered out because of insufficient coverage. If the separate parts had satisfied filtering criteria, *firestar* could have predicted 2 different sites. Here the detailed results section was still informative, and actually our human group in CASP10 was able to submit the correct prediction based on this (best MCC for the target, see appendix table 4). In general sites made up of residues distant in sequence can be problematic if detected in remote homologs.

5.5.3 Alignments quality and position conservation

The SQUARE estimation of local alignment reliability is a fundamental step in the *firestar* pipeline; the first acceptance filter is based on the score assigned to every single aligned residue (section 4.4.1). The importance of this evaluation is even greater for remote homologs where it is not possible to distinguish between true remote homologs and random hits solely based on the algorithm scores. Here SQUARE is still able to detect conservation and to extract useful binding information. For example in figures 28 and 29 of section 4.5.6.2 we show how SQUARE is able to spot confirmed conserved residues even from templates with an e-value over 30. However, if the alignment quality is poor, SQUARE cannot correctly evaluate positions.

Improvement in sequence alignment quality could be brought by the use structural alignments^{195,196}, but in most cases we would have to align a model structure of the target protein. This option has some drawbacks, since we cannot be sure of the 3D arrangement of the binding site in the model. Furthermore when we are superimposing two rigid bodies the more flexible regions may be displaced, and this could be very relevant since often binding sites include flexible or disordered regions. In addition, sequence alignments derived from the overlapping structures do not take into account evolutionary information.

In CASP10 experiment *firestar* was one of just two pure sequence-based methods, with HHpred (FN430). The other automatic methods^{119,197,198} used a mixed strategy, where sequence analysis was combined with structural information. These methods generated three-dimensional models for target proteins and searched libraries to find structurally homologous binding sites. Sequence information was used to refine the prediction. This approach can improve some predictions, but present limitations in others. The templates used for structure prediction may not be ideal for transferring information from functional sites (for example sites may undergo structural changes on binding). Erroneous side chain positioning is common in models and this is another limitation.

Results suggest that a structural approach can be useful in some cases, such as target T0682. Here we detected all binding residues for Retinal (PDB ligand id: RET), but we also added 7 false positive residues (table 10, section 4.5.6.1). We refined the *firestar* prediction using structural information for our CNIO human prediction. But in other cases, such as the reduced flavin mononucleotide (PDB ligand id: FNR) binding site for target T0744 *firestar* had the best MCC (0.803, appendix table 4) and the methods based on structure had worse predictions. CASP10 results suggest that while structural information can be useful for some targets, the overall improvement is minimal.

5.6 The effect of filters on *firestar* predictions

Much of the improvements in *firestar* accuracy came from reducing false positive predictions. In CASP8 experiment we detected a tendency to overpredict in larger binding sites such as target T0431 (bound to HEM), T0450 (bound to FAD), and T0490 (bound to FAD). But we also found false positives among metal binding sites, as in the case of

T0406 (bound to NI), T0410 (bound to FE) or T0444 (bound to FE). This *firestar* behavior was corrected introducing filters based on the site type (metal or non metal) or focused on the selection of the most relevant information coming from FireDB (section 4.5.4). As a result of this, we reduced the number of false positive by 53% (table 7, section 4.5.5).

CASP10 was a good testing ground to evaluate these filters against a completely new dataset, and to understand if the method was too conservative in comparison with the others. As shown in table 12 (section 4.5.6.3), *firestar* is the best automatic server in terms of false positive rate and the third best automatic server in terms of true positive detected (187, against 206 of SP-ALIGN and 188 of I-TASSER function).

However, target T0715 is interesting. *firestar* did not find a close homolog even though there were many NAD templates and the occurrence and conservation filters discarded all but the core binding residues, with the result that *firestar* predicted less than 50% of the site. Other methods performed significantly better for this target. We would need to evaluate further cases to confirm whether this was an isolated case or a general trend for bulky ligands and distant templates.

5.7 Applications of FireDB and *firestar* in large-scale projects

In this work we explored the use of *firestar* and FireDB in large scale functional annotation applications. The results confirm the importance of using functional information along with sequence and structural information to disentangle cellular and evolutive relationships.

firestar and FireDB were included as functional annotation tools in the APPRIS database, part of the GENCODE consortium. *firestar* was included in APPRIS because it was shown to be a reliable prediction method. *firestar* has the lowest coverage of the human genome among the 4 core programs in APPRIS, but this is to be expected, since not all the proteins contain ligand binding residues. At the same time, *firestar* is the method that has the best agreement with the gold standard for principal isoforms among the four core methods (section 4.6.1.1), supporting the use of functional information for this purpose. Indeed for cases such as the read-through transcripts or the incorrectly tagged nonsense_mediated_decay transcript of RP4-583P15.15, *firestar* is helping to refine the human genome annotation

The second project was centered on the annotation of Pfam functional protein domains. These domains, or families, are defined regions related by homology. For this reason conservation of functionally important residues within the family, and within higher-level groupings of families, or clans, is to be expected. Results from our preliminary analysis suggested that the *firestar* and FireDB annotations are consistent for most clans, supporting the aggregation of related families. The case of the P-loop Ntpase (section 4.6.2) demonstrates how functional residues can be used as additional criteria in clan generation, and how this kind of annotation could be included in the Pfam database.

In the third project we integrated *firestar* with SIAM, a functional annotation algorithm based on homology. The two methods in principle are not overlapping, since SIAM makes global functional predictions while *firestar* identifies local functional characteristics. The strategy used to map GO terms on FireDB compounds allowed us to cover a small fraction of the sites contained in the database, since only 112 compounds were annotated. We also used the catalytic site residues to annotate GO terms that are directly related to catalytic activity since catalytic sites are associated with enzymatic numbers. The GOA database¹⁹⁹ provide a list of correspondences between EC numbers and GO terms; in this way 99% of Master Sequence catalytic sites are associated with at least one GO term.

| DISCUSSION

Over the entire CAFA2 experiment, *firestar* coverage was similar to SIAM in terms of number of sequences annotated, and the limited overlap of terms supported the combined use of the two algorithms to cover a wider range of proteins. However, the results from the enzyme (EFI) dataset, showed that *firestar* obtained almost a 96% coverage and 17% more annotations than SIAM. A deep analysis of the results would shed light on the performance of the combined strategy.

6 CONCLUSIONS

1. We developed an automatic pipeline to map all the compounds contained in FireDB to 8 external specialized repositories, obtaining a final coverage of almost 94%. Additionally we retrieved, when available, associated bioactivity information and we further expanded the annotation dataset with manual literature mining for 326 database entries.
2. We manually assessed biological relevance of 664 ligand compounds and we tagged as ambiguous 56 compounds. Furthermore we classified a group of 31 molecules as metals, and we studied their characteristics at binding site level. The use of this information in the construction of the database supposed an important improvement in information quality.
3. Biological relevance for annotated binding sites has been automatically assessed using residue-level evolutionary conservation and manual ligand annotations. This protocol has been integrated in database construction pipeline.
4. *firestar* usability has been improved with the introduction of a fully automatic protocol able to evaluate and merge results from many templates into a an easy-to-read ranked list of consensus predictions.
5. The specificity and sensitivity of *firestar* have been improved with the introduction of a new sequence search method and individual filters for different candidate sites. Improvements have been tested in the context of an independent blind experiment, where *firestar* was demonstrated to be a state-of-the-art method for the prediction of functionally important residues and bound ligands.
6. We have developed a stand-alone version that allows *firestar* to be incorporated into large-scale pipelines. Reliable protein functional residues and ligand prediction has been used in three different annotation projects; the results presented here show that FireDB and *firestar* provide important information for proteome-wide functional and biomedical projects.

1. En este trabajo se ha desarrollado un protocolo automático capaz de vincular cerca del 94% de las moléculas no proteicas contenidas en FireDB con sus anotaciones en 8 repositorios especializado externos. Adicionalmente se ha recuperado la información disponible para estos compuestos relativa a su bioactividad, recurriendo en 326 compuestos a la anotación manual a partir de la literatura.
2. Hemos determinado manualmente la relevancia biológica de 664 compuestos y anotado otros 56 como ambiguos. Así mismo, hemos clasificado 31 ligandos como metales, para los que hemos estudiado las características de sus sitios de unión. La incorporación de esta información en el proceso de generación de FireDB, ha mejorado sustancialmente la calidad de la información recuperada.
3. Hemos establecido un protocolo de evaluación automática de la relevancia biológica de los sitios de unión anotados, basado en la conservación evolutiva de los residuos en sitios de unión homólogos y en la propia relevancia biológica del correspondiente ligando. Esta metodología también ha sido integrada en el proceso de generación de FireDB.
4. Se ha mejorado la funcionalidad de *firestar*, con la introducción de un protocolo automático capaz de evaluar e integrar los resultados de todos las estructuras homólogas encontradas en un listado priorizado de predicciones consenso de fácil interpretación.
5. Se han mejorado la especificidad y la sensibilidad de *firestar*, con la integración de un nuevo método de búsqueda de homólogos y la introducción de filtros específicos. Estas mejoras se han evaluado en el contexto de un experimento ciego internacional, cuyos resultados sitúan a *firestar* entre los mejores métodos de predicción de residuos funcionalmente importantes y de ligandos unidos.
6. Hemos desarrollado una versión autónoma de *firestar* que facilita su integración en protocolos de análisis a gran escala. De hecho, FireDB y *firestar* han aportado predicciones fiables de residuos y de ligandos unidos a tres grandes proyectos diferentes. Los resultados presentados avalan la capacidad de ambas herramientas para proporcionar información esencial para proyectos proteómicos a gran escala y de interés biomédico.

