Universidad Autónoma de Madrid

Programa de Doctorado en Biociencias Moleculares

La coevolución en regiones de interacción entre proteínas: estudio y desarrollo de métodos computacionales

Juan Rodríguez Rivas

Madrid, 2019

Departamento de Biología Molecular

Facultad de Ciencias

Universidad Autónoma de Madrid

La coevolución en regiones de interacción entre proteínas: estudio y desarrollo de métodos computacionales

TESIS DOCTORAL

Juan Rodríguez Rivas

Ingeniero en Informática

Directores de tesis:

Dr. Alfonso Valencia Herrera

Dr. Michael Liam Tress

Centro Nacional de Supercomputación

Barcelona Supercomputing Center

AGRADECIMIENTOS

He tenido la gran suerte de compartir esta aventura con magníficas personas y excelentes científicos.

Alfonso, muchas gracias por el apoyo, la paciencia y los buenos consejos que siempre has tenido. Echaré de menos tu visión y las discusiones francas y estimulantes. Si ya lo intuía al empezar, este tiempo no ha hecho más que confirmar que esta es mi vocación. Guardaré con cariño la experiencia de haber trabajado en un ambiente tan cordial, interesante y excelente.

Michael, muchas gracias por guiarme pacientemente en mis primeros y vacilantes pasos científicos. Por desvelarme con generosidad el fascinante mundo de las estructuras de proteínas y por poder tachar otro deporte exótico de mi lista.

David y Simo, muchas gracias. Nuestras discusiones a dos bandas y media han sido una de las experiencias que más he disfrutado y con las que más he aprendido durante esta tesis. Habéis dejado secuelas profundas en mi forma de pensar y de plantearme y afrontar las preguntas. Estoy en deuda con vosotros por el tiempo que habéis sacrificado, lo que he aprendido gracias a vuestra generosidad y los buenos ratos que hemos compartido.

A mis infatigables compañeros de doctorado: Jon y María. Jon, tienes el corazón más grande que la cabeza. Solo comparable a tu capacidad para complicar un boceto en el *pictionary*. En verdad, el recuerdo más indeleble de esta tesis será el momento de la alimaña oficinista. O los de la canasta. O el de los experimentos de choque elástico. O el de la telequinesis de María con la celulosa. María, espero que se me haya pegado algo de tu gran bondad y empatía. Aunque no comparta tus gustos por *tappers* con restos innombrables. Echaré de menos tu capacidad para sorprendernos con frases inocentemente intempestivas. He tenido la gran suerte de que hayáis sido mis compañeros y que hayamos compartido experiencias, frustración y, sobre todo, risas y buenos momentos. Sin vosotros, esto no hubiera sido lo mismo.

Y a un montón de gente de la que he aprendido muchas cosas, con las que he compartido buenos momentos y han contribuido a esta tesis sin pretenderlo. Gracias a Fede, un monumento de persona que me introdujo con tanta amabilidad como generosidad en el maravilloso y arrebatador mundo de la evolución. Gracias a Miguel Ponce, mi irremediable y prolijo compañero filosofal y proveedor oficial de las referencias y nuevos mundos más cautivadores. Gracias a los auténticos, en particular a Kiko, Marcelo y Julia. A los erasmus madriditas y algunos más: Miguel, María, Ciro, Rosa, Edu, Manuel, Candi, Marta, Manolo, Caro, Marina y Pilar. A mis compañeros en Barcelona: Davide, Victoria, Vera, Iker, Carlos, Hugo, Eva, Fransuá, Arnau, Eduard, Mónica, Laure, Alba Jené, Alba Lepore, Miguel Madrid, Miguel Vázquez, Victor, Mattia, Salva, Jose María, Patricia, Martin, así como a los mineros y compañeros del departamento. Y a mis compañeros en Madrid: Paolo (un grande, en todos los sentidos), Kristina (Kursk!), Elena, Enrique, Dani, Cesar, Iakes, Mark, Osvaldo, Txema, Ángel, Edu, David, Gonzalo, Mar, Felipe, Curi, Raquel, Florian, Javier, José Manuel, Tirso, Victor y Fátima. *Et al.*

Sobre todo, a los pilares sin los que esto no hubiera sido posible. A mis padres, Chus y Servando, con todo mi cariño. Gracias por vuestro incansable sacrificio, generosidad y abnegación para con nosotros. No podría tener más apoyo ni mejores brújulas para sortear los escollos del camino. A mi gran hermana, Ana, una supernova de vitalidad y entusiasmo que siempre me ha servido de guía y de la que he aprendido algunas de las lecciones más trascendentes. A Francis y a los pequeñines, Héctor y Malena. Al que ha sido, sin duda, el mayor de mis descubrimientos durante esta tesis. Silvia, la determinación forjada a base de bondad. Una luz que ilumina el camino de todos los que tienen la suerte de cruzarse con ella. No puedo haber mayor privilegio que el de haber sido tocado por la fortuna y teneros cerca.

A todos, muchas gracias. Solo espero que mi interacción con vosotros durante este tiempo haya sido la mitad de buena a como ha sido la mía con vosotros. Porque, no lo dudéis, hemos coevolucionado mutuamente durante este camino. Pero esa, es otra historia.

RESUMEN

El funcionamiento celular se sustenta en intrincadas redes de interacciones moleculares. Una de las más comunes e importantes de estas interacciones moleculares son las interacciones físicas entre proteínas. La correcta asociación de proteínas impone fuertes restricciones a la evolución de las correspondientes secuencias. En este contexto, el término coevolución engloba a las interdependencias evolutivas entre proteínas que interaccionan generadas por restricciones estructurales, entre otros factores. Se han desarrollado varios métodos para predecir contactos físicos entre proteínas a partir de covariaciones en alineamientos de secuencias. En la última década, el desarrollo de nuevas metodológicas computacionales y el crecimiento de los datos de secuencias han permitido su mejora. Los objetivos principales de esta tesis son una mayor comprensión del fenómeno de la coevolución en regiones de interacción entre proteínas y la mejora de este tipo de métodos, atendiendo a dos de los problemas que más limitan su ámbito de aplicación: la imposibilidad de predecir contactos sistemáticamente entre proteínas en especies eucariotas y la falta de suficiente información de secuencias en muchas familias.

La primera parte de la tesis se concentra en el desarrollo de métodos computacionales para estudiar la relación existente entre coevolución y conservación estructural de las interfaces a largas distancias evolutivas. La comparación de la señal coevolutiva detectada en alineamientos en procariotas con las divergencias estructurales entre complejos homólogos en procariotas y eucariotas nos ha llevado a descubrir que la señal de coevolución está asociada a un alto grado de conservación estructural. Esto permite proyectar con acierto los contactos predichos en procariotas, donde existen abundantes datos de secuencias, a complejos en eucariotas distantes pero relacionados evolutivamente. De esta forma resulta posible extender el ámbito de aplicación de metodologías basadas en coevolución a complejos de proteínas eucariotas.

En una segunda parte, investigamos el efecto que tienen los factores limitantes de la predicción de contactos: la insuficiente cantidad de secuencias disponibles, los sesgos derivados de la conservación de las posiciones y la falta de independencia entre las secuencias debidas a la filogenia subyacente. Nuestros resultados muestran que existen predicciones de interacciones correctas en casos con pocas secuencias que son difícilmente recuperables sin una metodología adecuada. Proponemos una metodología que, gracias al uso de distribuciones empíricas nulas obtenidas mediante la aleatorización de los alineamientos de partida, nos permite obtener un umbral específico para cada caso haciendo más comparable la señal entre casos. Este procedimiento mejora la calidad de las predicciones de forma notable, a la vez que permite rescatar predicciones correctas a partir de alineamientos con pocas secuencias.

Nuestro trabajo realza el papel de la coevolución en la evolución de las proteínas, en procesos como la divergencia en secuencia y la conservación de la estructura, así como su potencial para la construcción de modelos tridimensionales de un considerable número de interacciones entre proteínas. Temas en los que queda aún un importante margen de progreso, especialmente en lo que respecta a un mejor tratamiento de las relaciones filogenéticas entre las secuencias.

ABSTRACT

Cellular functions are based on convoluted networks of molecular interactions. Protein-protein interactions are one of the most important and prevalent of these interactions. The correct association of proteins imposes strong constraints on the evolution of proteins. In this context, the term coevolution encompasses the evolutive interdependence between interacting proteins due to existing structural constraints, among other factors. Several methods have been developed to predict contacts between proteins from sequence covariation in multiple sequence alignments. In the last decade, the development of new computational methods and the increase of available sequences have improved the contact prediction performance remarkably.

The main objectives of this thesis are a better understanding of sequence coevolution at protein interfaces and the improvement of contact prediction between proteins, with a focus on two of the main challenges in this field: the impossibility of predicting contacts in eukaryotes and the insufficient number of sequences for many protein families.

In the first part of this thesis, we present the development of a computational approach to study the relation between coevolution and structural conservation at protein interfaces over a large evolutionary scale. The comparison of the coevolutionary signal detected in prokaryotic alignments to the structural divergence between prokaryotic and eukaryotic homologs shows that the coevolutionary signal is associated with high structural conservation. This finding enables the correct projection of contact predictions from prokaryotes, where there is abundant sequence data, to distant but evolutionary related eukaryotic complexes. Thus, it is possible to extend the scope of application of coevolutionary methods to eukaryotic complexes.

In the second part, we study the limiting factors of contact prediction between proteins: the reduced number of sequences available, the biases induced by sequence conservation and the lack of independence between sequences due to the underlying phylogeny. Our results show that correct predictions in cases with few sequences are hard to recover using current methodologies. Here we propose a method that uses empirical null distributions obtained through randomizations of the input alignments to estimate a specific threshold for each case that makes the signal more comparable between cases. This method significantly improves the quality of the predictions and recovers correct predictions even for alignments with few sequences.

This work underlines the crucial role of coevolution in protein evolution, in processes such as sequence divergence and structural conservation, as well as its potential to build three-dimensional models for a considerable number of protein-protein interactions. These are areas in which there is still room for improvement, especially in handling the phylogenetic relations among sequences.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

CLAVE DE ABREVIATURAS

DCA: Análisis de acoplamiento directo (del inglés *Direct Coupling Analysis*).

APC: Corrección por el producto promedio (del inglés *Average Product Correction*).

HMM: Modelos ocultos de Markov (del inglés *Hidden Markov Models*).

PCA: Análisis de componentes principales (del inglés *Principal Component Analysis*).

CASP: Evaluación analítica de la predicción de estructura (del inglés *Critical Assessment of Structure Prediction*).

NMR: Resonancia magnética nuclear (del inglés *Nuclear Magnetic Resonance*).

MEND: Máximo obtenido con distribuciones empíricas nulas (del inglés Maximum from Empirical Null Distributions).

PCPC: Proporción de Casos Predichos Correctamente.

IDDI: Interacción entre dominios interproteina (del inglés *Inter-protein Domain-Domain Interaction*)

Nota sobre el uso de anglicismos

Aunque he tratado de minimizar en lo posible el uso de anglicismos, he optado por utilizar algunos anglicismos como *background* (distribución de fondo) y *score* (puntación) ya que tienen asociados significados específicos y resultan más identificables en el ámbito de estudio que nos ocupa, favoreciendo la comprensión de las materias tratadas en esta tesis. En cualquier caso, se acompañan en su primera ocurrencia del mejor equivalente posible en castellano.

1. INTRODUCCIÓN

Esta tesis tiene como su ámbito de estudio la coevolución molecular, más concretamente la coevolución entre proteínas y entre dominios de proteínas y la relación con su estructura. Introduciremos progresivamente las mejoras metodológicas y conceptuales que se han ido produciendo en campo de la coevolución molecular.

1.1 Introducción a la coevolución

1.1.1 Definición y coevolución entre especies

El origen del concepto de coevolución se encuentra en el estudio de las interacciones entre especies en sistemas ecológicos. El término coevolución alude al fenómeno por el cual los cambios en una especie pueden influir en las presiones de selección en otras especies debido a las interacciones entre ellas. Atendiendo a la definición de Thompson la coevolución se refiere a "el cambio evolutivo recíproco en especies que interaccionan" [1]. La prevalencia e importancia de las interacciones entre especies en sistemas ecológicos hace que la coevolución sea una parte fundamental de la teoría evolutiva. Sin embargo, muchas interacciones entre especies están muy acotadas en el tiempo y el espacio, o forman parte de un complejo sistema de interacciones, por lo que no dan lugar a interacciones ecológicas lo suficientemente fuertes y estables para que se produzca un proceso coevolutivo, o éste sea extremadamente difícil de detectar.

Los primeros estudios sobre coevolución se remontan al propio Darwin que observó la relación existente entre las longitudes de los nectarinos de las orquídeas y la proboscis de sus polinizadores [2,3]. La correspondencia en longitud define la viabilidad, intensidad y especificidad de las interacciones entre estas especies, de forma que los polinizadores pueden verse beneficiados por el alimento obtenido y las orquídeas pueden conseguir una forma más efectiva de fecundación. Bajo el influjo de la selección natural, estas interacciones producen patrones de covariación de los caracteres implicados en la interacción, reflejo de su importancia en la evolución de estas especies. Este trabajo precedió a estudios posteriores que utilizan la covariación entre caracteres fenotípicos, como medida observable, para estudiar las interacciones entre especies [4–6]. El término coadaptación se refiere a la coordinación de cambios en caracteres específicos en dos (o más) especies relacionados con la interacción entre ellas [7–9], influyendo en la supervivencia de ambas. Las interacciones asociadas a coadaptaciones suelen ser fuertes, específicas y mantenidas en el tiempo dando lugar a señales fuertes y relativamente fáciles de detectar. Un ejemplo extremo es la coadaptación entre la orquídea de Darwin y su polinizadora cuya proboscis mide unos 30cm [10,11].