7 BIBLIOGRAPHY

1. Tiessen A, Pérez-Rodríguez P, Delaye-Arredondo LJ. Mathematical modeling and comparison of protein size distribution in different plant, animal, fungal and microbial species reveals a negative correlation between protein size and protein number, thus providing insight into the evolution of proteomes. *BMC Res Notes*. 2012;5(1):85. doi:10.1186/1756-0500-5-85.
2. Meyer LC, Wright NT. Structure of giant muscle proteins. *Front Physiol*. 2013;4(December):368. doi:10.3389/fphys.2013.00368.
3. Lagerkvist U. "Two out of three": an alternative method for codon reading. *Proc Natl Acad Sci U S A*. 1978;75(4):1759-1762. doi:10.1073/pnas.75.4.1759.
4. Xu X-M, Carlson B a, Irons R, *et al*. Selenophosphate synthetase 2 is essential for selenoprotein biosynthesis. *Biochem J*. 2007;404(1):115-120. doi:10.1042/BJ20070165.
5. Gaston M a, Zhang L, Green-Church KB, Krzycki J a. The complete biosynthesis of the genetically encoded amino acid pyrrolysine from lysine. *Nature*. 2011;471(7340):647-650. doi:10.1038/nature09918.
6. Catherman AD, Skinner OS, Kelleher NL. Top Down proteomics: Facts and perspectives. *Biochem Biophys Res Commun*. 2014;445(4):683-693. doi:10.1016/j.bbrc.2014.02.041.
7. Pagani I, Liolios K, Jansson J, *et al*. The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res*. 2012;40(Database issue):D571-9. doi:10.1093/nar/gkr1100.
8. The Uniprot Consortium. Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res*. 2014;42:D191-8. doi:10.1093/nar/gkt1140.
9. Kosuge T, Mashima J, Kodama Y, *et al*. DDBJ progress report: a new submission system for leading to a correct annotation. *Nucleic Acids Res*. 2014;42(Database issue):D44-9. doi:10.1093/nar/gkt1066.
10. Benson D a, Cavanaugh M, Clark K, *et al*. GenBank. *Nucleic Acids Res*. 2013;41(Database issue):D36-42. doi:10.1093/nar/gks1195.
11. Levinthal C. Are there pathways for protein folding. *J Chim phys*. 1968.
12. Dill K a, Ozkan SB, Shell MS, Weikl TR. The protein folding problem. *Annu Rev Biophys*. 2008;37:289-316. doi:10.1146/annurev.biophys.37.092707.153558.
13. Feng Y, De Franceschi G, Kahraman A, *et al*. Global analysis of protein structural changes in complex proteomes. *Nat Biotechnol*. 2014;32(10). doi:10.1038/nbt.2999.
14. Groban ES, Narayanan A, Jacobson MP. Conformational changes in protein loops and helices induced by post-translational phosphorylation. *PLoS Comput Biol*. 2006;2(4):e32. doi:10.1371/journal.pcbi.0020032.
15. Cox S, Radzio-Andzelm E, Taylor SS. Domain movements in protein kinases. *Curr Opin Struct Biol*. 1994;4:893-901. doi:10.1016/0959-440X(94)90272-0.
16. Gohara DW, Di Cera E. Allostery in trypsin-like proteases suggests new therapeutic strategies. *Trends Biotechnol*. 2011;29(11):577-585. doi:10.1016/j.tibtech.2011.06.001.Allostery.
17. Wrabl JO, Gu J, Liu T, Schrank TP, Whitten ST, Hilser VJ. The role of protein conformational fluctuations in allostery, function, and evolution. *Biophys Chem*. 2011;159:129-141. doi:10.1016/j.bpc.2011.05.020.
18. Dunker a K, Silman I, Uversky VN, Sussman JL. Function and structure of inherently disordered proteins. *Curr Opin Struct Biol*. 2008;18(6):756-764. doi:10.1016/j.sbi.2008.10.002.
19. Yu H. Extending the size limit of protein nuclear magnetic resonance. *Proc Natl Acad Sci*. 1999;96(January):332-334.
20. Boekema EJ, Folea M, Kouřil R. Single particle electron microscopy. *Photosynth Res*. 2009;102(2-3):189-196. doi:10.1007/s11120-009-9443-1.
21. Berman HM, Westbrook J, Feng Z, *et al*. The Protein Data Bank. *Nucleic Acids Res*. 2000;28:235-242. doi:10.1093/nar/28.1.235.

22. Wang Z. How Many Fold Types of Protein Are There in Nature? *Proteins Struct Funct Bioinforma*. 1996;191:186-191. doi:10.1002/(SICI)1097-0134(199610)26:2<186::AID-PROT8>3.0.CO;2-E.
23. Schaeffer RD, Daggett V. Protein folds and protein folding. *Protein Eng Des Sel*. 2011;24(1-2):11-19. doi:10.1093/protein/gzq096.
24. Murzin a G, 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(4):536-540. doi:10.1006/jmbi.1995.0159.
25. Andreeva A, Howorth D, Chothia C, Kulesha E, Murzin AG. SCOP2 prototype: A new approach to protein structure mining. *Nucleic Acids Res*. 2014;42(D1):310-314. doi:10.1093/nar/gkt1242.
26. Sillitoe I, Lewis TE, Cuff A, *et al*. CATH: Comprehensive structural and functional annotations for genome sequences. *Nucleic Acids Res*. 2015;43(D1):D376-D381. doi:10.1093/nar/gku947.
27. Burgess a, Labbé J-C, Vigneron S, *et al*. Chfr interacts and colocalizes with TCTP to the mitotic spindle. *Oncogene*. 2008;27(42):5554-5566. doi:10.1038/onc.2008.167.
28. Koziol MJ, Garrett N, Gurdon JB. Tpt1 activates transcription of oct4 and nanog in transplanted somatic nuclei. *Curr Biol*. 2007;17(9):801-807. doi:10.1016/j.cub.2007.03.062.
29. Bheekha-Escura R, MacGlashan DW, Langdon JM, MacDonald SM. Human recombinant histamine-releasing factor activates human eosinophils and the eosinophilic cell line, AML14-3D10. *Blood*. 2000;96(6):2191-2198.
30. Haga A, Niinaka Y, Raz A. Phosphohexose isomerase/autocrine motility factor/neuroleukin/maturation factor is a multifunctional phosphoprotein. *Biochim Biophys Acta - Protein Struct Mol Enzymol*. 2000;1480:235-244. doi:10.1016/S0167-4838(00)00075-3.
31. Gurney ME, Apatoff BR, Spear GT, *et al*. Neuroleukin: a lymphokine product of lectin-stimulated T cells. *Science (80-)*. 1986;234(4776):574-81. doi:10.1126/science.3020690.
32. Fink AL. Natively unfolded proteins. *Curr Opin Struct Biol*. 2005;15(1):35-41. doi:10.1016/j.sbi.2005.01.002.
33. Bairoch A. The ENZYME database in 2000. *Nucleic Acids Res*. 2000;28:304-305. doi:10.1093/nar/28.1.304.
34. Ashburner M, Ball C, Blake J, Botstein D. Gene Ontology: tool for the unification of biology. *Nature*. 2000;25(may):25-29.
35. Heard E, Martienssen RA. Transgenerational Epigenetic Inheritance: Myths and Mechanisms. *Cell*. 2014;157(1):95-109. doi:10.1016/j.cell.2014.02.045.
36. Smith TF, Waterman MS. Identification of common molecular subsequences. *J Mol Biol*. 1981;147:195-197. doi:10.1016/0022-2836(81)90087-5.
37. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990;215(3):403-410. doi:10.1016/S0022-2836(05)80360-2.
38. Zemla a. LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Res*. 2003;31(13):3370-3374. doi:10.1093/nar/gkg571.
39. Holm L, Rosenström P. Dali server: conservation mapping in 3D. *Nucleic Acids Res*. 2010;38(Web Server issue):W545-9. doi:10.1093/nar/gkq366.
40. del Pozo A, Pazos F, Valencia A. Defining functional distances over gene ontology. *BMC Bioinformatics*. 2008;9:50. doi:10.1186/1471-2105-9-50.
41. Benabderrahmane S. IntelliGO: a new vector-based semantic similarity measure including annotation origin. *BMC Bioinformatics*. 2010;11(1):588. doi:10.1186/1471-2105-11-588.
42. Ramírez F, Lawyer G, Albrecht M. Novel search method for the discovery of functional relationships. *Bioinformatics*. 2012;28(2):269-276. doi:10.1093/bioinformatics/btr631.

43. Chothia C, Lesk A. The relation between the divergence of sequence and structure in proteins. *EMBO J*. 1986;5(4):823-826.
44. Rost B. Protein structures sustain evolutionary drift. *Fold Des*. 1997:19-24.
45. Rost B. Twilight zone of protein sequence alignments. *Protein Eng*. 1999;12(2):85-94.
46. Chung SY, Subbiah S. A structural explanation for the twilight zone of protein sequence homology. *Structure*. 1996;4(10):1123-1127.
47. Valencia A, Kjeldgaard M, Pai EF, Sander C. GTPase domains of ras p21 oncogene protein and elongation factor Tu: analysis of three-dimensional structures, sequence families, and functional sites. *Proc Natl Acad Sci U S A*. 1991;88(12):5443-5447.
48. Wood TC, Pearson WR. Evolution of protein sequences and structures. *J Mol Biol*. 1999;291(4):977-995. doi:10.1006/jmbi.1999.2972.
49. Koehl P, Levitt M. Sequence Variations within Protein Families are Linearly Related to Structural Variations. *J Mol Biol*. 2002;323(3):551-562. doi:10.1016/S0022-2836(02)00971-3.
50. Panchenko AR, Wolf YI, Panchenko L a, Madej T. Evolutionary plasticity of protein families: coupling between sequence and structure variation. *Proteins*. 2005;61(3):535-544. doi:10.1002/prot.20644.
51. Zhang Z, Wang Y, Wang L, Gao P. The combined effects of amino acid substitutions and indels on the evolution of structure within protein families. *PLoS One*. 2010;5(12):e14316. doi:10.1371/journal.pone.0014316.
52. Kabsch W, Holmes K. The actin fold. *FASEB J*. 1995:167-174.
53. Lattman E, Rose G. Protein folding--what's the question? *Proc Natl Acad Sci U S A*. 1993;90(January):439-441.
54. Sutto L, Marsili S, Valencia A, Gervasio FL. From residue coevolution to protein conformational ensembles and functional dynamics. *Proc Natl Acad Sci*. 2015;112(44):13567-13572. doi:10.1073/pnas.1508584112.
55. Boeckmann B, Blatter MC, Famiglietti L, et al. Protein variety and functional diversity: Swiss-Prot annotation in its biological context. *Comptes Rendus - Biol*. 2005;328(10-11):882-899. doi:10.1016/j.crv.2005.06.001.
56. Schnoes AM, Ream DC, Thorman AW, Babbitt PC, Friedberg I. Biases in the experimental annotations of protein function and their effect on our understanding of protein function space. *PLoS Comput Biol*. 2013;9(5):e1003063. doi:10.1371/journal.pcbi.1003063.
57. Todd a E, Orengo C a, Thornton JM. Evolution of function in protein superfamilies, from a structural perspective. *J Mol Biol*. 2001;307(4):1113-1143. doi:10.1006/jmbi.2001.4513.
58. Devos D, Valencia A. Practical limits of function prediction. *Proteins Struct Funct Bioinforma*. 2000;41(1):98-107.
59. Quackenbush J, Cho J, Lee D, et al. The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Res*. 2001;29(1):159-164.
60. Holm L, Sander C. The FSSP database: fold classification based on structure-structure alignment of proteins. *Nucleic Acids Res*. 1996;24(1):206-209.
61. Sangar V, Blankenberg DJ, Altman N, Lesk AM. Quantitative sequence-function relationships in proteins based on gene ontology. *BMC Bioinformatics*. 2007;8(1):294. doi:10.1186/1471-2105-8-294.
62. Rost B. Enzyme function less conserved than anticipated. *J Mol Biol*. 2002;318(2):595-608. doi:10.1016/S0022-2836(02)00016-5.
63. Holm L, Sander C. Mapping the protein universe. *Science*. 1996;273(5275):595-603.
64. Dayhoff M, Schwartz R. A Model of Evolutionary Change in Proteins. *Atlas protein Seq Struct*. 1978:345-352.

65. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*. 1992;89(22):10915-10919.
66. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res*. 2011;39(Web Server issue):W29-37. doi:10.1093/nar/gkr367.
67. Söding J. Protein homology detection by HMM-HMM comparison. *Bioinformatics*. 2005;21(7):951-960. doi:10.1093/bioinformatics/bti125.
68. Mariani V, Kiefer F, Schmidt T, Haas J, Schwede T. Assessment of template based protein structure predictions in CASP9. *Proteins*. 2011;79 Suppl 1:37-58. doi:10.1002/prot.23177.
69. De Filippo C, Ramazzotti M, Fontana P, Cavalieri D. Bioinformatic approaches for functional annotation and pathway inference in metagenomics data. *Brief Bioinform*. 2012;13(6):696-710. doi:10.1093/bib/bbs070.
70. Schnoes AM, Brown SD, Dodevski I, Babbitt PC. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput Biol*. 2009;5(12):e1000605. doi:10.1371/journal.pcbi.1000605.
71. Bork P, Koonin E. Predicting functions from protein sequences—where are the bottlenecks? *Nat Genet*. 1998:313-318.
72. Worth CL, Gong S, Blundell TL. Structural and functional constraints in the evolution of protein families. *Nat Rev Mol Cell Biol*. 2009;37(4):727-733. doi:10.1038/nrm2762.
73. Guharoy M, Chakrabarti P. Conserved residue clusters at protein-protein interfaces and their use in binding site identification. *BMC Bioinformatics*. 2010;11:286. doi:10.1186/1471-2105-11-286.
74. Ouzounis C, Pérez-Irratxeta C. Are binding residues conserved? *Pacific Symp Biocomput*. 1997:401-412.
75. Nelson DL, Cox MM. Lehninger Principles of Biochemistry 5th ed. *Book*. 2008:1-1294. doi:10.2307/1309148.
76. Holliday GL, Andreini C, Fischer JD, *et al*. MACiE: exploring the diversity of biochemical reactions. *Nucleic Acids Res*. 2012;40(Database issue):D783-9. doi:10.1093/nar/gkr799.
77. Holliday GL, Almonacid DE, Mitchell JBO, Thornton JM. The chemistry of protein catalysis. *J Mol Biol*. 2007;372(5):1261-1277. doi:10.1016/j.jmb.2007.07.034.
78. Holliday GL, Mitchell JBO, Thornton JM. Understanding the functional roles of amino acid residues in enzyme catalysis. *J Mol Biol*. 2009;390(3):560-577. doi:10.1016/j.jmb.2009.05.015.
79. Porter CT, Bartlett GJ, Thornton JM. The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res*. 2004;32(Database issue):D129-33. doi:10.1093/nar/gkh028.
80. Furnham N, Holliday GL, de Beer T a P, Jacobsen JOB, Pearson WR, Thornton JM. The Catalytic Site Atlas 2.0: cataloging catalytic sites and residues identified in enzymes. *Nucleic Acids Res*. 2014;42(Database issue):D485-9. doi:10.1093/nar/gkt1243.
81. Gherardini PF, Wass MN, Helmer-Citterich M, Sternberg MJE. Convergent evolution of enzyme active sites is not a rare phenomenon. *J Mol Biol*. 2007;372(3):817-845. doi:10.1016/j.jmb.2007.06.017.
82. Hubbard TJ, Ailey B, Brenner SE, Murzin AG, Chothia C. SCOP: a Structural Classification of Proteins database. *Nucleic Acids Res*. 1999;27(1):254-256.
83. Smirnova I, Chang H-Y, von Balmoos C, Ädelroth P, Gennis RB, Brzezinski P. Single mutations that redirect internal proton transfer in the ba3 oxidase from *Thermus thermophilus*. *Biochemistry*. 2013;52(40):7022-7030.

84. Boucher CA, Cammack N, Schipper P, *et al.* High-level resistance to (-) enantiomeric 2'-deoxy-3'-thiacytidine in vitro is due to one amino acid substitution in the catalytic site of human immunodeficiency virus type 1 reverse transcriptase. *Antimicrob Agents Chemother.* 1993;37(10):2231-2234. doi:10.1128/AAC.37.10.2231.
85. Andreeva NS, Rumsh LD. Analysis of crystal structures of aspartic proteinases: On the role of amino acid residues adjacent to the catalytic site of pepsin-like enzymes. *Protein Sci.* 2001:2439-2450.
86. Fujimori A, Harker WG, Kohlhagen G, Hoki Y, Pommier Y. Mutation at the Catalytic Site of Topoisomerase I in CEM / C2 , a Human Leukemia Cell Line Resistant to Camptothecin. *Cancer Res.* 1995:1339-1346.
87. Ivanisenko V a, Pintus SS, Grigorovich D a, Kolchanov N a. PDBSite: a database of the 3D structure of protein functional sites. *Nucleic Acids Res.* 2005;33(Database issue):D183-7. doi:10.1093/nar/gki105.
88. Kleywegt GJ, Jones T a. Databases in protein crystallography. *Acta Crystallogr D Biol Crystallogr.* 1998;54(Pt 6 Pt 1):1119-1131.
89. Laskowski R a. PDBsum: summaries and analyses of PDB structures. *Nucleic Acids Res.* 2001;29(1):221-222.
90. de Beer T a P, Berka K, Thornton JM, Laskowski R a. PDBsum additions. *Nucleic Acids Res.* 2014;42(Database issue):D292-6. doi:10.1093/nar/gkt940.
91. Stuart A, Ilyin V, Sali A. LigBase: a database of families of aligned ligand binding sites in known protein sequences and structures. *Bioinformatics.* 2002;18(1):200-201.
92. Hendlich M, Bergner A, Günther J, Klebe G. Relibase: Design and Development of a Database for Comprehensive Analysis of Protein–Ligand Interactions. *J Mol Biol.* 2003;326(2):607-620. doi:10.1016/S0022-2836(02)01408-0.
93. Benson ML, Smith RD, Khazanov N a, *et al.* Binding MOAD, a high-quality protein-ligand database. *Nucleic Acids Res.* 2008;36(Database issue):D674-8. doi:10.1093/nar/gkm911.
94. Dessailly BH, Lensink MF, Orengo C a, Wodak SJ. LigASite--a database of biologically relevant binding sites in proteins with known apo-structures. *Nucleic Acids Res.* 2008;36(Database issue):D667-73. doi:10.1093/nar/gkm839.
95. Yang J, Roy A, Zhang Y. BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic Acids Res.* 2013;41(Database issue):D1096-103. doi:10.1093/nar/gks966.
96. Bairoch A, Boeckmann B, Ferro S, Gasteiger E. Swiss-Prot: juggling between evolution and stability. *Brief Bioinform.* 2004;5(1):39-55.
97. Schomburg I, Chang A, Placzek S, *et al.* BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA. *Nucleic Acids Res.* 2013;41(Database issue):D764-72. doi:10.1093/nar/gks1049.
98. Bartlett GJ, Porter CT, Borkakoti N, Thornton JM. Analysis of Catalytic Residues in Enzyme Active Sites. *J Mol Biol.* 2002;324(1):105-121. doi:10.1016/S0022-2836(02)01036-7.
99. Nagano N. EzCatDB: the Enzyme Catalytic-mechanism Database. *Nucleic Acids Res.* 2005;33(Database issue):D407-12. doi:10.1093/nar/gki080.
100. Lopez G, Valencia A, Tress ML. FireDB--a database of functionally important residues from proteins of known structure. *Nucleic Acids Res.* 2007;35(Database):D219--D223. doi:10.1093/nar/gkm297.
101. Puvanendrapillai D, Mitchell JBO. Protein Ligand Database (PLD): additional understanding of the nature and specificity of protein-ligand complexes. *Bioinformatics.* 2003;19(14):1856-1857. doi:10.1093/bioinformatics/btg243.
102. Feng Z, Chen L, Maddula H, *et al.* Ligand Depot: a data warehouse for ligands bound to macromolecules. *Bioinformatics.* 2004;20(13):2153-2155. doi:10.1093/bioinformatics/bth214.