Los patrones observables relativos a procesos coevolutivos no se encuentra restringidos a la evolución de caracteres concretos, también se puede observar en escenarios más globales de la coevolución entre especies. Ejemplo de ello es la Regla de Fahrenholz que establece que, en caso de existir coespeciación, existe una elevada similitud entre las filogenias de los parásitos y sus hospedadores [12–14]. Cabe mencionar que este tipo de relaciones, como el mutualismo y parasitismo obligado, se produce una dependencia fuerte

entre las especies [14,15], que no suelen ser comunes [16], debidas a interacciones muy intensas. Por un lado, la mayor parte de interacciones ecológicas, como veremos más adelante, se producen entre multitud de especies en complejas redes ecológicas. Por otro lado, la evolución de las especies se produce en sistemas ecológicos complejos con multitud de factores, no sólo interacciones ecológicas. Por ello, la existencia de covariaciones entre especies que interaccionan puede ser debidas a multitud de factores o fenómenos que afecten de forma similar a la adaptación de estas especies [17] sin relación alguna con sus interacciones.

El término coevolución con entidad propia fue establecido más tarde por Ehrlich y Raven en 1964 en su trabajo sobre la interacción entre mariposas y plantas [4]. Este trabajo puso las bases para el posterior desarrollo del campo motivando un área de conocimiento específica para el estudio de las interacciones entre especies y su importancia para la comprensión de los procesos evolutivos. Supone una generalización del concepto de coadaptación, ya que se refiere a los cambios evolutivos recíprocos debido a la interacción entre especies, sin que esto se vea traducido necesariamente en cambios observables de caracteres concretos. Por lo que se suele asumir que la coevolución engloba a la coadaptación [18].

Uno de los conceptos más influyentes en el contexto de la coevolución es la hipótesis de la Reina Roja formulada por Van Valen en 1973 [19]. Van Valen reconcilia dos ideas aparentemente contradictorias como son la conservación y el cambio. En analogía con la historia en “Alicia a través del espejo y lo que encontró allí”, las especies deben estar en continuo cambio para poder mantenerse en el mismo lugar en un contexto siempre cambiante. Las interacciones con otras especies constituyen parte de este cambio continuo del entorno. La congruencia entre conservación y cambio se debe a que actúan a distintos niveles. Por un lado, se conserva la interacción entre las especies. Por otro, es necesario para adaptarse a los cambios que se producen en otras especies y, de forma más general fuera de un proceso coevolutivo, al resto del entorno. Un ejemplo son las carreras armamentísticas entre presa y depredador, donde la especie depredadora está en constante evolución para contrarrestar los cambios en la presa para evadir al depredador y viceversa [20].

El avance en el campo hizo patente que la coevolución no se limitaba a casos donde existe una interacción específica entre dos (o pocas) especies, sino que en los sistemas ecológicos se suelen encontrar una gran cantidad de interacciones entre especies a diferentes escalas de intensidad. En el extremo, se habla de coevolución difusa cuando el número de interacciones es elevado y la intensidad de las interacciones relativamente débiles, provocando que sean mucho más difícil de detectar [21]. En realidad, es razonable asumir que existe un gradiente de posibilidades desde interacciones muy específicas a más promiscuas, pudiendo ser de distintas intensidades, aunque es esperable que las más específicas tiendan a corresponderse con interacciones más intensas.

En esta línea, la teoría del mosaico geográfico coevolutivo [22] propone que los diferentes ambientes geográficos y temporales afectan al modo en que las especies interactúan pudiendo incluso cambiar el tipo de interacción. Distingue también entre zonas calientes (del inglés *hot-spots*) donde estas relaciones se dan con intensidad y zonas frías (del inglés *cold-spots*) donde sólo está presente una de las especies o no hay

selección recíproca y que suelen englobar geográficamente a zonas calientes [23]. Además, diversos factores genéticos (mutación, flujo genético, deriva, etc.) y ambientales moldean continuamente los rasgos coevolutivos sobre los que la selección natural puede actuar tanto dentro de poblaciones como entre distintas poblaciones [23]. En consecuencia, la teoría de mosaico coevolutivo representa un marco de trabajo más general que integra una mayor diversidad de escenarios y factores dando lugar a una perspectiva más amplia del fenómeno coevolutivo.

1.1.2 Coevolución molecular y jerarquía de niveles

El surgimiento e incremento de datos moleculares durante el siglo XX, en especial datos de secuencias de proteínas y genomas de organismos, permitió empezar a explorar la coevolución a escalas moleculares. Existen ejemplos documentados donde la interacción entre especies está asociada a la coevolución de dos genes concretos, uno por cada especie [24]. En estos casos, tales genes están íntimamente ligados a la interacción entre el par de especies. Se ha mostrado que la coevolución entre genes puede estar asociada a la coevolución entre los residuos de las regiones de interacción de las correspondientes proteínas [25]. De esta forma, se producen cambios recíprocos en las secuencias de los genes implicados que preservan el correcto funcionamiento del complejo de proteínas resultante. Numerosos trabajos también han documentado una coevolución entre residuos de una misma proteína [26]. A partir de estas evidencias se puede hacer una clasificación de los niveles en los que opera la coevolución en función del principal agente involucrado. Podemos distinguir entre la coevolución entre especies, la coevolución entre proteínas y la coevolución entre residuos [26,27] (Figura 1.1). Utilizaremos esta jerarquía para referirnos de forma concisa a estos distintos escenarios, pero no pretende ser una clasificación de todos los niveles existentes ya que otros muchos niveles podrían ser considerados (ecosistemas, poblaciones, nucleótidos, etc.).

Por último para integrar la coevolución molecular en el contexto general de la coevolución, se puede definir la coevolución como “el proceso por el que las interacciones de agentes evolutivos evolucionan acumulando cambios dirigidos por la selección natural en dichos agentes” [27]. Donde estos agentes pueden ser, entre otros, las especies, las proteínas y los residuos de éstas.

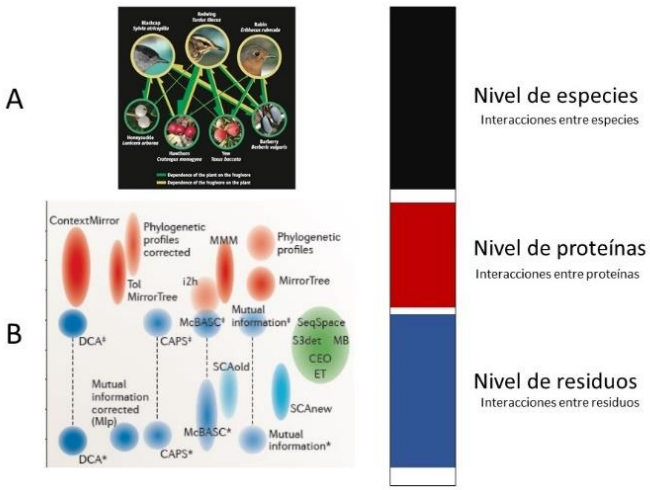


Figura 1.1 Esquema de niveles coevolutivos y métodos de coevolución a nivel molecular. A) Nivel de especies. B) Niveles de coevolución molecular. En la parte izquierda se muestran los métodos diseñados para detectar la coevolución entre proteínas (en rojo) y aquellos para detectar contactos físicos (en azul). En verde, se muestran los métodos relacionados con la detección de residuos funcionales. Fuentes: A) Imagen extraída de [28] B) Imagen extraída de [26].

1.2 Coevolución entre proteínas

Dado que la coevolución entre proteínas se debe a una interdependencia funcional entre ellas, se han desarrollo métodos capaces de predecir interacciones funcionales mediante la detección de señales coevolutivas utilizando información genómica. Se pueden distinguir tres familias de métodos: basados en *perfiles filogenéticos*, basados en similitud de árboles filogenéticos y basados en la fusión de genes.

Los métodos conocidos como *perfiles filogenéticos* se basan en la observación de que las proteínas que interaccionan tienden a aparecer o desaparecer de forma conjunta en los genomas [29,30]. Utilizan genomas de especies representativas para detectar patrones de ausencia y presencia de genes. Muchas interacciones funcionales requieren un conjunto mínimo de proteínas para ser funcional, por lo que si una de ellas desaparece la presión selectiva sobre el resto se reduce. De forma similar, la adquisición y mantenimiento de un gen en un genoma puede depender de la presencia de un conjunto de genes. Las mayores diferencias entre métodos se encuentran en los modelos evolutivos utilizados, la selección de las especies representativas y el tratamiento de los sesgos [26].

Los patrones de aparición y desaparición representan eventos drásticos que obvian relaciones más sutiles de coevolución. Las interdependencias entre proteínas que interaccionan funcionalmente se suelen reflejar en tasas de evolución similares. En analogía con los árboles de especies huéspedes y hospedadoras (sección 1.1.1), sus árboles filogenéticos tienden a ser similares. Esta observación, detectada primero en casos concretos [31], ha constituido la base de un importante grupo de métodos [26]. La primera metodología sistemática de este tipo estimaba la similitud de los árboles filogenéticos mediante la comparación de las distancias entre las secuencias de los alineamientos de ortólogos [32]. Contribuciones posteriores redujeron el impacto del árbol de las especies en el cálculo de estas similitudes [33]. Otra mejora importante calcula primero todas las similitudes entre pares de proteínas y extrae la parte de la señal de cada par que es explicable por una tercera, obteniendo mejores predicciones [34].

Por su parte, los métodos basados en la fusión de genes [29,35] se basan en la observación de que, partiendo de un par de proteínas que interaccionan, es posible encontrar pares de proteínas homologas a las de partida codificadas en genes distintos en algunos genomas y en el mismo gen en otros genomas. Estos eventos de fusión pueden deberse a que al encontrarse fusionados como dominios de una misma proteína se facilita su interacción, permitiendo mayores concentraciones del complejo activo, mientras que si se encuentran codificados en genes distintos tendrán de encontrarse por difusión [29].

1.3 Coevolución entre residuos

1.3.1 Cambios coordinados y contactos físicos

La selección natural impone fuertes restricciones sobre la evolución de las proteínas para mantener su función. Las variaciones en secuencia se encuentran restringidas tanto por motivos estructurales (p. ej. estabilidad, plegamiento, interacciones) como funcionales (p. ej. sitios catalíticos y unión de ligandos). La

eclosión e incremento de datos moleculares durante el siglo XX permitió la exploración de la evolución de las distintas posiciones de las familias de proteínas. Entre las aportaciones más relevantes se encuentra la observación de que las posiciones más conservadas en alineamientos múltiple de secuencias homólogas se encuentran asociadas a roles funcional o estructurales importantes [36]. La conservación de las posiciones continúa siendo una fuente inagotable de información evolutiva y funcional. Por ejemplo, sigue siendo la fuente de información más importante en los métodos de predicción del efecto de mutaciones [37].

La primera observación relacionada a la coevolución a nivel de residuos se remonta a 1987. Altschuh *et al.* alinearon las secuencias de 7 proteínas homólogas de la cápside del virus del mosaico del tabaco, estudiaron sus patrones de variación y sus posiciones en la única estructura disponible a la época [38]. Como resultado, observaron que las posiciones que mostraban patrones idénticos de variación, es decir un cambio de aminoácido en una de esas posiciones estaba asociado a un cambio en la otra, se encontraban en su mayoría espacialmente cercanos en la estructura. En buena parte de estas variaciones se podía explicar la posible causa de estas covariaciones [38]. Por ejemplo, los cambios en parejas de posiciones en el núcleo de la proteína presentaban complementariedad de tamaño. En otras regiones, un buen número de combinaciones consistían o bien en parejas de residuos hidrofóbicos o en parejas con carga. Esto llevó a plantear que estas covariaciones eran debidas a cambios compensatorios que mantuvieran la interacción física. Estudios posteriores generalizaron esta observación concluyendo que se trata de un fenómeno común [39–42].

Estas contribuciones dieron lugar al establecimiento de una importante línea de investigación sobre la correlación entre cambios coordinados en alineamientos de secuencias y contactos físicos en estructuras [40–42]. Entre ellos, cabe destacar el estudio de Göbel *et al.* que proponía la existencia de una relación directa entre coevolución y contactos en estructuras y su potencial para mejorar las predicciones de estructura de proteínas [40], como luego se demostró [43,44]. A estos cambios coordinados se les denominó mutaciones correlacionadas o coordinadas. Los primeros trabajos permitieron la automatización de la aproximación y su generalización a un número creciente de familias de proteínas.

Cabe preguntarse cuál es la causa de esta relación entre covariación en secuencia y contactos físicos. Numerosos estudios han observado que mutaciones deletéreas pueden ser compensadas por otras mutaciones en su entorno espacial próximo, evitando su efecto perjudicial en adecuación biológica (*fitness*) [45]. Esto es, se producen cambios mutuos y específicos en pares residuos bajo el influjo de la selección natural. En analogía con la terminología en el contexto de especies (sección 1.1.1), nos referimos a estas adaptaciones específicas de residuos como coadaptaciones [11]. Sin embargo, la existencia de una fuerte covariación asociada a un contacto físico no implica necesariamente que exista coadaptación, aunque esto sea probable. Por otro lado, dado que estas mutaciones compensatorias han sido observadas y proveen de un mecanismo biofísico plausible, es la interpretación dominante en el campo. Aunque es necesario recalcar que es posible que haya otros factores que puedan explicar esta relación, como tasas evolutivas similares [46,47]. Una perspectiva más clara sobre este asunto se ve entorpecida por la gran cantidad de escenarios

posibles, la probable presencia de señales sutiles indetectables y la dificultad para distinguir la parte de la covariación debida a la señal filogenética como veremos más adelante.

1.3.2 Predicción de estructura

La estructura de las proteínas es una información de valor incalculable para entender su función, evolución y las patologías asociadas a ellas. Sin embargo, la determinación experimental de la estructura de las proteínas sigue siendo un proceso costoso y laborioso. Tras más de 50 años de investigación sobre el plegamiento de las proteínas, es sabido que, en casi la totalidad de los casos, las proteínas adaptan de forma espontánea su conformación nativa en condiciones fisiológicas [48]. Esto es, la secuencia de las proteínas codifica su estructura tridimensional. De este principio se deduce que toda la información necesaria para determinar la estructura de las proteínas está contenida en su secuencia [49]. Dada la importancia de la estructura de las proteínas y la posibilidad de deducirla solo a partir de su secuencia, la comunidad científica ha realizado un ingente esfuerzo para tratar este resolver este problema.

Prueba de ello es la existencia continuada durante más 25 años de CASP (del inglés *Critical Assessment of Structure Prediction*). CASP es un experimento comunitario para establecer de forma lo más objetiva posible el progreso en el campo de la predicción de estructura. En cada edición de CASP, los asesores del experimento obtienen estructuras tridimensionales de proteínas obtenidas experimentalmente gracias a la colaboración de grupos experimentales. Durante la duración del experimento, de aproximadamente 4 meses, las secuencias de las proteínas cuya estructura es conocida solo por los asesores son hechas públicas progresivamente. Los grupos inscritos en el experimento envían modelos tridimensionales de las proteínas objetivo y, posteriormente, los asesores evalúan la bondad de los modelos enviados. Las diversas ediciones de CASP han constatado un progresivo avance caracterizado por pequeños saltos cuantitativos [50], aunque aún estamos lejos de una solución generalizada del problema. Han servido también para poner de relieve los principales avances entre los que podemos destacar el uso de HMM para la detección de homólogos remotos [51], el uso de bibliotecas de fragmentos [52], la predicción de contactos mediante coevolución y el uso del aprendizaje profundo [53,54].

La estructura de las proteínas está altamente conservada en comparación con sus secuencias. Es bien sabido que, en la inmensa mayoría de casos, las estructura de proteínas homólogas es similar incluso si sus secuencias son muy divergentes [55]. Dada la secuencia de una proteína, podemos utilizar la estructura de una proteínas homóloga para, a partir de ella, realizar un modelo de la estructura de la proteína objetivo [56]. La proteína homóloga usada se denomina plantilla (*template*). La búsqueda de una plantilla se puede realizar comparando la secuencia de la proteína objetivo contra una base de datos de secuencias de proteínas con estructura conocida. El grado de similitud entre la estructura de proteínas homólogas es inversamente proporcional, en términos generales, al grado divergencia. En consecuencia, la calidad de los modelos generados de esta forma es proporcional a la identidad en secuencia, es decir, a la proporción de aminoácidos idénticos en el alineamiento de las secuencias de las proteínas objetivo y de la plantilla [56–59]. De forma

orientativa, para identidades de secuencias superiores al 50% se consiguen modelos cercanos a los que se podrían obtener mediante métodos experimentales, mientras que para identidades en secuencia inferiores al 30%, la calidad de los modelos decae rápidamente [56]. En cualquier caso, si es posible encontrar una plantilla, la tarea de modelado será más sencilla y el modelo resultante de una mayor calidad [60]. Debido a ello, en CASP se distinguen dos categorías: modelado basado en plantillas o TBM (del inglés *Template Based Modeling*) y modelado libre o FM (del inglés *Free Modelling*).

Dos conclusiones principales se pueden extraer de estos experimentos comunitarios de acuerdo con las últimas ediciones: i) La forma más fiable de realizar un modelo estructural sigue siendo por medio del uso de plantillas [61]; ii) en tiempos recientes la predicción de contactos, principalmente basada en coevolución y/o aprendizaje profundo que detallaremos más adelante, está incrementando considerablemente la capacidad de realizar buenos modelos estructurales sin la ayuda de plantillas [62,63].

Sabemos que con unas buenas restricciones espaciales es posible determinar con un alto nivel de detalle la estructura de las proteínas [64,65]. Por lo que la inclusión de predicciones de contactos en los métodos de modelado de estructura ha sido común. Históricamente, a pesar de que la calidad de las predicciones de contactos era limitada, su integración dentro de protocolos de predicciones de estructura ha resultado útil [66,67], en particular con las progresivas mejoras metodológicas en la predicción de contactos y la disponibilidad de una mayor cantidad de secuencias y mejores programas de alineamiento. En este sentido cabe destacar el trabajo de Ortíz *et al*., primero demostrando que es posible deducir el plegamiento de un buen número de proteínas conociendo en pequeño conjunto de restricciones (contactos) no cercanos en secuencia [68,69] y, luego, extendiendo esta observación cuando se incluyen contactos predichos, con mayor incertidumbre, mediante coevolución [44,69,70]. En la última década se ha mejorado sustancialmente la predicción de contactos con la introducción de DCA (del inglés *Direct Coupling Analysis*) [71,72], posibilitando la construcción sistemática de buenos modelos estructurales sin la necesidad de utilizar plantillas [73,74]. Aunque para ello se requiere una gran cantidad de secuencias en los alineamientos de partida lo cual limita severamente su ámbito de aplicación actualmente. La gran diferencia entre los métodos “clásicos” y DCA, es que mientras anteriormente se utilizaban directamente correlaciones entre columnas de los alineamientos, en DCA se construye un modelo global de la familia, considerando el posible efecto del resto de columnas en la correlación de un par de columnas dado. Por ello, distinguiremos entre métodos locales y globales.

1.3.3 Predicción de contactos mediante coevolución: Métodos locales

Se pueden definir tres grandes familias en cuanto a métodos locales puramente coevolutivos se refiere: basados en correlación, basados en información mutua y aquellos que utilizan árboles filogenéticos. En todos ellos la información básica es la misma, un alineamiento múltiple de secuencias homólogas a la proteína objetivo.

1.3.3.1 MÉTODOS BASADOS EN CORRELACIONES

Estos métodos se basan en medir la correlación de cambios observados. Por ejemplo, en uno de los primeros y más influyentes métodos [40], que denotaremos por McBASC [26,75], se calcula la correlación entre la magnitud de los cambios ocurridos entre cada par de posiciones a lo largo de todas las secuencias en el alineamiento múltiple de secuencias homólogas. Siguiendo la formulación de Olmea *et al.* [76], la correlación r_{ij} se mide como

$$r_{ij} = \frac{1}{N^2} \sum_{kl} \frac{(s_{ikl} - \langle s_i \rangle)(s_{jkl} - \langle s_j \rangle)}{\sigma_i \sigma_j}$$

donde s_{ikl} se refiere al valor de en la matriz de substitución entre el aminoácido que se encuentra en la columna i de la secuencia k y el que se encuentra en la misma columna de la secuencia l; $\langle s_i \rangle$ es la media de las substituciones en la posición i y σ_i su desviación típica. Obviamente, este valor solo está definido para columnas que no están completamente conservadas, por lo que para este análisis se eliminan todas las posiciones completamente conservadas.

De esta forma, se obtiene un valor de correlación entre cada par de posiciones del alineamiento que permite ordenar las predicciones. En conjuntos de validación, es posible contrastar si las correspondientes posiciones en la estructura tridimensional son contactos físicos o no. La Figura 1.2 muestra un esquema explicativo del proceso. Esta aproximación permite obtener una estimación de la correlación de los patrones de cambio en las posiciones sin asumir unas pautas concretas en los cambios compensatorios, y tiende a favorecer a posiciones relativamente conservadas con respecto a posiciones más variables [75]. Esto es probablemente debido a que, dado que cada cambio es considerado en la correlación, el patrón de covariación tiene que ser muy consistente, algo particularmente difícil en posiciones variables. La tendencia de favorecer posiciones conservadas diferencia a McBASC frente a otras aproximaciones que tienden a favorecer posiciones

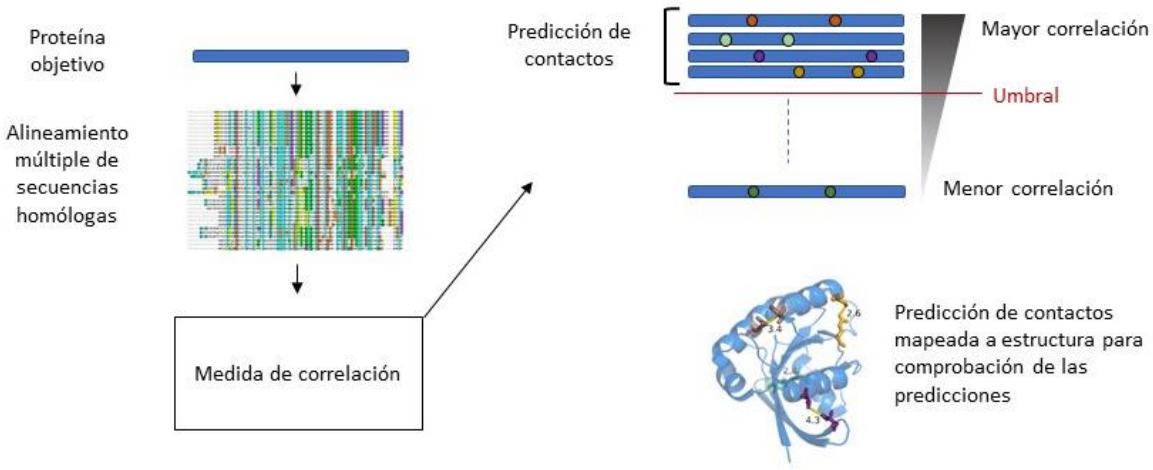


Figura 1.2 Esquema explicativo de una predicción de contactos. Partiendo de la secuencia de aminoácidos de la proteína objetivo, se construye un alineamiento múltiple de secuencias homólogas a la proteína objetivo. A partir de este alineamiento se computa la medida de correlación entre posiciones del alineamiento, ordenándose las predicciones por este valor. Usualmente se utiliza un umbral de la medida de correlación para definir el conjunto de predicciones. Durante el desarrollo de métodos, se pueden realizar estas predicciones sobre conjuntos de proteínas con estructura conocida para poder mapear las predicciones a las estructuras y comprobar si son realmente contactos físicos en ella.

relativamente variables [75], algo esperable de métodos basados en covariación. Desarrollos subsecuentes minimizaron problemas asociados a la calidad del alineamiento [76] y al efecto filogenético [77].

1.3.3.2 MÉTODOS BASADOS EN INFORMACIÓN MUTUA

La segunda familia se basa en la teoría de la información y utiliza la información mutua (MI, del inglés *Mutual Information)* como medida de covariación. La información mutua cuantifica la dependencia mutua entre dos variables aleatorias por medio de la reducción en entropía que se produce en una de las variables al conocer la otra. En el caso de la predicción de contactos, la información mutua se calcula para cada par de columnas en base a las distribuciones de aminoácidos de cada columna y la distribución conjunta de aminoácidos de las dos columnas.

$$MI(i,j) = H(i) + H(j) - H(i,j) = -\sum_{a_i} P(s_i)\log\ P(s_i)) - \sum_{a_j} P(s_j)\log\ P(s_j)) + \sum_{a_{ij}} P(s_i,s_j)\log\ P(s_i,s_j))$$

donde $P(s_i)$ corresponde a la frecuencia de cada posible aminoácido a_i en la posición i del alineamiento, y de forma equivalente para $P(s_j)$. $P(s_i,s_j)$ corresponde a la frecuencia de cada pareja de aminoácidos en las posiciones i y j.

En este caso las variables son categóricas (aminoácidos) sin que cada categoría tenga un significado particular aparte de distinguir los elementos que pertenecen a cada categoría. Es decir, no se tiene en consideración las propiedades fisicoquímicas de cada aminoácido ni las posibles similitudes entre distintos aminoácidos. En esencia, cuantifica la sobrerrepresentación de parejas de aminoácidos con respecto a lo que es esperable dada la variabilidad de las posiciones. Los primeros métodos mostraron la utilidad de la aproximación [78] aunque su capacidad predictiva era algo inferior a los métodos basados en correlaciones [75]. El mejor rendimiento de los métodos basados en correlaciones parece ser debida a su tendencia a dar mayor preponderancia a posiciones más conservadas [75] que tienen una mayor tendencia a estar próximas y deberían estar asociadas a mayores presiones de selección. Esto contrasta con MI que se basa exclusivamente en la variabilidad sin considerar la conservación de las posiciones, favoreciendo posiciones relativamente variables.