103. Shin J-M, Cho D-H. PDB-Ligand: a ligand database based on PDB for the automated and customized classification of ligand-binding structures. *Nucleic Acids Res.* 2005;33(Database issue):D238-41. doi:10.1093/nar/gki059.
104. Lipinski C a, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev.* 2001;46(1-3):3-26.
105. He Y, Yi W, Suino-Powell K, *et al.* Structures and mechanism for the design of highly potent glucocorticoids. *Cell Res.* 2014;24(6):713-726. doi:10.1038/cr.2014.52.
106. Günther S, Senger C, Michalsky E, Goede A, Preissner R. Representation of target-bound drugs by computed conformers: implications for conformational libraries. *BMC Bioinformatics.* 2006;7:293. doi:10.1186/1471-2105-7-293.
107. Bourne PE, Berman HM, McMahon B, Watenpaugh KD, Westbrook J, Fitzgerald PMD. The Macromolecular Crystallographic Information File (mmCIF). *Methods Enzymol.* 1997;277:571-590.
108. McNaught AD. The IUPAC International Chemical Identifier (InChI). *Chem Int.* 2007;2007.
109. Brown AS, Patel CJ. A review of validation strategies for computational drug repositioning. *Brief Bioinform.* 2016;(November):1-4. doi:10.1093/bib/bbw110.
110. Casari G, Sander C, Valencia A. A method to predict functional residues in proteins. *Nat Struct Biol.* 1995:171-178.
111. Lichtarge O, Bourne HR, Cohen FE. An evolutionary trace method defines binding surfaces common to protein families. *J Mol Biol.* 1996;257(2):342-358. doi:10.1006/jmbi.1996.0167.
112. Pazos F, Bang J-W. Computational Prediction of Functionally Important Regions in Proteins. *Curr Bioinform.* 2006;1(1):15-23. doi:10.2174/157489306775330633.
113. Rausell A, Juan D, Pazos F, Valencia A. Protein interactions and ligand binding: from protein subfamilies to functional specificity. *Proc Natl Acad Sci U S A.* 2010;107(5):1995-2000. doi:10.1073/pnas.0908044107.
114. de Juan D, Pazos F, Valencia A. Emerging methods in protein co-evolution. *Nat Rev Genet.* 2013;(March). doi:10.1038/nrg3414.
115. Binkowski TA, Joachimiak A. Protein functional surfaces: global shape matching and local spatial alignments of ligand binding sites. *BMC Struct Biol.* 2008;8:45. doi:10.1186/1472-6807-8-45.
116. Ghersi D, Sanchez R. EasyMIFs and SiteHound: A toolkit for the identification of ligand-binding sites in protein structures. *Bioinformatics.* 2009;25(23):3185-3186. doi:10.1093/bioinformatics/btp562.
117. Glaser F, Morris RJ, Najmanovich RJ, Laskowski RA, Thornton JM. A method for localizing ligand binding pockets in protein structures. *Proteins Struct Funct Genet.* 2006;62(2):479-488. doi:10.1002/prot.20769.
118. Wallace A, Borkakoti N, Thornton J. TESS: A geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. *Protein Sci.* 1997:2308-2323.
119. Wass MN, Kelley LA, Sternberg MJE. 3DLigandSite: Predicting ligand-binding sites using similar structures. *Nucleic Acids Res.* 2010;38(SUPPL. 2):469-473. doi:10.1093/nar/gkq406.
120. Oh M, Joo K, Lee J. Protein-binding site prediction based on three-dimensional protein modeling. *Proteins Struct Funct Bioinforma.* 2009;77(S9):152-156. doi:10.1002/prot.22572.
121. Roche DB, Tetchner SJ, McGuffin LJ. FunFOLD: an improved automated method for the prediction of ligand binding residues using 3D models of proteins. *BMC Bioinformatics.* 2011;12(1):160. doi:10.1186/1471-2105-12-160.
122. Roy A, Yang J, Zhang Y. COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. *Nucleic Acids Res.* 2012;40(Web Server issue):W471-7. doi:10.1093/nar/gks372.

123. Chitale M, Hawkins T, Park C, Kihara D. ESG: extended similarity group method for automated protein function prediction. *Bioinformatics*. 2009;25(14):1739-1745. doi:10.1093/bioinformatics/btp309.
124. Saraç Ö, Atalay V, Cetin-Atalay R. GOPred: GO molecular function prediction by combined classifiers. *PLoS One*. 2010;5(8):1-11. doi:10.1371/journal.pone.0013711.
125. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*. 2005;21(18):3674-3676. doi:10.1093/bioinformatics/bti610.
126. Punta M, Ofran Y. The rough guide to in silico function prediction, or how to use sequence and structure information to predict protein function. *PLoS Comput Biol*. 2008;4(10):e1000160. doi:10.1371/journal.pcbi.1000160.
127. Pei J, Grishin N. AL2CO: calculation of positional conservation in a protein sequence alignment. *Bioinformatics*. 2001;17(8):700-712.
128. Capra J a, Singh M. Predicting functionally important residues from sequence conservation. *Bioinformatics*. 2007;23(15):1875-1882. doi:10.1093/bioinformatics/btm270.
129. Wass MN, Sternberg MJE. ConFunc--functional annotation in the twilight zone. *Bioinformatics*. 2008;24(6):798-806. doi:10.1093/bioinformatics/btn037.
130. Weingart U, Lavi Y, Horn D. Data mining of enzymes using specific peptides. *BMC Bioinformatics*. 2009;10:446. doi:10.1186/1471-2105-10-446.
131. Arakaki AK, Huang Y, Skolnick J. EFICAZ2: enzyme function inference by a combined approach enhanced by machine learning. *BMC Bioinformatics*. 2009;10:107. doi:10.1186/1471-2105-10-107.
132. López G, Valencia A, Tress ML. Firestar--Prediction of Functionally Important Residues Using Structural Templates and Alignment Reliability. *Nucleic Acids Res*. 2007;35(Web Server issue):W573-7. doi:10.1093/nar/gkm297.
133. Tress ML, Jones D, Valencia A. Predicting Reliable Regions in Protein Alignments from Sequence Profiles. *J Mol Biol*. 2003;330(4):705-718. doi:10.1016/S0022-2836(03)00622-3.
134. Tress ML, Grana O, Valencia A. SQUARE--determining reliable regions in sequence alignments. *Bioinformatics*. 2004;20(6):974-975. doi:10.1093/bioinformatics/bth032.
135. Stockwell GR, Thornton JM. Conformational diversity of ligands bound to proteins. *J Mol Biol*. 2006;356(4):928-944. doi:10.1016/j.jmb.2005.12.012.
136. Soro S, Tramontano A. The prediction of protein function at CASP6. *Proteins*. 2005;61 Suppl 7(April):201-213. doi:10.1002/prot.20738.
137. Pellegrini-Calace M, Soro S, Tramontano A. Revisiting the prediction of protein function at CASP6. *FEBS J*. 2006;273(13):2977-2983. doi:10.1111/j.1742-4658.2006.05309.x.
138. Lopez G, Rojas A. Assessment of predictions submitted for the CASP7 function prediction category. *Proteins Struct Funct Bioinforma*. 2007;(69(Suppl 8)):165-174. doi:10.1002/prot.
139. Lopez G, Ezkurdia I, Tress ML. Assessment of ligand binding residue predictions in CASP8. *Proteins Struct Funct Bioinforma*. 2009;77(S9):138-146. doi:10.1002/prot.22557.
140. Schmidt T, Haas J, Gallo Cassarino T, Schwede T. Assessment of ligand-binding residue predictions in CASP9. *Proteins*. 2011;79 Suppl 1:126-136. doi:10.1002/prot.23174.
141. Gallo Cassarino T, Bordoli L, Schwede T. Assessment of ligand binding site predictions in CASP10. *Proteins*. 2014;82 Suppl 2(June):154-163. doi:10.1002/prot.24495.
142. Radivojac P, Clark WT, Oron TR, et al. A large-scale evaluation of computational protein function prediction. *Nat Methods*. 2013;10(3):221-227. doi:10.1038/nmeth.2340.