1.3.3.3 INFLUENCIA DE LA FILOGENÍA Y LA CONSERVACIÓN DE LAS POSICIONES

Una dificultad muy importante de los métodos basados en covariación para detectar coevolución radica en la distinción entre la covariación debida a las relaciones filogenéticas entre las secuencias que conforman el alineamiento y la covariación debida a las restricciones estructurales, tal y como han discutido numerosos trabajos [46,77,79,80]. Las covariaciones observables se ven ampliamente influenciadas por la filogenia subyacente ya que muchas de las similitudes entre las secuencias provienen de su pasado común [75,81]. Un ejemplo extremo es clarificador al respecto. Si imaginamos dos grupos de secuencias dentro de una familia que son altamente similares dentro de su grupo, pero muy distintas a las del otro grupo, es obvio que las covariaciones que podamos observar se deben principalmente a las relaciones filogenéticas existes entre

ellas y no a un proceso coevolutivo. En general, la relación filogenética entre las secuencias afecta profundamente a las covariaciones observables haciendo muy difícil distinguir cuando las covariaciones se deben a este hecho o un proceso coevolutivo.

Las relaciones filogenéticas, por tanto, suponen una dificultad en este contexto y deben ser entendidas como un factor de confusión. Es necesario aclarar que la información filogenética puede ser útil en otros contextos como la detección de interacciones entre proteínas (como vimos en la sección 1.2) o de residuos funcionalmente importantes [82,83], pero en el contexto de esta tesis representa un factor de confusión.

Muchos métodos, independientemente de la aproximación, tratan de reducir el efecto de la señal filogenética. En uno de los trabajos más exitoso a este respecto [79], se minimiza la señal filogenética mediante la introducción de una corrección llamada APC (del inglés *Average Product Correction*)

$$APC(i,j) = \frac{MI(i,\bar{x})MI(j,\bar{x})}{\overline{MI}}$$

donde $MI(i,\bar{x}) = \frac{1}{m}\sum_{x\neq i} MI(i,x)$, para las m columnas del alineamiento. Y de forma equivalente para $MI(j,\bar{x})$. Mientras que $\overline{MI}$ es la información mutua media de todos los pares de columnas. La medida o *score* de coevolución final, MI_p, se obtienen sustrayendo APC del valor de la información mutua de cada par de columnas, $MI_p(i,j) = MI(i,j) - APC(i,j)$.

Como se deduce de su forma analítica, esta corrección reduce el valor de la covariación observada para pares de posiciones que tienden a covariar con en el resto de las posiciones, lo que es esperable para señales debidas a la filogenia. En concreto, Dunn *et al.* demostraron que permite estimar una distribución de fondo (*background*) que considera la conservación de las posiciones y la señal filogenética para cada par de posiciones cuya corrección mejora sistemáticamente las predicciones [79,84].

En cuanto a los métodos basados en correlaciones, cabe destacar la contribución de Fares *et al.* analizando la robustez de los resultados tras quitar clados del alineamiento [77] y la contribuciones de Ortiz *et al*. apoyándose en técnicas de reducción de dimensionalidad [44,85]. En este último caso, se minimizaba la parte de la señal de covariación relacionada con las componentes principales ya que está bien establecido que éstas capturan la estructura filogenética de la familia [82,86].

Otro factor de confusión importante son los sesgos introducidos por la conservación de las posiciones, ya que afectan a la probabilidad de encontrar al azar covariaciones que no están relacionadas con restricciones estructurales [79]. Es importante mencionar que la conservación de las posiciones y las relaciones filogenéticas no son problemas independientes ya que la filogenia afecta a la conservación observable. Por otro lado, el trabajo de Fodor *et al.* aclaró que cada método tiene tendencias distintas con respecto a la conservación, dando más relevancia a distintos regímenes de conservación [87]. Esto, unido a cuanto es esperable encontrar un contacto dependiendo de la conservación de las posiciones, explica en buena medida las diferencias de rendimiento [75].

1.3.3.4 MÉTODOS QUE UTILIZAN ÁRBOLES FILOGENÉTICOS

La distinción entre señal filogenética y coevolutiva es la principal motivación del tercer grupo de métodos de predicción de contactos, aquellos que utilizan árboles filogenéticos. En estos métodos se construyen árboles a partir de alineamientos múltiples de secuencias homólogas que permiten estimar las secuencias ancestrales y, por tanto, los cambios en las secuencias acaecidos. En el trabajo de Yeang y Haussler [88], generan un modelo nulo (sin coevolución) mediante un proceso de Markov de tiempo continuo donde las transiciones de aminoácidos para cada par de posiciones es proporcional a las frecuencias de pares de amino ácidos observadas en el alineamiento. Desde esta matriz de transiciones, se modifican las probabilidades de las transiciones recorriendo los estados ancestrales y dando mayor peso a cambios que hayan ocurrido simultáneamente. Finalmente, se obtiene el *score* coevolutivo contrastando (*log-odd ratio*) el modelo coevolutivo con respecto al independiente. Este trabajo ejemplifica las principales líneas de los trabajos que utilizan árboles donde, por un lado, se puede reconstruir los estados ancestrales para tratar de encontrar eventos que muestren dependencia entre las posiciones [88,89] (mediante cambios simultáneos) y, por otro, permite contrastar con respecto a un modelo nulo [90,91]. Este tipo de aproximación tiene la ventaja de ser menos susceptible a la señal filogenética ya que se considera explícitamente la relaciones entre las secuencias. Sin embargo, la construcción de árboles es computacionalmente costosa y no está exenta de artefactos. De hecho, se ha mostrado que los resultados de este tipo de aproximación tiende a no ser robusta a la reconstrucción filogenética y el modelo de substituciones asumido [92,93]. Por todo ello, este tipo de aproximación ha recibido menos atención por parte de la comunidad [26].

1.3.3.5 PREDICCIÓN DE CONTACTOS MEDIANTE APRENDIZAJE AUTOMÁTICO

En 2001, Fariselli *et al.* propusieron la utilización de un método de aprendizaje automático que incluía las mutaciones correlacionadas junto con otras variables para el aprendizaje [94]. Desde entonces, numerosas aproximaciones para la predicción de contactos basadas en aprendizaje automático y considerando covariación han sido propuestas incluyendo redes neurales [76,94–96], máquinas de vector de soporte [97,97] y bosques aleatorios [98]. Típicamente incluyen otras fuentes de información y características como la conservación de la posición [76,94,96,97], separación en la secuencia [76], potenciales de contacto [76,97], predicción de estructura secundaria [94,97,99], predicción de accesibilidad [76,97], propiedades fisicoquímicas [96,97], uso de ventanas deslizantes [94,97,99], entre otras posibilidades.

1.3.4 Predicción de contactos mediante coevolución: Métodos globales

1.3.4.1 ANÁLISIS DE ACOPLAMIENTO DIRECTO

Las proteínas se plieguen a su forma nativa gracias a la formación de complejas redes de interacciones entre sus aminoácidos constituyentes [100]. En este tipo de sistemas de alta interactividad es común que las correlaciones observadas entre pares de variables no sean debidas a interacción directas, sino que se deriven de cadenas de correlaciones provenientes de la red de interacciones (Figura 1.3). Estas correlaciones indirectas son bien conocidas en física estadística, en particular en la física de sistemas de espines [101]. Las correlaciones indirectas pueden llegar a ser más fuertes que las directas [102]. Lapedes *et al*. propusieron en 1999 aproximar el problema usando los modelos desarrollados en el contexto de la física estadística, primero en un escenario simulado [103] y, más tarde, con secuencias reales [104]. Algunos años después, aproximaciones similares fueron propuestas mostrando una notable mejora en la predicción de contactos [71,72,105,106].

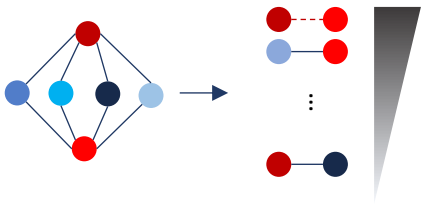


Figura 1.3 Esquema del efecto de las interacciones indirectas en la correlación. Los nodos en rojo no interaccionan directamente, pero pueden tener una correlación observable mayor que las interacciones directas al influenciarse mutuamente por múltiples caminos en la red.

Daremos aquí una breve introducción sobre el análisis de acoplamiento directo (DCA, del inglés *Direct Coupling Analysis*), se pueden encontrar más detalles en la sección 3.4.3 de Materiales y Métodos. En estas aproximaciones se propone modelar la distribución de secuencias en una familia de proteínas con un modelo estadístico $P(A_1, \dots, A_L)$, donde A_i es el aminoácido en la posición i y L es el número de columnas en el alineamiento de la familia de proteínas en cuestión. Para que el modelo sea acorde con los datos, sus marginales han de cumplir las frecuencias de aminoácidos de cada columna y las frecuencias de pares de aminoácidos de cada par de columnas. Existen una infinidad de modelos posibles bajo estas restricciones. De los posibles modelos, podemos tomar el más general, esto es el menos sesgado o de máxima entropía [107,108], deduciéndose

$$P(A_1, \dots, A_L) = \frac{1}{Z} \exp\left\{ \sum_i h_i(A_i) + \sum_{i<j} e_{ij}(A_i, A_j) \right\}$$

donde el factor de normalización Z es también conocido como función de partición. Dado el número habitual de posiciones en los alineamientos no es posible realizar el computo de forma exacta, por lo que es necesario recurrir a inferencias aproximadas. Existen diversos esquemas aproximados con distintas asunciones y costes computacionales que pueden ser usados incluyendo *mean field* [72], *message passing* [71], *pseudolikelihood maximization* [109], regresión logística multinomial [110], *adaptative cluster expansion* [111] y máquinas de Boltzmann [112,113], entre otros [114]. Otras aproximaciones no directamente motivadas desde el principio de máxima entropía pero similares con respecto a la modelización probabilística incluye métodos basados

redes bayesianas [105] y la inversión de la matriz de covarianza [115]. Durante esta tesis hemos optado principalmente por plmDCA asimétrico [110] (Materiales y Métodos, sección 3.4.3), basado en la regresión logística multinomial, por su buen balance entre la precisión de las predicciones de contactos y el tiempo de computo.

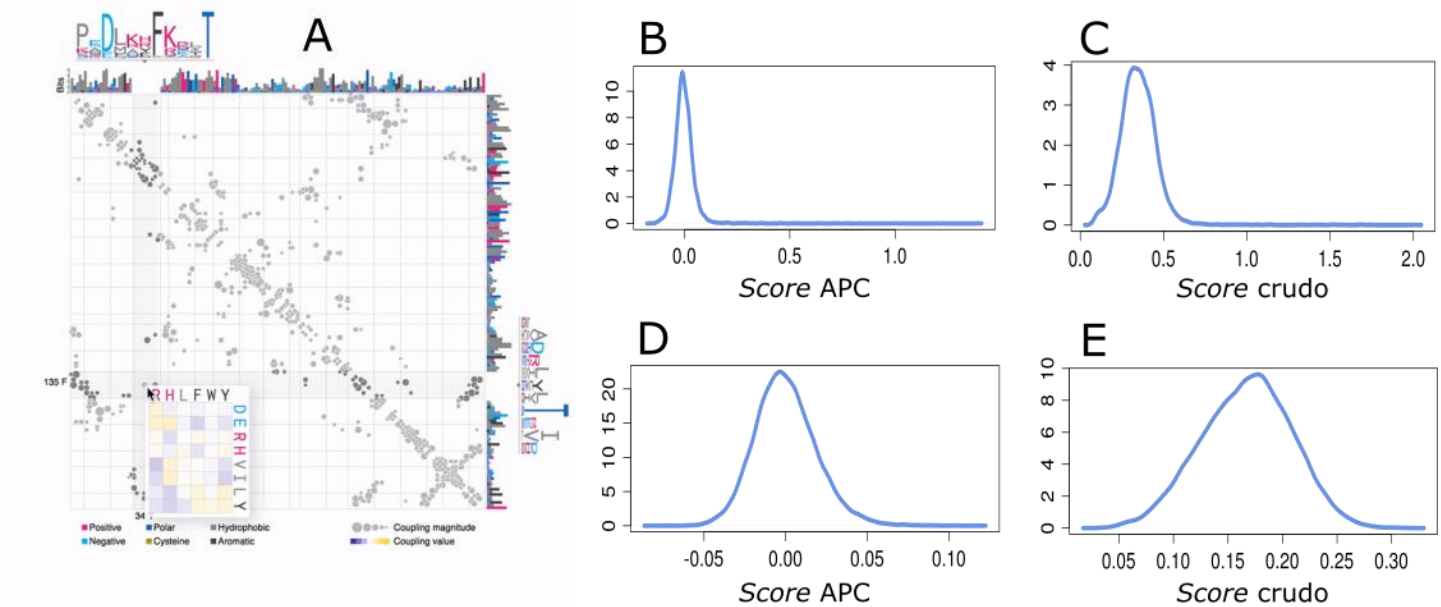


Figura 1.4 Parámetros de acoplamientos y *scores*. A) Matriz de parámetros de acoplamientos e_{ij}. En gris se muestran los pares de posiciones cuyos parámetros de acoplamiento tienen una magnitud relevante, donde el tamaño del punto es proporcional a la magnitud. En los marginales se puede observar el histograma de conservación de las posiciones y los logos HMM de conservación para las posiciones cercanas al par de posiciones seleccionadas. Se muestran en distintos colores los diversos tipos de aminoácidos. En la zona ampliada, se muestran los parámetros de acoplamiento para el par de posiciones seleccionada, el gradiente azul y amarillo muestra la magnitud de cada parámetro de acoplamiento. Nótese que solo se muestran el subconjunto de parámetros que tienen una magnitud no cercana a 0, ya que la mayor parte de los 400 pares de combinaciones de aminoácidos tienen un valor muy cercano a 0. B y C) Distribución de *scores* para todos los pares de posiciones interdominio para un caso con señales coevolutivas muy fuertes y muy buenas predicciones de contacto (dominios Pfam HisKA y Response_reg), el panel B muestra la distribución del *score* APC y el panel C el del *score* crudos. D y E) Distribución de scores para todos los pares de posiciones interdominio para un caso donde no se observan señales coevolutivas fuertes y las predicciones de contacto son malas (dominios Pfam DNA_pol3_delta y DNA_pol3_delta2), el panel B muestra la distribución del *score* APC y el panel C el del *score* crudo. Los valores atípicos (*outliers*) en la cola derecha de estas distribuciones suelen estar asociados a señales fuertes de coevolución y predicciones de contacto correctas, como muestra las diferencias entre los dos casos. Fuentes: La imagen del panel A fue extraída de EVZoom.org.