143. Altschul SF, Madden TL, Schäffer a a, *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 1997;25(17):3389-3402.
144. Schäffer AA, Wolf YI, Ponting CP, Koonin E V, Aravind L, Altschul SF. IMPALA : matching a protein sequence against a position-specific score matrices. *Bioinformatics.* 1999;15(12):1000-1011.
145. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 2004;32(5):1792-1797. doi:10.1093/nar/gkh340.
146. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics.* 2006;22(13):1658-1659. doi:10.1093/bioinformatics/btl158.
147. Remmert M, Biegert A, Hauser A, ding JS ouml. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods.* December 2011:1-6. doi:10.1038/nmeth.1818.
148. Heller S, McNaught A, Stein S, Tchekhovskoi D, Pletnev I. InChI - the worldwide chemical structure identifier standard. *J Cheminform.* 2013;5:7. doi:10.1186/1758-2946-5-7.
149. Chambers J, Davies M, Gaulton A, *et al.* UniChem: a unified chemical structure cross-referencing and identifier tracking system. *J Cheminform.* 2013;5(1):3. doi:10.1186/1758-2946-5-3.
150. Weininger D. SMILES, A Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J Chem Inf Model.* 1988;28:31-36. doi:10.1021/ci00057a005.
151. Berman H, Henrick K, Nakamura H. Announcing the worldwide Protein Data Bank. *Nat Struct Biol.* 2003;10:980. doi:10.1038/nsb1203-980.
152. Tress ML, Cozzetto D, Tramontano A, Valencia A. An analysis of the Sargasso Sea resource and the consequences for database composition. *BMC Bioinformatics.* 2006;7:213. doi:10.1186/1471-2105-7-213.
153. Venter JC, Remington K, Heidelberg JF, *et al.* Environmental genome shotgun sequencing of the Sargasso Sea. *Science.* 2004;304(5667):66-74. doi:10.1126/science.1093857.
154. Bolton EE, Wang Y, Thiessen PA, Bryant SH. PubChem: Integrated Platform of Small Molecules and Biological Activities. *Annu Rep Comput Chem.* 2008;4:217-241. doi:10.1016/S1574-1400(08)00012-1.
155. Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.* 2014;42:D199-205. doi:10.1093/nar/gkt1076.
156. Hastings J, de Matos P, Dekker A, *et al.* The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Res.* 2013;41:D456-63. doi:10.1093/nar/gks1146.
157. Gaulton A, Bellis LJ, Bento a P, *et al.* ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* 2012;40(Database issue):D1100-7. doi:10.1093/nar/gkr777.
158. Knox C, Law V, Jewison T, *et al.* DrugBank 3.0: a comprehensive resource for “omics” research on drugs. *Nucleic Acids Res.* 2011;39(Database):D1035-D1041. doi:10.1093/nar/gkq1126.
159. Caspi R, Altman T, Billington R, *et al.* The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res.* 2014;42(Database issue):D459-71. doi:10.1093/nar/gkt1103.
160. Whirl-Carrillo M, McDonagh EM, Hebert JM, *et al.* Pharmacogenomics knowledge for personalized medicine. *Clin Pharmacol* 2012;92(4):414-417. doi:10.1038/clpt.2012.96.

161. Vujcic-Zagar A, Pijning T, Kralj S, *et al.* Crystal structure of a 117 kDa glucansucrase fragment provides insight into evolution and product specificity of GH70 enzymes. *Proc Natl Acad Sci U S A.* 2010;107:21406-21411. doi:10.1073/pnas.1007531107.
162. Kasampalidis IN, Pitas I, Lyroudia K. Conservation of metal-coordinating residues. *Proteins Struct Funct Bioinforma.* 2007;68(1):123-130. doi:10.1002/prot.21384.
163. Alberts IL, Nadassy K, Wodak SJ. Analysis of zinc binding sites in protein crystal structures. *Protein Sci.* 1998;7(8):1700-1716. doi:10.1002/pro.5560070805.
164. Yang W, Lee H-W, Hellinga H, Yang JJ. Structural analysis, identification, and design of calcium-binding sites in proteins. *Proteins.* 2002;47(3):344-356. doi:10.1002/prot.10093.
165. Fariselli P, Rossi I, Capriotti E, Casadio R. The WWWH of remote homolog detection: The state of the art. *Brief Bioinform.* 2006;8(2):78-87. doi:10.1093/bib/bbl032.
166. Saripella GV, Sonnhammer ELL, Forslund K. Sequence analysis Benchmarking the next generation of homology inference tools. *Bioinformatics.* 2016;32(June):2636-2641. doi:10.1093/bioinformatics/btw305.
167. Harrow J, Denoeud F, Frankish A, *et al.* GENCODE: producing a reference annotation for ENCODE. *Genome Biol.* 2006;7(Suppl 1):S4. doi:10.1186/gb-2006-7-s1-s4.
168. Lipscomb WN, Sträter N. Recent Advances in Zinc Enzymology. *Chem Rev.* 1996;96(7):2375-2434.
169. Berg JM, Shi Y. The galvanization of biology: a growing appreciation for the roles of zinc. *Science.* 1996;271(5252):1081-1085.
170. Batsanov SS. Van der Waals Radii of Elements. *Inorg Mater Transl from Neorg Mater Orig Russ Text.* 2001;37(9):871-885. doi:10.1023/A:1011625728803.
171. Alvarez S. A cartography of the van der Waals territories. *Dalt Trans.* 2013;42(24):8617. doi:10.1039/c3dt50599e.
172. Stamatoyannopoulos J, Dutta A, Guigó R, *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature.* 2007;447:799-816.
173. Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M. An integrated encyclopedia of DNA elements in the human genome. *Nature.* 2012;489:57-74. doi:10.1038/nature11247.
174. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet.* 2008;40(12):1413-1415. doi:10.1038/ng.259.
175. Flicek P, Amode MR, Barrell D, *et al.* Ensembl 2014. *Nucleic Acids Res.* 2014;42(D1):749-755. doi:10.1093/nar/gkt1196.
176. Tress ML, Wesselink JJ, Frankish A, *et al.* Determination and validation of principal gene products. *Bioinformatics.* 2008;24(1):11-17. doi:10.1093/bioinformatics/btm547.
177. Rodriguez JM, Maietta P, Ezkurdia I, *et al.* APPRIS: annotation of principal and alternative splice isoforms. *Nucleic Acids Res.* 2013;41(Database issue):D110-7. doi:10.1093/nar/gks1058.
178. Ezkurdia I, Juan D, Rodriguez JM, *et al.* Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. *Hum Mol Genet.* 2014;23(22):5866-5878. doi:10.1093/hmg/ddu309.
179. Maietta P, Lopez G, Carro A, *et al.* FireDB: a compendium of biological and pharmacologically relevant ligands. *Nucleic Acids Res.* 2014;42(Database issue):D267-72. doi:10.1093/nar/gkt1127.
180. Lopez G, Maietta P, Rodriguez JM, Valencia A, Tress ML. firestar--advances in the prediction of functionally important residues. *Nucleic Acids Res.* 2011;39(Web Server issue):W235--41. doi:10.1093/nar/gkr437.