El número de parámetros del modelo, del orden de q^2L^2 es en la practica muy superior al número de secuencias en los alineamientos de entrada. Para una proteína pequeña de 100 aminoácidos, el número de parámetros es del orden de 10^6, mientras que el número de secuencias rara vez excede del orden de 10^5 y es normalmente muy inferior. Además es necesario eliminar o repesar secuencias similares (p. ej. más de un 70% o un 90% de identidad en secuencia) para reducir el efecto filogenético [71,72]. Utilizaremos el término "secuencias efectivas" para referirnos al número de secuencias no redundantes al 80% en un alineamiento. Dado que el número de parámetros excede los datos disponibles, es necesario realizar una regularización para evitar sobreajuste (*overfitting*) [114], utilizándose comúnmente la norma L_2 [110].

Los parámetros de acoplamiento (*coupling*) e_{ij} del modelo capturan la dependencia estadística entre pares de aminoácidos para cada par de posiciones minimizando la influencia de otras posiciones (Figura 1.4A). Para conseguir una predicción de contacto es necesario combinar los parámetros de los pares de aminoácidos,

obteniendo una medida o *score* entre pares de posiciones. Para ello, actualmente se utiliza principalmente la norma de Frobenius (Materiales y Métodos, sección 3.4.3, [109]). Llamaremos *score* crudo al *score* obtenido aplicando la norma de Frobenius sobre los parámetros e_{ij}. Sobre esta norma de Frobenius se aplica APC (Materiales y Métodos, sección 3.4.3, [109,115]), algo que mejora sistemáticamente las predicciones [109]. Al igual que el contexto que con información mutua, esta mejora se debe a que se minimiza el *score* de pares de posiciones que muestran más correlación con un elevado número de posiciones, atribuible a la señal filogenética, así como a sesgos provenientes de la conservación de las posiciones [115,116]. La predicción final es equivalente en formato a la de los métodos locales, los pares de posiciones se ordenan mediante el *score* obtenido para dar lugar a un ranking de predicciones de contacto (Figura 1.2). Nos referiremos al *score* antes de aplicar APC, extraído directamente desde el modelo DCA, como *score* crudo (ver distribución de *scores* crudos en Figura 1.4C y E). Después de aplicar APC, como *score* APC (ver distribución de *scores* APC en Figura 1.4B y D). Es posible estandarizar el *score APC* y obtener un *z-score* mediante la desviación mediana absoluta (*Median Absolute Deviation*, sección 3.4.3 de Materiales y Métodos) [117], al que nos referiremos como *z-score* APC.

En la práctica totalidad de estudios se establece un umbral sobre los *scores* para obtener predicciones suficientemente fiables. Por debajo del umbral, se considera que la señal no es lo suficientemente fuerte y se evita una mayor proporción de falsos positivos. Para referirnos a predicciones asociadas a pares de posiciones con un *score* coevolutivo por encima del umbral, diremos que ese par de posiciones *han coevolucionado.* Utilizaremos esta expresión en el texto por simplicidad, implicando que detectamos una fuerte covariación específica entre ese par de posiciones que probablemente esté relacionada con un fenómeno coevolutivo.

Cabe destacar que, a pesar de que APC es usado actualmente en la práctica totalidad de las metodologías DCA, no se comprende con exactitud el motivo de su éxito en el contexto de DCA [116]. En el caso de información mutua, APC proporciona un *background* para cada par de posiciones que considera la conservación de las posiciones y la señal filogenética. Tras sustraer este *background* de los *scores* se mejoran sistemáticamente las predicciones, tanto si se usa información mutua [79] como si se usa DCA [109,115]. Según Vorberg *et al*., "es ampliamente reconocido en el campo que nuestro limitado entendimiento sobre qué ruido está corrigiendo APC y por qué es una corrección tan efectiva está impidiendo de forma importante el progreso de desarrollo de métodos estadísticos mejorados para la predicción de pares de residuo en contacto" [116]. En concreto, recalcan que el éxito de APC puede deberse más a su capacidad para corregir los sesgos debidos a la conservación que a relaciones filogenéticas. Aunque es importante recordar que no se trata de dos problemas independientes.

En los últimos años se han propuesto muchas mejoras ulteriores. Una línea particularmente prolífica está basada en el uso del aprendizaje automático, en particular con métodos de aprendizaje profundo (d*eep learning*) ya sea mediante el aprendizaje de la asociación entre covariación y otras variables con los contactos físicos o combinando los *scores* obtenidos por DCA con otra fuentes de información [118–123]. En ambos

casos son necesarios amplios conjuntos de entrenamiento no redundantes. Es destacable el hecho de que solo utilizando datos de covariación con aprendizaje profundo se obtengan una calidad en las predicciones en línea con otros métodos que incluyen otras informaciones [124], subrayando el papel crucial de la covariación en estas aproximaciones. El aprendizaje profundo permite capturar patrones complejos en los datos más allá de la covariación. Un ejemplo de ello es que permite capturar patrones típicos en los mapas de contacto. Señales relativamente fuertes pero aisladas en el mapa de contactos pueden ser filtradas mientras que señales del mismo tipo pero coherentes con otras señales en posiciones vecinas se pueden ver reforzadas [119]. Recientemente, métodos basados en aprendizaje profundo han mostrado que es posible predecir distancias entre posiciones [125–128], no solo contactos. Aparte del aprendizaje profundo, otros trabajos han mostrado que resulta beneficioso incluir consideraciones geométricas de forma que los contactos sean compatibles con ellas [129], contrastar la covariación a nivel de nucleótidos con respecto a la de aminoácidos [130] o combinar las predicciones de contacto con datos experimentales de resonancia magnética nuclear [121] o reticulación química (*chemical crosslinking*) acoplada a espectrometría de masas [131].

1.3.4.2 APLICACIONES ACTUALES DE LA PREDICCIÓN DE CONTACTOS

La introducción de DCA ha supuesto una mejora notable en cuanto a la predicción de contactos [71,72]. Esto ha posibilitado la construcción de buenos modelos estructurales sin la necesidad de plantillas estructurales [73,74], en particular para proteínas de membrana que son difíciles de cristalizar [74]. La predicción de contactos mediante coevolución ha seguido jugando un papel fundamental en el progreso de la predicción de estructura en los últimos años, tal y como atestiguan las últimas ediciones de CASP [62,63]. Mención especial merece la última edición de CASP celebrada en 2018, donde el método AlphaFold logró un notorio avance en la categoría de modelado libre sin plantillas [132]. Al tiempo de escribir esta tesis, parece que este avance se debe principalmente al uso de predicción de distancias, basada en coevolución, junto al hecho de que estas distancias constituyen un espacio diferenciable apto para minimización por descenso de gradiente así como al uso de sofisticadas redes neuronales en buena parte de las etapas del protocolo [132].

Las mejora de las predicciones de contacto también han sido útiles en el estudio de conformaciones alternativas [133,134], estados de oligomerización [72,74,135,136], dinámica molecular y/o proceso de plegamiento [112,137,138], distinguir entre plegamientos nativos e incorrectos [139], proteínas desordenadas [140], estructura de RNAs [141–143] e interacciones entre proteínas [117,122,144,145].

Dado que en DCA se construye un modelo estadístico de la familia de proteínas, sus aplicaciones van más allá de la predicción de contactos. Ya que podemos asociar una probabilidad a cada secuencia, aunque concretamente se utiliza la energía estadística del Hamiltoniano que es proporcional a la probabilidad, podemos medir el efecto de las mutaciones comparando la energía entre la secuencia objetivo y la misma secuencia con la mutación a medir. Estas medidas computacionales correlacionan bien con experimentos donde se miden el *fitness* de miles de combinaciones de mutaciones [146–148] y pueden mejorar la

predicción del efecto de la mutaciones [146–148]. También se ha mostrado que estos modelos son capaces de reproducir frecuencias de amino ácidos de orden superior a dos a pesar de incluir solo parámetros de primer y segundo orden [149,150], así como reproducir la estructura de subfamilias de la familia en cuestión [112,149].

La mayor limitación de DCA es la ingente cantidad de datos en secuencia necesarios. La calidad de las predicciones crece con el número de secuencias en los alineamientos [151]. Como cantidad orientativa, se estima que se necesitan del orden miles de secuencias no redundantes en los alineamientos [137] o, más precisamente, al menos 5 secuencias no redundantes por cada posición del alineamiento [151]. Aunque para determinadas aplicaciones 500 secuencias no redundantes pueden suficientes [117,135,152]. Esta necesidad limita enormemente el ámbito de aplicación de DCA, ya que relativamente pocas familias de proteínas son lo suficientemente grandes. No obstante, se han realizado trabajos de modelización a larga escala importantes para proteínas de membrana [74], difíciles de cristalizar, o utilizando grandes base de datos que incluyen secuenciación de metagenomas [153]. Pese al crecimiento de bases de datos de secuencias, este sigue siendo el mayor obstáculo para ampliar el ámbito de aplicación de este tipo de aproximaciones.

1.3.5 Predicción estructural de complejos de proteínas

Al igual que ocurre con la estructura terciaria de las proteínas, la estructura cuaternaria de las proteínas está también altamente conservada [154]. Por ello, es posible realizar buenos modelos tridimensionales utilizando complejos homólogos como plantillas [155]. De forma similar a las proteínas, los complejos homólogos con identidades de secuencia superiores al 30-40% interactúan en la mayor parte de los casos de forma muy similar [154]. Sin embargo, a diferencia de que ocurre con la estructura terciaria, esto no suele ser así por debajo de ese umbral (identidad en secuencia < 30%) [154]. Una forma de minimizar este hecho es tratar de buscar distintos complejos homólogos divergentes y, si la región de interfaz está estructuralmente conservada, entonces la probabilidad de que nuestro complejo objetivo tenga una interfaz similar es mucho mayor [156]. A esto se suma el hecho de que nuestro desconociendo de la estructura de los complejos es mucho mayor que para proteínas en proporción al número estimado de interacciones que se producen [157]. Por ejemplo, para humanos se estima que se conoce la estructura, o se pueden realizar modelos fiables, para entre ~6% [122] y el 11.7% (según las estadísticas de la versión 2019_01 de interactome3D [155]) de las todas las interacciones.

Cuando las estructuras de las proteínas que forman el complejo son conocidas, es posible realizar modelos del complejo mediante métodos de acoplamiento proteína a proteína (*protein-protein docking*). Estos métodos se componen generalmente de dos pasos: generación de poses y puntación de poses. En el primer paso, se generan una gran cantidad de posibles asociaciones entre las dos proteínas normalmente tratándolas como cuerpos rígidos y, típicamente, alguna función de puntuación sencilla para descartar poses muy desfavorables. En términos generales, se encuentra una pose cercana a la real en un 80% de los casos habiendo generado 10000 poses [158]. En el segundo paso se puntúa cada una de las posibles soluciones y

se ordenan de acuerdo con esta métrica. Las funciones de puntuación generalmente tienen en consideración la complementariedad química o geométrica, interacciones de enlace de hidrógeno o Van der Waals y potenciales de interacción empíricos. Es difícil obtener una función que sea capaz de identificar estructuras nativas, por lo que existen numerosos algoritmos y funciones de puntuación [159,160].

Desgraciadamente, las aproximaciones basadas en cuerpos rígidos tienen series dificultades con respecto a los cambios conformacionales que ocurren en la interfaz durante la unión, algo muy común en los complejos de proteínas [161]. Otra fuente de dificultad es el uso de modelos en vez de estructuras para las proteínas individuales reduciendo la tasa éxito conforme se utilizan plantillas más divergentes [162].

Por otro lado, los métodos de acoplamiento proteína a proteína (*protein-protein docking*) se pueden combinar con otras fuentes de información para mejorar la tasa de éxito en el modelado de la interacción. En particular, se puede aumentar notablemente la tasa de éxito combinándolos con informaciones experimentales como mutagénesis, restricciones de interfaz por NMR (del inglés *Nuclear Magnetic Resonance*), microscopía, entre otros [163,164]. Otra fuente de información particularmente interesante con la que se pueden combinar son los contactos predichos por los métodos de predicción de contactos entre proteínas, que introduciremos a continuación.

1.3.6 Predicción de contactos entre proteínas

La predicción de contactos entre proteínas es esencialmente idéntica a la predicción de contactos para proteínas, recibiendo de entrada un alineamiento, que en este caso contiene las familias de las dos proteínas cuya interacción se quiere predecir, y devolviendo los *scores* coevolutivos para cada posible combinación de posiciones de las dos proteínas. La diferencia es la necesidad de generar un alineamiento emparejado donde cada secuencia perteneciente a la primera familia esté emparejada con una secuencia de la segunda familia (Figura 1.5). Para ello se recuperan primero secuencias homólogas alineadas para cada proteína (o dominio) por separado. A continuación, se procede a intentar emparejar aquellos pares de secuencias que realmente interactúen físicamente, ya que en caso contrario estaremos introduciendo ruido.

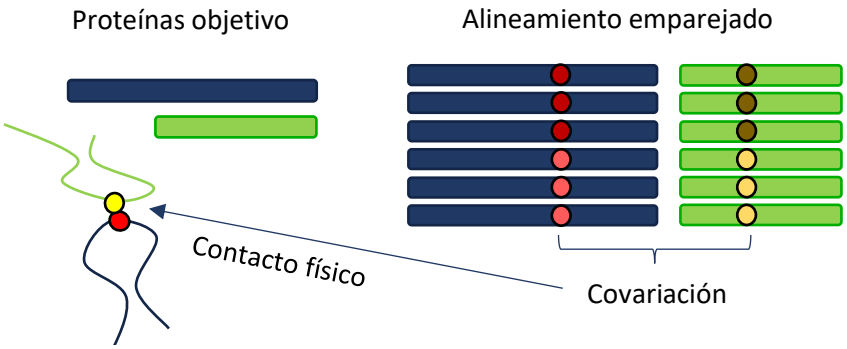


Figura 1.5 Esquema sobre alineamientos emparejados. En el caso de las predicciones de contactos entre proteínas es imprescindible construir alineamientos emparejados de tal forma que las secuencias de la primera familia estén emparejadas con secuencias de la segunda familia con las que interaccionan físicamente, o no se podrá recuperar adecuadamente las covariaciones entre las dos familias.

Una parte clave para la predicción de contactos interproteína es encontrar una forma lo suficientemente fiable de encontrar qué pares de proteínas interaccionan para realizar estos emparejamientos. Una posibilidad es utilizar, en procariotas, la adyacencia genómica ya que las proteínas que interacción físicamente en procariotas suelen estar cercanas en el genoma formando parte del mismo operón [144,145]. Partiendo de una pareja de proteínas cuya interacción es conocida (o existe evidencia de ello), se pueden utilizar relaciones de ortología uno a uno para recuperar pares de proteínas emparejadas [32,34,165], aunque este tipo de procedimientos tiende a generar alineamientos con pocas secuencias. Sin embargo, un estudio reciente muestra que es factible en bacterias [166] gracias a la gran cantidad de genomas secuenciados en bacterias. El uso de criterios menos estrictos de ortología permite extender el ámbito de aplicación pero a costa de obtener predicciones menos fiables [167,168]. Otra posibilidad es maximizar la correlación entre los correspondientes árboles de las dos familias [169,170] o maximizar la covariación ya sea medida mediante DCA [171–173] o mediante información mutua (MI) [174]. A diferencia de la predicción de contactos, MI es competitiva con DCA en este contexto [174], posiblemente debido a que la señal filogenética no es un obstáculo en este ámbito, ya que los emparejamientos deben ser coherentes con los árboles de las dos familias. Las aproximaciones basadas en la maximización tienen la ventaja de conseguir más emparejamientos, algo importante si luego se pretende utilizar DCA para realizar la predicción de contactos final. Por otro lado, introducen una mayor incertidumbre en los emparejamientos en comparación con adyacencia genómica, por lo que sigue siendo un campo de investigación activo [171–174].

Aunque la predicción de contactos entre proteínas basada en coevolución ha recibido mucha menos atención que la predicción dentro de proteínas, hay varios trabajos sobre ello. Ya en 1997, Pazos y Valencias mostraron que era posible realizar predicciones de contacto entre proteínas y que éstas contenían información sobre la interacción al ser capaces de distinguir el complejo real entre una serie de señuelos (*decoys*) [175]. A pesar de ser potencialmente útiles, los métodos locales presentaban limitaciones importantes para ser aplicados en la práctica, como mostró un estudio posterior donde se analizaban una serie de métodos locales y su combinación [176]. Posteriormente se demostró que la utilización de DCA favorecía claramente la calidad de las predicciones, aunque los métodos locales podían ser útiles en determinadas condiciones [152]. El uso la adyacencia genómica para realizar los emparejamientos en procariotas permitió, finalmente, obtener predicciones de contactos entre proteínas muy fiables [144,145] al permitir obtener alineamientos con muchas secuencias y emparejamientos fiables. Otros estudios han mostrado, en sistemas concretos, el potencial de la combinación de la predicción de contacto intermolecular con datos experimentales de microscopía electrónica [177] y la espectrometría de masas [178].

Varios estudios han propuesto el uso de aleatorizaciones de los alineamientos para mejorar la relación señal/ruido en el contexto de interacciones entre proteínas [179,180]. En estas aproximaciones, los aminoácidos de cada columna se reordenan aleatoriamente mientras se deja intacto el resto de alineamiento, repitiendo el proceso un elevado número de veces. Esto permite obtener un transformar los *scores* de coevolución en un *p-valor* empírico a partir de las aleatorizaciones, mejorando notablemente las

predicciones de los métodos locales [180]. Pero no ha sido aplicado a métodos globales dado el elevado coste computacional (p. ej. 2 000 aleatorizaciones para cada columna) [180]. En la segunda parte de esta tesis, exploramos aleatorizaciones menos costosas computacionalmente.

Recientemente se ha aplicado un método de aprendizaje profundo entrenado para intraproteína para predecir contactos interproteína funcionando mejor que otros modelos de tipo DCA [168]. Esto es interesante, ya que la limitada cantidad de datos estructurales sobre interacciones impide realizar aprendizajes profundos directamente. Sugiere, también, que los patrones en intraproteína e interproteína son lo suficientemente similares. Sin embargo, la calidad de las predicciones es relativamente baja y, además, contrasta con un estudio más reciente donde este tipo de aproximación se comportó peor que un modelo de DCA [166], por lo que aconsejable tomar el asunto con la debida cautela.

Utilizando los emparejamientos por adyacencia genómica y combinando la predicción de contacto mediante DCA junto con el acoplamiento de proteínas (*protein-protein docking)*, se ha logrado realizar modelos precisos de complejos de forma sistemática en procariotas [144,145,166]. Cabe destacar que estos estudios muestran que con restricciones espaciales fiables, aunque sean muy pocas, se pueden producir buenos modelos en la gran mayoría de casos [144,145]. Pese a este éxito, hay que mencionar que el ámbito de aplicación es bastante reducido, ya que son necesarias gran cantidad de secuencias en el alineamiento y solo en procariotas. Aunque también es cierto que el crecimiento del número de genomas está permitiendo la reconstrucción sistemática y bastante amplia de complejos en procariotas, tal y como muestra un estudio muy reciente [166].

El problema de necesitar un gran número de secuencias es aún más grave para la predicción de contactos interproteína (en comparación con intraproteína) ya que el proceso de emparejamiento reduce el número de secuencias en los alineamientos emparejados. Y este problema es aún mucho más grave en eucariotas ya que si nos fijamos en el número de genomas no redundantes (según el criterio utilizado en UniProt [181]), solo hay 1 520 en eucariotas por 26 691 en bacterias (aun cuando los criterios de UniProt reducen mucho más la redundancia en bacterias que en eucariotas; datos obtenidos en julio 2019 mediante el portal www.UniProt.org). Aún peor, los criterios de adyacencia genómicas no son aplicables a eucariotas y el gran número de duplicaciones en los genomas de eucariotas, en comparación con procariotas, hace que sea extremadamente difícil construir alineamientos emparejamiento fiables de forma sistemática. En particular, cuando existan un elevado número de duplicaciones. En este escenario, experimentalmente no es posible, o es extremadamente difícil, establecer sistemáticamente que pares de proteínas de las dos familias interaccionan. Además, cuando las paralogías son abundantes, las relaciones entre las familias son complejas y es muy habitual que existan relaciones uno a muchos o muchos a muchos, en vez de uno a uno. Esto es algo difícil de capturar en un alineamiento por lo que se tiende a diseñar estrategias que minimicen el problema evaluando el resultado tras manipular las combinaciones de secuencias. Las recientes los avances en el proceso de emparejamiento [171,172,174] pueden ser de utilidad a este respecto, aunque éste es un problema que no tiene visos de resolverse en un futuro próximo.

Por otro lado, es más que razonable pensar que varios factores afectan en diverso grado a la coevolución y su detección en distintos complejos de proteínas. Prueba de ello es la introducción de correcciones *ad hoc* a los *scores* coevolutivos considerando el número de secuencias y posiciones [144,145]. Dos problemas centrales y relacionados en este contexto es el uso del mismo umbral para todos los casos [135,144,145] y la falta de un tratamiento de la conservación de las posiciones y de las relaciones filogenéticas entre las secuencias específica de cada caso. Factores como el número de secuencias, posiciones, conservación y filogenia afectan de forma diversa a cada familia de proteínas e interacciones, por lo que es esperable que un mismo umbral para todos no sea la estrategia óptima. Una asunción razonable es que la aleatorización de los alineamientos emparejados pueda ser beneficioso para corregir, al menos parcialmente, estos factores.

Trataremos en esta tesis tanto el problema de la aplicabilidad de los métodos a secuencias eucariotas como la necesidad de un umbral o una corrección específica para cada caso. Respecto al primer problema, analizamos la relación entre coevolución y conservación estructural de las superficies de interacción para establecer un procedimiento que permita la proyección de las predicciones de contactos a secuencias remotas, aplicándolo a la predicción de contactos en secuencias de eucariotas basados en alineamientos de secuencias homólogas pero distantes en procariotas. Con respecto al problema del uso de un umbral único para todos los casos, íntimamente relacionado con las características de cada caso incluyendo sus características de conservación y filogenéticas, utilizamos aleatorizaciones de los alineamientos emparejados para obtener un umbral específico de cada caso. Este umbral específico, adicionalmente, nos permite rescatar predicciones correctas provenientes de alineamiento con pocas secuencias, indetectables para los métodos actuales.

2. OBJETIVOS

Los principales objetivos de esta tesis son el estudio de la coevolución en regiones de interacción entre proteínas, investigando la relación entre coevolución y conservación estructural, y la mejora de los métodos de predicción de contactos entre proteínas con especial énfasis en la extensión del ámbito de aplicación a un mayor número de casos. Los objetivos concretos son:

1. Desarrollar una nueva metodología computacional para mejorar la detección de señales de coevolución entre proteínas y entre dominios de proteínas a partir de alineamientos emparejados en procariotas.

2. Investigar si la coevolución está asociada a una mayor conservación estructural mediante la comparación de la intensidad de la señal coevolutiva y la conservación estructural en interfaces. Analizar las implicaciones para la predicción de contactos entre proteínas en eucariotas.

3. Analizar los factores más limitantes de los métodos de predicción de contactos entre proteínas basados en coevolución.

4. Desarrollar una nueva metodología para mejorar las predicciones de contactos entre proteínas en casos con alineamientos emparejados poco poblados.

5. Evaluar la mejora de la capacidad predictiva de la nueva metodología respecto a metodologías previas.

3. MATERIALES Y MÉTODOS

3.1 Recopilación de datos

3.1.1 CONJUNTO DE CASOS

La base de datos 3did es una colección de pares de dominios Pfam para los cuales existe información estructural sobre su interacción física [182–185], obtenida mediante el escaneo de dominios Pfam (usando los perfiles HMM) sobre PDB [186]. Esta base de datos es ampliamente utilizada para estudios que relacionan secuencias y estructura [187–191] y el uso de dominios Pfam permite recuperar homólogos muy distantes en secuencia, algo fundamental para el estudio de divergencias entre procariotas y eucariotas. De 3did (versión 06_2014) obtuvimos una lista de 8 651 pares de dominios Pfam [192] que interaccionan en alguna estructura de PDB. Tras filtrar homodímeros, esta lista se reduce a 4 556 pares de dominios Pfam heterodímericos. Para cada elemento de esta lista, extrajimos de 3did los identificadores de estructuras PDB que contienen el par de dominios en interacción, que fueron utilizados durante el proceso de obtención de los datos estructurales (sección 3.1.2).

Para estos 4 556 pares de dominios Pfam, ejecutamos el protocolo para la detección de señales coevolutivas descrito en la sección 3.4, obteniendo un alineamiento emparejado para cada caso. Clasificamos estos casos en intraproteína o interproteína (Figura 4.1), dependiendo de si la mayor parte de parejas de dominios Pfam del alineamiento emparejado se encontraban en la misma proteína o en distintas (Figura 4.1A Y B), respectivamente.

Para el estudio llevado a cabo en la primera parte de la tesis (sección 4.1), descartamos los casos cuyos alineamientos emparejados tienen menos 500 secuencias [135] (tras quitar secuencias redundantes al 80% de identidad en secuencia para reducir el efecto filogenético [72], ver sección 3.4.2), obteniendo 559 de los 4 556 pares de dominios Pfam. La Figura A1 del Anexo I muestra el histograma de secuencias en los alineamientos emparejados para estos casos. Este es el conjunto de casos que fue utilizado en el primer estudio descrito en la sección 4.1. De estos 559 casos, 401 son intraproteína y 158 interproteína (Figura 4.1C). Posteriormente, clasificamos cada caso en función de si tenían estructuras en especies procariotas, eucariotas o en ambas (sección 3.3). A continuación, se especifican el número de casos dependiendo de las estructuras disponibles y el tipo de interacción entre dominios (intraproteína o interproteína):

- Intraproteína:
    - Casos donde encontramos estructura en procariotas, pero no en eucariotas: 239.
    - Casos donde encontramos estructura en eucariotas, pero no en procariotas: 10.
    - Caso donde encontramos estructura tanto en procariotas como en eucariotas: 152.