181. Apweiler R, Bairoch A, Wu CH, *et al.* UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.* 2004;32(Database issue):D115-9. doi:10.1093/nar/gkh131.
182. Moriaud F, Richard SB, Adcock SA, *et al.* Identify drug repurposing candidates by mining the Protein Data Bank. *Brief Bioinform.* 2011;12(4):336-340. doi:10.1093/bib/bbr017.
183. Lamb J. The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease. *Science (80-)*. 2006;313(5795):1929-1935. doi:10.1126/science.1132939.
184. Kissa M, Tsatsaronis G, Schroeder M. Prediction of drug gene associations via ontological profile similarity with application to drug repositioning. *Methods.* 2015;74:71-82. doi:10.1016/j.ymeth.2014.11.017.
185. Gottlieb A, Stein GY, Ruppin E, Sharan R. PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Mol Syst Biol.* 2014;7(1):496-496. doi:10.1038/msb.2011.26.
186. Ahmed A, Smith RD, Clark JJ, Jr JBD, Carlson HA. Recent improvements to Binding MOAD: A resource for protein-ligand Binding affinities and structures. *Nucleic Acids Res.* 2015;43(D1):D465-D469. doi:10.1093/nar/gku1088.
187. Murakami Y, Omori S, Kinoshita K. NLDB: a database for 3D protein- ligand interactions in enzymatic reactions. *J Struct Funct Genomics.* 2016;17(4):101-110. doi:10.1007/s10969-016-9206-0.
188. Pan Z, Wang B, Zhang Y, *et al.* dbPSP: a curated database for protein phosphorylation sites in prokaryotes. *Database (Oxford).* 2015;2015:bav031. doi:10.1093/database/bav031.
189. Yu J-F, Dou X-H, Sha Y-J, *et al.* DisBind: A database of classified functional binding sites in disordered and structured regions of intrinsically disordered proteins. *BMC Bioinformatics.* 2017;18(1):206. doi:10.1186/s12859-017-1620-1.
190. Yan C, Wu F, Jernigan RL, Dobbs D, Honavar V. Characterization of protein-protein interfaces. *Protein J.* 2008;27(1):59-70. doi:10.1007/s10930-007-9108-x.
191. Carro A, Tress M, de Juan D, *et al.* TreeDet: A web server to explore sequence space. *Nucleic Acids Res.* 2006;34(WEB. SERV. ISS.):110-115. doi:10.1093/nar/gkl203.
192. Rodriguez-Rivas J, Marsili S, Juan D, Valencia A. Conservation of coevolving protein interfaces bridges prokaryote–eukaryote homologies in the twilight zone. *Proc Natl Acad Sci.* 2016;113(52):15018-15023. doi:10.1073/pnas.1611861114.
193. Reva B, Antipin Y, Sander C. Determinants of protein function revealed by combinatorial entropy optimization. *Genome Biol.* 2007;8(11):R232. doi:10.1186/gb-2007-8-11-r232.
194. Grabowski M, Niedzialkowska E, Zimmerman MD, Minor W. The impact of structural genomics: the first quinquennial. *J Struct Funct Genomics.* 2016;17(1). doi:10.1007/s10969-016-9201-5.
195. Di Tommaso P, Moretti S, Xenarios I, *et al.* T-Coffee: A web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. *Nucleic Acids Res.* 2011;39(SUPPL. 2):13-17. doi:10.1093/nar/gkr245.
196. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol Biol Evol.* 2013;30(4):772-780. doi:10.1093/molbev/mst010.
197. Brylinski M, Feinstein WP. eFindSite: improved prediction of ligand binding sites in protein models using meta-threading, machine learning and auxiliary ligands. *J Comput Aided Mol Des.* 2013;27(6):551-567. doi:10.1007/s10822-013-9663-5.
198. Zhang C, Freddolino PL, Zhang Y. COFACTOR: improved protein function prediction by combining structure, sequence and protein–protein interaction information. *Nucleic Acids Res.* 2017:1-9. doi:10.1093/nar/gkx366.

199. Huntley RP, Sawford T, Mutowo-Meullenet P, *et al.* The GOA database: Gene Ontology annotation updates for 2015. *Nucleic Acids Res.* 2015;43(D1):D1057-D1063. doi:10.1093/nar/gku1113.

8 APPENDIX

| APPENDIX

COMPOUND	
COMPID	varchar(3)
NAME	text
CHEMCS	varchar(50)
CLASS	varchar(7)
FORMULA	varchar(50)
WEIGHT	float
ATOMSUM	smallint(5)
SYNONIM	text
PARENT	char(11)
FREQ_35	int(6)
FREQ_40	int(6)
FREQ_45	int(6)

INFOACC	
CADID	varchar(5)
PDBID	varchar(4)
UNIACC1	varchar(6)
UNIACC2	varchar(6)
UNIACC3	varchar(6)
EC1	varchar(15)
EC2	varchar(15)
EC3	varchar(15)
EC4	varchar(15)
HEADER	text
PDBTITLE	text
UNITITLE1	text
UNITITLE2	text
UNITITLE3	text
MUTANTS	text
OTHER	text
BIUNIT	text
GOIDS	text
GONAMES	text
GOEVIDENCE	text

CONSENSUS	
CADID	varchar(5)
CLUSTID	varchar(5)
PDBID	varchar(5)
CONSEQ	text
RELENT	text
PDBSEQ	text
CONNUM	text
PDBNUM	text
PROPRT	text
CONLENGTH	int(5)
PDBLENGTH	int(5)
NUMSEQS	int(4)
NRDB97	text
CLUSTCODES	text
LINKEDCHADS	text

SITE35	
SITEID	int(9)
CADID	varchar(5)
PDBID	varchar(4)
CLUSTID	varchar(5)
PDBRES	text
NUMPDBRES	text
CONRES	text
NUMCONRES	text
SITETYPE	char(3)
CSANOTE	text
COMPID	varchar(3)
LIGPDBNUM	int(5)
LIGCHAIN	varchar(1)
GOMATCH	text
BINDID	int(9)

SITE40	
SITEID	int(9)
CADID	varchar(5)
PDBID	varchar(4)
CLUSTID	varchar(5)
PDBRES	text
NUMPDBRES	text
CONRES	text
NUMCONRES	text
SITETYPE	char(3)
CSANOTE	text
COMPID	varchar(3)
LIGPDBNUM	int(5)
LIGCHAIN	varchar(1)
GOMATCH	text
BINDID	int(9)

SITE45	
SITEID	int(9)
CADID	varchar(5)
PDBID	varchar(4)
CLUSTID	varchar(5)
PDBRES	text
NUMPDBRES	text
CONRES	text
NUMCONRES	text
SITETYPE	char(3)
CSANOTE	text
COMPID	varchar(3)
LIGPDBNUM	int(5)
LIGCHAIN	varchar(1)
GOMATCH	text
BINDID	int(9)

CSITE35	
CSITEID	int(6)
CLUSTID	varchar(5)
PDBID	varchar(4)
NUMCONRES	text
CONRES	text
OCCUPANCY	text
CSITESUM	int(3)
SITETYPE	char(3)
EVIDENCY	char(3)
COMPIDS	text
NUMSEQS	int(3)
SITEIDS	text
GOIDS	text
GONAMES	text
GOMATCH	text
SCORE	int(1)
LIGTYPE	varchar(4)

CSITE40	
CSITEID	int(6)
CLUSTID	varchar(5)
PDBID	varchar(4)
NUMCONRES	text
CONRES	text
OCCUPANCY	text
CSITESUM	int(3)
SITETYPE	char(3)
EVIDENCY	char(3)
COMPIDS	text
NUMSEQS	int(3)
SITEIDS	text
GOIDS	text
GONAMES	text
GOMATCH	text
SCORE	int(1)
LIGTYPE	varchar(4)

CSITE45	
CSITEID	int(6)
CLUSTID	varchar(5)
PDBID	varchar(4)
NUMCONRES	text
CONRES	text
OCCUPANCY	text
CSITESUM	int(3)
SITETYPE	char(3)
EVIDENCY	char(3)
COMPIDS	text
NUMSEQS	int(3)
SITEIDS	text
GOIDS	text
GONAMES	text
GOMATCH	text
SCORE	int(1)
LIGTYPE	varchar(4)

BINDSITE35	
BINDID	int(9)
CADID	varchar(5)
PDBID	varchar(4)
CLUSTID	varchar(5)
PDBRES	text
NUMPDBRES	text
CONRES	text
NUMCONRES	text
COMPIDS	text
LIGPDBNUMS	text
LIGCHAINS	text
SITEIDS	text

BINDSITE40	
BINDID	int(9)
CADID	varchar(5)
PDBID	varchar(4)
CLUSTID	varchar(5)
PDBRES	text
NUMPDBRES	text
CONRES	text
NUMCONRES	text
COMPIDS	text
LIGPDBNUMS	text
LIGCHAINS	text
SITEIDS	text

BINDSITE45	
BINDID	int(9)
CADID	varchar(5)
PDBID	varchar(4)
CLUSTID	varchar(5)
PDBRES	text
NUMPDBRES	text
CONRES	text
NUMCONRES	text
COMPIDS	text
LIGPDBNUMS	text
LIGCHAINS	text
SITEIDS	text

COMPARE35	
ID	int
CLUSTID	varchar(5)
TEMPID	varchar(5)
CLUSTSITEID	int
TEMPSITEID	int
SIZE1	int(2)
MOTIF1	text
MOTIFNUMS1	text
SIZE2	int(2)
MOTIF2	text
MOTIFNUMS2	text
SQSCORES	text
SQGLOBALMEAN	float
OVERLAP	int(2)
PCENTID	int(2)
COMPIDS1	text
COMPIDS2	text
SITEIDS1	text
SITEIDS2	text

COMPARE40	
ID	int
CLUSTID	varchar(5)
TEMPID	varchar(5)
CLUSTSITEID	int
TEMPSITEID	int
SIZE1	int(2)
MOTIF1	text
MOTIFNUMS1	text
SIZE2	int(2)
MOTIF2	text
MOTIFNUMS2	text
SQSCORES	text
SQGLOBALMEAN	float
OVERLAP	int(2)
PCENTID	int(2)
COMPIDS1	text
COMPIDS2	text
SITEIDS1	text
SITEIDS2	text