- Interproteína:
    - Casos donde encontramos estructura en procariotas, pero no en eucariotas: 106.
    - Casos donde encontramos estructura en eucariotas, pero no en procariotas: 9.
    - Caso donde encontramos estructura tanto en procariotas como en eucariotas: 43.

En el segundo estudio (sección 4.2), rebajamos el número de secuencias mínimo en los alineamientos emparejados de 500 a 40 (inclusive), para poder estudiar casos con alineamientos con pocas secuencias. Para reducir el coste computacional y los requerimientos de memoria filtramos todos los casos con más de 600 posiciones. Por otro lado, nos enfocamos solo en interacciones interproteína, dado que el objetivo principal de este estudio eran interacciones entre proteínas y el coste computacional de 1000 aleatorizaciones es elevado. Filtramos todos los casos que no tuvieran estructura en especies procariotas, para evitar potenciales problemas derivados de alineamientos de baja calidad en el mapeo de las posiciones a las estructuras PDB. Finalmente, obtuvimos con estos criterios 490 pares de dominios Pfam que conforman el conjunto de casos utilizado en el segundo estudio (Figura 3.1).

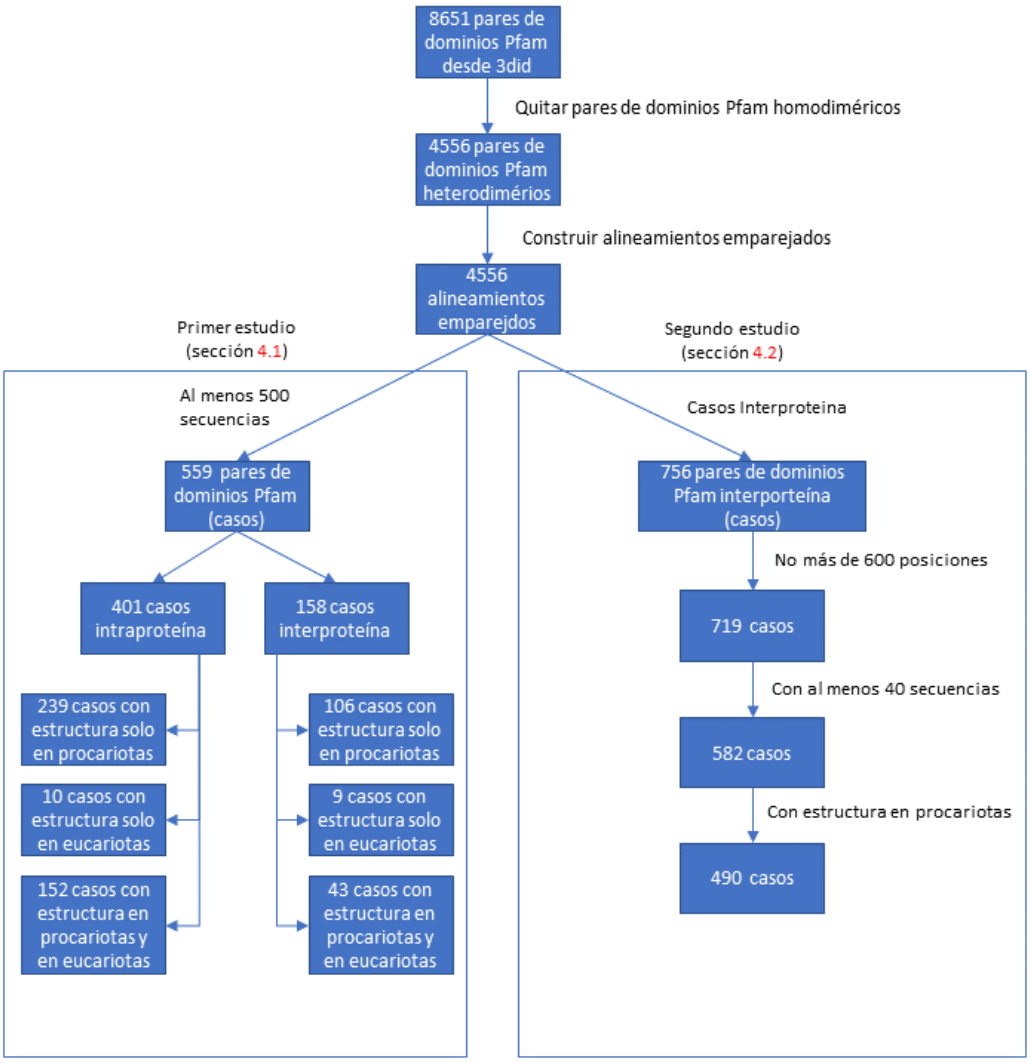


Figura 3.1 Esquema de la obtención de los dos conjuntos de casos utilizados en los dos estudios realizados.

3.1.2 OBTENCIÓN DE DATOS ESTRUCTURALES

Partiendo del conjunto de identificadores de PDB obtenidos en la sección anterior, descargamos los ficheros de las unidades biológicas en formato PDB a través del repositorio http://www.rcsb.org/. Para aquellas estructuras con más de una unidad biológica, descargamos la anotada como primera unidad biológica (que es la seleccionada por los autores del experimento). En caso de no existir unidad biológica, descargamos la unidad asimétrica. Para estructuras de NMR, solo consideramos el primer modelo. A partir de estos ficheros, extrajimos las secuencias de las cadenas implicadas en la interacción (conteniendo el dominio Pfam en cuestión). Los aminoácidos análogos a alguno de los 20 aminoácidos estándar (como la selenocisteína), fueron traducidos a su análogo estándar. Alineamos estas secuencias a su correspondiente dominio Pfam mediante el programa HMMER [193] (versión 3; programa hmmalign, con parámetro "--trim") usando los perfiles HMM de Pfam, obteniendo un mapeo entre la numeración de las posiciones en el PDB y en el dominio Pfam. En este mapeo se incluyó las posiciones que alinean con estados de *match* (estados no asociados a posiciones donde los huecos no son habituales) del perfil de Pfam, esto es, que no son inserciones con respecto al perfil HMM. Descartamos aquellas estructuras donde al menos uno de los dos dominios no aparece en la unidad biológica. Los detalles sobre la definición de interfaces y contactos se encuentran en la sección 3.2.

3.1.3 GENOMAS

Se obtuvieron 15 271 genomas procariotas secuenciados completamente de Ensembl bacteria (versión 23) [194]. Para cada uno ellos descargamos tanto los ficheros en formato FASTA conteniendo las secuencias de aminoácidos de los genes de todo el genoma como los ficheros GTF con las anotaciones genómicas (p. ej. codón de inicio y final de cada gen).

3.2 Definición de interfaces y contactos

Para todos los pares de dominios Pfam y sus correspondientes cadenas en las estructuras, extrajimos la distancia entre cada par de aminoácidos posibles de ambos dominios. Esta distancia fue obtenida considerando todas las posibles combinaciones de átomos pesados (descartando el hidrógeno y el deuterio) de ambos aminoácidos y guardando la distancia más pequeña. En caso de encontrar el mismo dominio Pfam en varias cadenas de una estructura, computamos las distancias para cada par de aminoácidos de todas las posibles combinaciones de cadenas que implicaban a ambos dominios Pfam, quedándonos con la distancia más pequeña. Anotamos todas las posiciones en regiones desordenas (no resueltas) o no alineadas, para las cuales no es posible obtener una distancia. Todos los análisis concernientes a los contactos físicos e interfaces no consideran estas posiciones.

Para definir los contactos que se producen en una interfaz, tomamos una distancia de 8 angstroms como umbral de referencia para considerar que un par de residuos (cada uno perteneciente a un dominio distinto) se encuentran en contacto, siguiendo el estándar *de facto* en el campo [144,145]. Como definiciones alternativas, usamos un umbral más restrictivo de 5 angstroms y uno más laxo de 12 angstroms. Gracias a los mapeos entre los dominios Pfam y las estructuras, anotamos las posiciones en los dominios Pfam en contacto en las interfaces.

A partir de esta información colapsamos la información de distintas estructuras PDB sobre una misma pareja de dominios Pfam con dos estrategias. En la primera estrategia, consideramos los contactos obtenidos en todas las estructuras en especies procariotas, por un lado, y en todas las estructuras en especies procariotas, por otro. Esta definición incorpora información de distintos escenarios biológicos, como la variabilidad debida a la flexibilidad, cambios conformacionales o divergencias dentro del mismo dominio, así como metodológicos (distintas técnicas de resolución de la estructura, condiciones de cristalización diversas, entre otros). Nos referimos a las interfaces obtenidas con esta estrategia como *interfaces completas*. En la segunda estrategia, obtuvimos la calidad de los alineamientos de todas las cadenas de las estructuras mediante el programa HMMER y los perfiles HMM de Pfam. Seleccionamos como estructura representativa aquella con un mayor *bitscore* (puntuación). Usando estos alineamientos colapsamos la información de las posiciones de contactos para todas las estructuras con, al menos, un 98% de identidad en secuencia con la estructura representativa. En este procedimiento también se consideraron por separado las estructuras en especies procariotas de aquellas en especies eucariotas. Nos referimos a las interfaces obtenidas con esta estrategia como *interfaces representativas*. Por tanto, la interfaz representativa es la colección de contactos encontrados en la estructura representativa o estructuras con la misma secuencia o muy cercana, mientras que la interfaz completa está compuesta por todos los contactos que se pueden encontrar en estructuras PDB que contengan los dominios de interés e interaccionen. En ambos casos se obtienen interfaces diferenciadas para eucariotas y procariotas, siempre dependiendo de la disponibilidad de estructuras. Esta información fue integrada en un conjunto de 559 interacciones entre dominios donde cada para par de posiciones interdominio se detalla tanto la información coevolutiva como la estructural y fue utilizado en el primer estudio (sección 4.1). El procedimiento es el mismo para el segundo estudio (sección 4.2), donde se utilizan las interfaces completas de 490 interacciones entre dominios interproteína.

3.3 Clasificación de estructuras como procariotas o eucariotas

Usamos las anotaciones de 3did para recuperar los identificadores de PDB, cadenas y rangos de posiciones dentro de las cadenas para 37 126 interacciones entre dominios con estructura conocida en PDB. Obtuvimos los identificadores taxonómicos de cada cadena mediante SIFT [195]. Clasificamos cada cadena recorriendo la taxonomía del *National Center for Biotechnology Information* [196] (NCBI) en sentido ascendente hasta los niveles taxonómicos "*Eukaryotes*", "*Bacteria*" o "*Archaea*" (los términos utilizados en NCBI *taxonomy*). Para

los casos donde los dos dominios pertenecen a la misma proteína (y su especie está especificada en SIFT) fueron clasificadas como eucariota si se alcanza el nivel "*Eukaryotes*" o como procariota si se alcanzan los niveles "*Bacteria*" o "*Archaea*". En caso de que los dos dominios pertenezcan a dos proteínas distintas, se anotaron de la misma forma que en el caso anterior siempre que ambas proteínas alcancen el mismo nivel taxonómico superior (eucariota si las dos alcanza el nivel "*Eukaryotes*" o como procariota si las dos alcanzan el nivel "*Bacteria*" o el nivel "*Archaea*") y la especie estuviera especificada en SIFT. Finalmente, extrajimos las secuencias de los ficheros PDB que no pudieron ser clasificadas con el procedimiento anterior y realizamos búsquedas contra la base datos de secuencias TrEMBL [197] (reduciendo redundancia al 80% de identidad en secuencia con cd-hit para acelerar las búsquedas) usando el programa blastp (versión 2.2.18 [198]). Anotamos como procariotas o eucariotas utilizando las anotaciones filogenéticas proporcionadas por TrEMBL de la secuencia representativa encontrada (homóloga a la secuencia objetivo y con un >80% identidad en secuencia). De nuevo, en caso de tratarse de dos proteínas, su anotación debe ser consistente. Los 387 pares de cadenas de PDB que no pudieron ser clasificados con los anteriores criterios fueron descartados.

3.4 Protocolo para la detección de señales coevolutivas

Con el objetivo de obtener una cuantificación de la señal coevolutiva, diseñamos un protocolo computacional que consta de 3 partes principales (Figura 3.2): la búsqueda de secuencias homólogas, la construcción de un alineamiento emparejado a partir de las secuencias homólogas y el computo del modelo estadístico sobre el alineamiento emparejado. Construimos alineamientos emparejados para los 4 556 pares de dominios Pfam descritos anteriormente (sección 3.1.1). Para el primer estudio (sección 4.1), computamos el modelo estadístico para 559 alineamientos emparejados. Para el segundo estudio (sección 4.2) computamos 490 modelos con los alineamientos emparejados, más un total de 980 000 modelos con los alineamientos emparejados aleatorizados (sección 4.2.1).