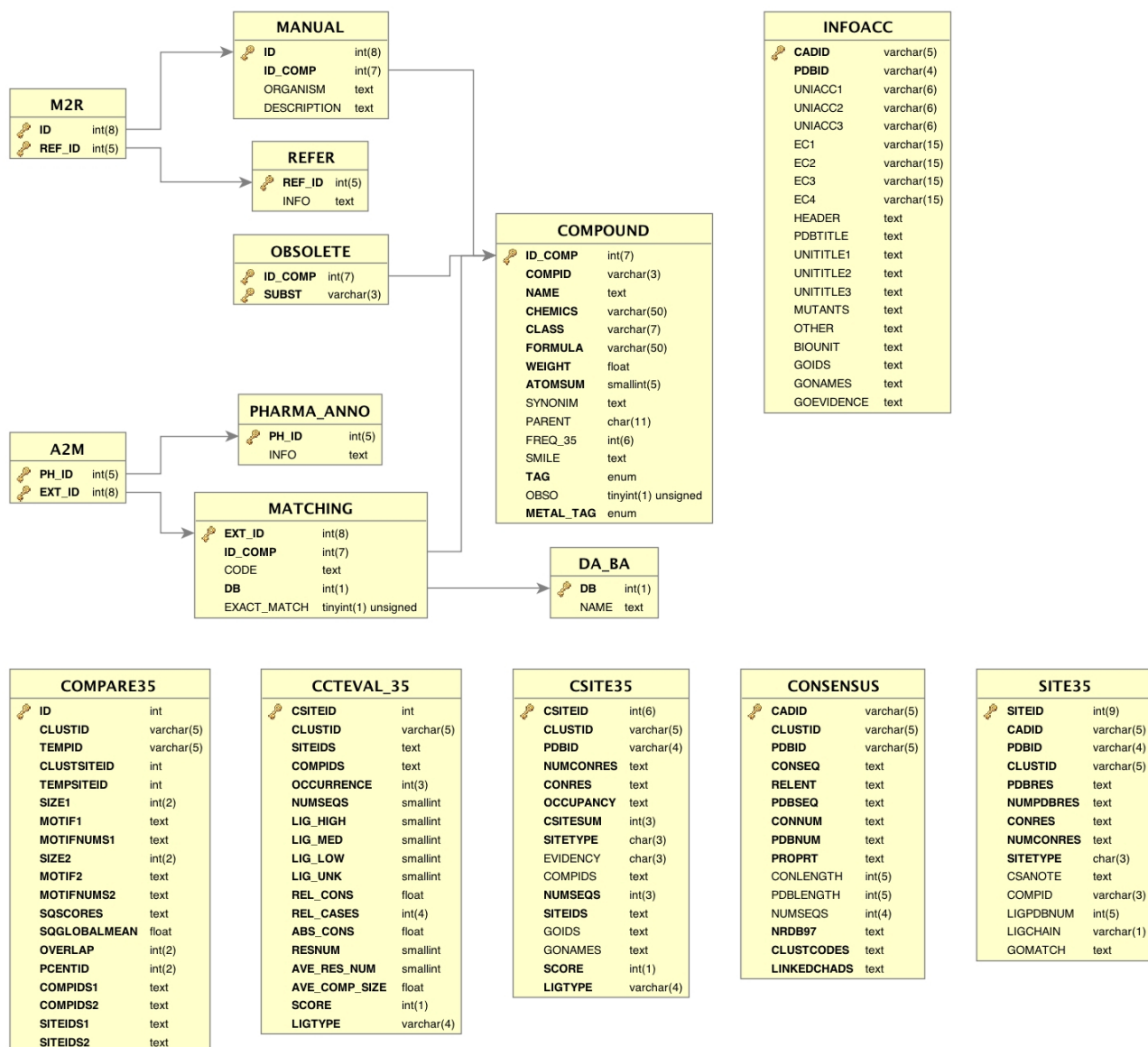
COMPARE45	
ID	int
CLUSTID	varchar(5)
TEMPID	varchar(5)
CLUSTSITEID	int
TEMPSITEID	int
SIZE1	int(2)
MOTIF1	text
MOTIFNUMS1	text
SIZE2	int(2)
MOTIF2	text
MOTIFNUMS2	text
SQSCORES	text
SQGLOBALMEAN	float
OVERLAP	int(2)
PCENTID	int(2)
COMPIDS1	text
COMPIDS2	text
SITEIDS1	text
SITEIDS2	text

CCTEVAL_35	
CSITEID	int
CLUSTID	varchar(5)
SITEIDS	text
COMPIDS	text
OCCURRENCE	int(3)
NUMSEQS	smallint
LIG_HIGH	smallint
LIG_MED	smallint
LIG_LOW	smallint
LIG_UNK	smallint
REL_CONS	float
REL_CASES	int(4)
ABS_CONS	float
RESNUM	smallint
AVE_RES_NUM	smallint
AVE_COMP_SIZE	float
SCORE	int(1)
LIGTYPE	varchar(4)

CCTEVAL_40	
CSITEID	int
CLUSTID	varchar(5)
SITEIDS	text
COMPIDS	text
OCCURRENCE	int(3)
NUMSEQS	smallint
LIG_HIGH	smallint
LIG_MED	smallint
LIG_LOW	smallint
LIG_UNK	smallint
REL_CONS	float
REL_CASES	int(4)
ABS_CONS	float
RESNUM	smallint
AVE_RES_NUM	smallint
AVE_COMP_SIZE	float
SCORE	int(1)
LIGTYPE	varchar(4)

CCTEVAL_45	
CSITEID	int
CLUSTID	varchar(5)
SITEIDS	text
COMPIDS	text
OCCURRENCE	int(3)
NUMSEQS	smallint
LIG_HIGH	smallint
LIG_MED	smallint
LIG_LOW	smallint
LIG_UNK	smallint
REL_CONS	float
REL_CASES	int(4)
ABS_CONS	float
RESNUM	smallint
AVE_RES_NUM	smallint
AVE_COMP_SIZE	float
SCORE	int(1)
LIGTYPE	varchar(4)

Appendix figure 1 FireDB SQL schema. Every yellow box represents a table; all the corresponding fields are listed with the assigned type. Primary keys are marked with a key symbol.



Appendix figure 2 FireDB new version SQL schema. Every yellow box represents a table; all the corresponding fields are listed with the assigned type. Primary keys are marked with a key symbol, and external keys are connected with a directional arrow.

Target	PDB ID	Ligand ID	Contact residues
T0284	3B8I	OXL-MG2	48,49,50,88,159,212,235
T0289	2GU2	ZN	20,23,115
T0292	2CL1	5Z5	12,33,34,66,84-88,90,146,160
T0293	2H00	SAH	34,36,41,69-71,75,76,91-93,97,119,121-124,143-145,147,186,189
T0308	2H57	GTP-MG	10-16,31,34,55,56,59,114,115,117,118,147-149
T0312	2H6L	ZN	89,91,104
T0313	2H58	ADP-MG	7,9,10,12,86-92
T0315	2GZX	NI-NI	6,8,92,128,153,204
T0316	2HMA	SAM-MG	12-14,16,18,19,36-38,100,104,108,126-128,152,155
T0318	2HB6	ZN-ZN	252,257,275,334,336
T0319	2J6A	ZN	11,16,112,115
T0320	2WSI	FAD	59,61,66,106,107,144,148,161,163-165,181,182,185,188,190,300
T0324	2HDO	PO4	9-11,104,105,137
T0329	2HL0	NA	9,11,189
T0330	2HCF	MG	9,11,177
T0332	2HA8	SAH	87-89,110-112,115,129,130,132,137-139,141,144
T0339	2HDY	PLR	71,72,75,117,119,166,205,207,208,228,230,231,267,268
T0341	2H04	PO4-MG	13,15,46,47,179,204
T0348	2HF1	ZN	11,14,29,32
T0369	2HKV	NI	48,123,127
T0371	2HX1	MG	19,21,232
T0372	2HQY	COA	175,190,246,270,272,275

Appendix table 1 List of the targets included in the CASP7 official assessment paper¹³⁸: for every target are listed the bound ligand PDB ID, and the numbering for the contacting residues (distance cut-off: Van der Waals atomic radii + 0.5 Å).

Target	PDB ID	Ligand ID	Contact residues	Neutrals
T0391	3D89	FES	57,59-62,80,82,83,85	-
T0394	3DCY	PO4	15,16,22,28,66,94,203,204	17-19,21
T0396	3GWL	FAD	3,4,7,8,10,11,15,44,48,49,52,75,77-79,81,82,84,85,87,90,95,98	-
T0406	3DI5	NI	48,127,131	-
T0407	3E38	ZN(3)	44,46,51,76,113,122,157,214,216	-
T0410	3D3L	FE	206,211,386	198,201-203,207,390,393,397,403,439,440,442,443
T0422	3D8B	ADP	78,79,85-87,127-132,259,288,289,292	183,184
T0425	3CZX	ZN	11,25,77	17,18,78-80,144
T0426	3DA2	ZN	117,119,142	144,166,221-223,232
T0430	3DLZ	AMP-MG	49-51,54,57,68,70,116,164-168,170,212,213,215,245,246	52,53
T0431	3DAX	HEM	84,112,116,268,269,272,273,276,343,419,420,425,427-429,432,433,436,471	-
T0440	3DCP	FE(2)-ZN	6,8,14,40,93,123,181,258,260	-
T0444	2VUX	FE	135,198,232,235	-
T0450	3DA1	FAD	24,27-29,47-49,54-57,60,61,63,65,191,193,228-231,235,252,254,292,338-340,372-375	277-280,293-295
T0453	3DED	CA(3)	76-78,83	-
T0457	3DEV	MG	29,83,106,158	-
T0461	3DH1	ZN	75,111,114	-
T0470	3DJB	MG	29,58,59,122	-
T0476	2K5C	ZN	4,7,47,50	-
T0477	3DKP	ADP	49,51,53,56,75-80	-
T0478	3D19	MG-FE	30,117,121,154,158,248,252	-
T0480	2K4X	ZN	21,24,39,42	-
T0483	3DLS	ADP-MG(2)	32,33,40,53,55,92,109-111,114,116,159,160,162,172,173	34-36,108,155,157,158
T0485	3DLC	SAM	8,16,28,50-53,72-74,77,99-101,117-119,122,123	-
T0487	3F73	MG	478,546,548,660	-
T0490	3DME	FAD	10,11,13-15,33-35,43,44,46-48,50,52,171-173,204-206,208,234,272,315,316,348-354	-
T0508	3DOU	SAM	22,46-52,67-69,82-85,111-113,151	-

Appendix table 2 List of the targets included in the CASP8 official assessment paper¹³⁹: for every target are listed the bound ligand PDB ID (between parentheses the occurrence of the ligand in the crystal), and the numbering for the contacting residues (distance cut-off: Van der Waals atomic radii + 0.5 Å).

Target	Ligand ID	Type	Interface	Contacting residues numbering
T0652	AMP	Non-metal	No	74,79,80,99,100-104,165,180,182,183
T0657	ZN	Metal	No	121,132,133,143
T0659	ZN(2)	Metal	No	43,48,63
T0675	ZN(2)	Metal	No	21,24,37,42,49,52,65,70
T0686	MG	Metal	No	28,30,103
T0696	NA	Metal	No	18,69,104
T0697	LLP(2)	Non-metal	A-A	91,150-152,190,243,245,247,272,274,301,303,304,351
T0706	MG(2)	Metal	A-A	25,27,99,101,129,130
T0720	MN(10)/SF4(10)	Metal	No	32,34,35,62,99,113-115,182,188,191,194,197,200
T0721	FAD(2)	Non-metal	No	10,12-14,33-39,42,45,46,60,78-80,109-111,114,126,136,235,237,268,269,277,278,281
T0726	ZN	Metal	No	273,277,307
T0737	FAD	Non-metal	No	37,40-42,44,45,49,78,83,114,117,118,120,121,123,124,128,130,135,138,174,237

Appendix table 3 List of the targets evaluated in the CASP10 experiment, as extracted from the official assessment paper¹⁴¹: for every target are listed the bound ligand PDB ID, type, if it is located at the interface in the crystal structure and the numbering for the contacting residues (distance cut-off: Van der Waals atomic radii + 0.5 Å)

Group	Name	T0652	T0657	T0659	T0661	T0675	T0682	T0686	T0687	T0694	T0696	T0697	T0706	T0715	T0720	T0721	T0726	T0732	T0737	T0738	T0744	T0745	T0754	MCC mean
FN119	firestar	0.897	0.793	-	0.572	0.929	0.821	1.000	1.000	0.660	1.000	0.855	1.000	0.505	0.542	0.744	1.000	0.662	0.824	0.911	0.803	-	0.936	0.823
FN475	CNIO	0.796	0.891	-	0.545	1.000	0.911	0.815	0.815	0.797	0.862	0.822	1.000	0.526	0.774	0.806	0.652	0.731	0.877	0.862	0.755	0.490	0.936	0.794
FN237	zhang	0.897	0.891	-0.057	0.477	0.755	0.938	1.000	1.000	0.743	0.767	0.963	1.000	0.721	0.502	0.760	0.610	0.662	0.902	0.915	0.745	1.000	0.942	0.779
FN349	I-TASSER FUNCTION	0.884	0.891	-0.057	0.477	1.000	0.938	1.000	1.000	0.699	0.767	0.899	0.489	0.730	0.252	0.626	0.610	0.662	0.902	0.915	0.700	1.000	0.942	0.742
FN326	SP-ALIGN	0.884	0.793	0.445	0.497	0.687	0.840	0.815	1.000	0.519	0.813	0.827	1.000	0.866	0.607	0.774	0.403	0.548	0.733	0.868	0.636	0.455	0.936	0.725
FN208	COFACTOR human	0.797	0.720	-	0.477	1.000	0.828	1.000	1.000	0.797	0.657	0.963	0.569	0.398	0.252	0.704	0.544	0.665	0.804	0.727	0.591	0.596	0.644	0.702
FN285	McGuffin	0.852	0.000	-	-	0.687	0.933	1.000	1.000	-	0.315	0.886	0.864	0.827	0.311	0.735	1.000	0.588	0.799	0.733	0.622	0.303	0.644	0.689
FN236	3DLigandSite	-	-0.015	-	-	-	0.799	1.000	-	0.716	0.315	0.855	-	-	-	0.778	0.705	0.638	0.880	0.829	0.601	-	-	0.675
FN227	COFACTOR	0.797	0.720	-0.043	0.477	0.438	0.821	0.437	1.000	0.797	0.657	0.963	1.000	0.519	0.252	0.704	0.574	0.665	0.804	0.727	0.591	0.596	0.644	0.643
FN473	Seok	0.742	-0.015	-	0.616	0.860	0.828	0.864	0.771	-	0.228	0.784	0.864	0.747	0.348	0.710	0.513	0.588	0.796	0.822	0.475	-	0.673	0.643
FN221	Atome2_CBS	0.842	-0.015	-	-	-	0.707	-	-	0.620	-	0.921	-	0.736	-	0.675	0.372	-	0.796	0.862	0.482	-	-	0.636
FN082	FNGUSHAK	-	0.000	0.175	0.419	0.495	0.748	0.606	1.000	0.615	0.697	0.871	0.662	0.733	0.293	0.735	0.574	0.622	0.697	0.842	0.747	0.716	0.936	0.628
FN273	IntFOLD2	0.852	0.000	-	-	0.687	0.860	0.864	0.771	0.241	0.315	0.886	0.864	0.827	0.257	0.795	0.865	0.506	0.827	0.733	0.629	-0.025	0.644	0.620
FN128	3DLigandSite2	0.773	-0.015	-	-	-	0.378	0.815	0.815	0.699	-	0.855	-	-	0.252	0.691	0.652	0.662	0.460	0.801	0.695	-	-	0.610
FN430	HHpredA	0.897	0.000	0.309	0.623	-	-0.063	0.772	0.771	0.776	0.439	0.878	0.864	0.678	0.591	0.886	0.610	0.662	0.687	0.821	0.458	0.102	0.936	0.605
FN059	ConPred-UCL	0.773	-0.037	-	0.377	0.627	0.527	0.864	0.864	0.464	0.561	0.150	0.864	0.747	-	0.593	0.773	0.622	0.683	0.702	0.515	-	-	0.593
FN261	Seok-server	0.742	-0.015	-	0.616	0.632	0.828	0.663	0.662	0.357	0.315	0.832	0.864	0.713	0.252	0.597	0.665	0.558	0.796	0.744	0.361	-	0.541	0.586
FN231	Binding Kihara	0.573	0.891	-	0.574	0.936	-0.025	0.223	-0.022	0.302	1.000	0.079	-0.014	0.508	0.301	0.327	-0.005	0.542	0.451	0.578	0.280	-	0.870	0.419
FN471	Chuo binding-sites	0.597	0.256	-	0.525	0.273	0.185	0.180	0.245	0.167	0.309	0.442	0.314	0.386	0.272	0.619	0.163	0.382	0.524	0.545	0.378	0.072	0.286	0.339

Appendix table 4 MCC calculated in our reassessment of CASP10 results for all participating groups. Here disqualified targets T0745 and T0754 are included. All the groups are ordered by descending mean MCC (last column, in red). Results from firestar server are highlighted in green, while results from our human group are highlighted in orange. Best MCCs for every target are highlighted in bold.

Name	Group	FN119	FN475	FN237	FN326	FN349	FN208	FN227	FN285	FN273	FN430	FN473	FN082	FN261	FN059	FN128	FN236	FN231	FN221	FN471
firestar	FN119	-																		
CNIO	FN475	0.64	-																	
zhang	FN237	0.59	0.94	-																
SP-ALIGN	FN326	0.32	0.43	0.39	-															
I-TASSER FUNCTION	FN349	0.18	0.33	0.22	0.87	-														
COFACTOR human	FN208	0.01	0.04	0.02	0.32	0.08	-													
COFACTOR	FN227	0.01	0.01	0.00	0.08	0.16	0.53	-												
McGuffin	FN285	0.02	0.06	0.03	0.10	0.13	0.37	0.63	-											
IntFOLD2	FN273	0.01	0.02	0.01	0.04	0.06	0.20	0.40	0.37	-										
HHpredA	FN430	0.03	0.04	0.03	0.06	0.16	0.30	0.40	0.78	0.94	-									
Seok	FN473	0.00	0.01	0.01	0.02	0.04	0.14	0.27	0.51	0.78	0.93	-								
FNGUSHAK	FN082	0.01	0.01	0.01	0.02	0.05	0.15	0.34	0.67	0.85	0.94	0.99	-							
Seok server	FN261	0.00	0.00	0.00	0.01	0.01	0.06	0.12	0.34	0.47	0.73	0.53	0.76	-						
ConPred-UCL	FN059	0.00	0.00	0.00	0.01	0.01	0.04	0.11	0.21	0.30	0.51	0.44	0.51	0.60	-					
3DLigandSite2	FN128	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.04	0.04	0.04	0.10	0.07	0.11	0.22	-				
3DLigandSite	FN236	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.05	0.05	0.09	0.10	0.04	0.10	0.21	0.76	-			
Binding Kihara	FN231	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.09	0.09	0.08	0.08	0.08	0.10	0.14	0.69	0.82	-		
Atome2 CBS	FN221	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.02	0.02	0.02	0.01	0.08	0.37	0.53	0.84	-	
Chuo binding-sites	FN471	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.27	0.48	0.60	0.87	-

Appendix table 5 Wilcoxon signed rank test results among all the groups. Yellow cells mean statistical significant differences (<0.06 pvalue) for results comparison of the two intersecting groups in the table. Grey cells contain no statistical significant results