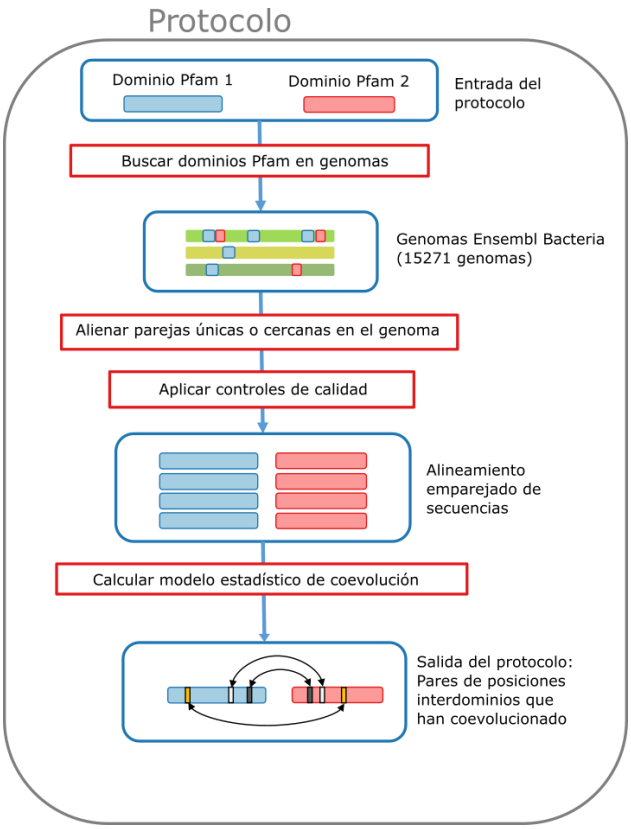


Figura 3.2 Esquema del protocolo desarrollado para la detección de coevolución entre dominios. Adaptada de [117].

3.4.1 BÚSQUEDAS DE SECUENCIAS HOMÓLOGAS

Dada una pareja de dominios Pfam de entrada (aunque potencialmente se puede usar cualquier pareja de modelos HMM), nuestro protocolo busca para cada dominio Pfam, utilizando su perfil HMM, proteínas que contengan miembros del dominio Pfam correspondiente en todas las regiones codificantes de los 15 271 genomas de especies procariotas usando el programa HMMER. Para ello se utiliza el programa hmmsearch con los parámetros “--noali”, “--domtblout” y “--cut_ga”. En concreto, el parámetro “--cut_ga” selecciona el umbral de significancia estadística establecido por los autores de los perfiles de Pfam, de tal forma que se trata de un umbral con un buen balance entre precisión y exhaustividad (*recall*). Dado el tamaño de la búsqueda, solo se guardan la información de aquellos casos donde se ha detectado una región homóloga con una significancia por encima del umbral establecido. Por el mismo motivo, se utiliza el parámetro “--noali” para no generar, en este paso, los alineamientos de las regiones homólogas con respecto del perfil HMM de Pfam. A los identificadores de proteína (provenientes de Ensembl bacteria) añadimos el identificador taxonómico o *taxid*, de la especie a la que pertenece (utilizando las anotaciones de Ensembl bacteria), para poder comprobar fácilmente las especies a la que pertenecen. Obviamente, los emparejamientos se realizan entre dominios de una misma especie, como se detalla a continuación.

3.4.2 CONSTRUCCIÓN DE ALINEAMIENTOS EMPAREJADOS

Como describimos en la introducción (sección 1.3.6), para buscar señales coevolutivas entre dominios es necesario construir alineamientos emparejados, donde cada secuencia (región homóloga) del primer dominio se empareja con una secuencia del segundo dominio de tal forma que estas secuencias pertenecen a dominios que realmente interaccionen físicamente o, al menos, es así en la gran mayoría de casos. Nuestro protocolo implementa tres estrategias para llevar a cabo esta tarea. La primera estrategia está basada en la pertenencia al mismo gen, la segunda en la adyacencia genómica y la tercera en la unicidad genómica:

- Mismo gen: para cada homólogo encontrado del primer dominio, el protocolo comprueba si se encuentra un homólogo del segundo dominio en la misma proteína, utilizando los identificadores de proteína obtenidos en el paso anterior. Si las dos secuencias se solapan, solo se emparejan cuando el solapamiento es inferior al 10% del tamaño del dominio más pequeño, o 20 aminoácidos como máximo. Los residuos solapantes son eliminados de uno de los dos dominios, ya que si no producirían correlaciones perfectas y espurias entre las columnas de la región solapante.
- Adyacencia genómica: se comprueba si para cada homólogo del primer dominio se encuentra algún homólogo del segundo en el mismo *contig* y a menos de 300 bases nucleicas de distancia en el genoma, utilizando las anotaciones de codones de inicio y fin proporcionadas por Ensembl bacteria.
- Unicidad genómica: se comprueba si los homólogos encontrados para ambos dominios Pfam son únicos en el genoma, esto es, para estos dominios Pfam solo se ha encontrado un solo homólogo en todo el genoma. Para utilizar esta estrategia, que puede ser activada o desactivada mediante un parámetro del protocolo, es aconsejable la existencia de evidencia externa sobre la interacción de una pareja de dominios Pfam, ya que empareja cualquier par de dominios que sean únicos en el genoma. En nuestro caso, esta evidencia viene del hecho que existen estructuras confirmando la interacción entre los dominios a través de 3did.

En todos los casos, si un dominio Pfam aparece más de una vez en una proteína, la proteína en cuestión es descartada ya que no es posible asegurar cuales de estos dominios interaccionan con el otro dominio Pfam de interés, evitando el ruido que podría introducirse de otro modo.

Por medio de estas estrategias, el protocolo recupera un conjunto de emparejamientos que se traducen en parejas de identificadores de proteínas. A partir de éstos, el protocolo recupera las secuencias codificantes para realizar los alineamientos múltiples de secuencias. Se construye un alineamiento por separado para cada dominio Pfam mediante el programa hmmalign (con parámetros “--trim” y “--allcol”) de HMMER. En estos alineamientos se descartan las columnas que no correspondan a un estado de *match* del perfil HMM, esto es, se eliminan las inserciones con respecto al perfil HMM. Los dos alineamientos resultantes constituyen el alineamiento emparejado, donde el orden de las secuencias asegura que cada secuencia del primer dominio se corresponde con una secuencia del segundo con la cual interacciona. Sobre este alineamiento emparejado se aplican los siguientes controles de calidad:

- Aunque no es habitual, puede ocurrir que la misma secuencia de un dominio aparezca en más de un emparejamiento. Por ejemplo, cuando un gen se encuentra flanqueado por dos genes que contienen secuencias pertenecientes a un mismo dominio Pfam. Esto puede dar lugar a que la secuencia del gen flanqueado se empareje dos veces, una vez con cada secuencia de los genes que le flanquean. En estos casos, es difícil determinar con cuál de los dos interactúa realmente o si lo hace con los. Por ello, si la misma secuencia de un dominio aparece en más de un emparejamiento, se eliminan todos los emparejamientos donde aparezca esta secuencia para evitar ambigüedades.
- Cualquier pareja donde uno de los dominios contenga más de un 20% de huecos (*gaps)* es descartada.
- Para secuencias emparejadas que tengan una identidad superior al 80% (considerando las secuencias de los dos dominios conjuntamente) con respecto a otras secuencias emparejadas, se elige una secuencia representativa de todas ellas. O, en otras palabras, se reduce la redundancia del alineamiento a un 80% de identidad en secuencia. Este paso se realiza mediante el programa cd-hit (versión 2) [199]. Este proceso reduce la señal filogenética, mejorando la predicción de contactos [137]. Siguiendo el lenguaje habitual en el campo [72], utilizamos el término secuencias efectivas para resaltar el hecho de que no son redundantes. Pero cabe recalcar que en toda la sección de resultados siempre que se menciona el número de secuencias, nos estamos refiriendo al número final de secuencias en los alineamientos emparejados, esto es, al número de secuencias en los alineamientos emparejados no redundantes al 80%.

Durante este procedimiento se hace uso del programa HHsuite [200] para traducir entre los distintos formatos de alineamientos múltiples de secuencias requeridos.

El resultado final es un alineamiento emparejado en formato FASTA.

3.4.3 COMPUTO DEL MODELO COEVOLUTIVO

Como se describió en la introducción (sección 1.3.4.1), en las metodologías DCA se propone modelar la familia de proteínas con un modelo de estadístico $P(A_1, \dots, A_L)$ que incluye interacciones de segundo orden (interacciones entre posiciones) a partir de un alineamiento múltiple de secuencias. Las correlaciones entre dos columnas pueden deberse efectos indirectos por medio de correlaciones con otras columnas. Mediante la construcción de un modelo que capture interacciones de segundo orden es posible minimizar estos efectos indirectos [71,103]. Sea

$$\mathbf{A} = (A_i^a), i = 1, \dots, L, \;\; a = 1, \dots, N$$

un alineamiento múltiple de N secuencias y L columnas. Consideraremos un alfabeto de 21 posibles caracteres q, los 20 aminoácidos naturales más el estado de hueco (*gap,* "-" en los alineamientos). Para ser

coherente con los datos, el modelo tiene que cumplir las frecuencias de aminoácidos por columnas y por pares de columnas observadas en el alineamiento

$$\forall i, A_i: \sum_{\{A_k | k \neq i\}} P(A_1, \dots, A_L) \equiv f_i(A_i)$$

$$\forall i, j, A_i, A_j: \sum_{\{A_k | k \neq i, j\}} P(A_1, \dots, A_L) \equiv f_{i,j}(A_i, A_j)$$

Nótese que en el modelo nos restringimos a interacciones de segundo orden, ya que órdenes superiores son, en principio, computacionalmente impracticables. Además de cumplir estas restricciones queremos el modelo más insesgado dado que existen una infinidad de modelos posibles bajo estas restricciones. Este modelo se puede obtener tomando la distribución que, entre las que son coherentes con los datos, maximiza la entropía

$$\mathrm{S} = -\sum_{\{A_i | i=1,\dots,L\}} P(A_i, \dots, A_L) \ln P(A_i, \dots, A_L)$$

La solución a este problema de optimización es bien conocida [107,108]. Una vez introducidos los multiplicadores de Lagrange, encontramos

$$P(A_1, \dots, A_L) = \frac{1}{Z} \exp\left\{\sum_i h_i(A_i) + \sum_{i<j} e_{ij}(A_i, A_j)\right\}$$

Los parámetros de este modelo tienen una interpretación simple, siendo h_i los sesgos aminoácidicos para cada posición y e_{ij} el acoplamiento estadístico entre pares de aminoácidos para cada par de posiciones. El factor de normalización Z, también conocido como función de partición, es

$$Z = \sum_{A_1,\dots,A_L} \exp\left\{\sum_i h_i(A_i) + \sum_{i<j} e_{ij}(A_i, A_j)\right\}$$

Un importante problema en este contexto es el computo de los marginales $P(A_i)$ y $P(A_i, A_j)$ desde el modelo $P(A_i, \dots, A_L)$ que son los que nos permiten contrastar el modelo con los datos. Considerando el tamaño habitual de las familias de proteínas y que para L posiciones existirían 21^L posibles secuencias, no es posible computar los marginales, debido al factor de normalización, de forma exacta y es necesario realizar inferencias aproximadas. Una de las diversas aproximaciones para afrontar este problema es mediante un modelo logístico de regresión multinomial [110] (ver la sección 1.3.4.1 de la introducción para otras posibles aproximaciones), que detallaremos a continuación.

Realizamos una regresión logística multinomial de cada posición A_i con respecto al resto de posiciones del alineamiento $\boldsymbol{A}_{\setminus i}$ [110,201]. La probabilidad condicionada de observar un aminoácido a_i en la posición i dados los aminoácidos en las otras posiciones viene dada por

$$P\big(A_i = a_i \mid \boldsymbol{A}_{\setminus i} = \boldsymbol{a}_{\setminus i}\big) \propto \exp\left\{h_i(a_i) + \sum_{j \neq i} e_{ij}(a_i, a_j)\right\}$$

Donde a_i y a_j son los aminoácidos en las posiciones i y j de la secuencia $\boldsymbol{a}$. Para cada posición i, los parámetros de acoplamiento e_{ij} que cuantifican la interacción de esta posición con el resto de las posiciones en el alineamiento son inferidas mediante una maximización de la verosimilitud (*likelihood*) condicionada logarítmica $LL_i(\boldsymbol{h}, \boldsymbol{e})$ con regularización L_2:

$$\boldsymbol{h}^*, \boldsymbol{e}^* = argmax_{\boldsymbol{h},\boldsymbol{e}}\left\{LL_i(\boldsymbol{h}, \boldsymbol{e}) - \lambda \sum_{\alpha} h_i^2(\alpha) - \lambda \sum_{j,\alpha,\beta} e_{ij}^2(\alpha, \beta)\right\}$$

donde,

$$LL_i(\boldsymbol{h}, \boldsymbol{e}) = \frac{1}{N}\sum_{s} log\ [\ (P(A_i = a_i^s | \boldsymbol{A}_{\setminus i} = \boldsymbol{a}_{\setminus i}^s)]$$

a_i^s es el aminoácido en la posición *i* de la secuencia $\boldsymbol{a}^s$ y el factor de regularización $\lambda = 0.01$ [110]. Siguiendo el protocolo propuesto por Ekeberg et al. [110], los valores de acoplamiento e_{ij} fueron centrados doblemente

$$e_{ij}^c(\alpha, \beta) = \ e_{ij}^*(\alpha, \beta) - e_{ij}^*(\alpha, :) - e_{ij}^*(:, \beta) + e_{ij}^*(:, :),$$

donde ":" representan el promedio sobre los correspondientes índices y los parámetros de acoplamientos fueron forzados a ser simétricos

$$e_{ij}^{sym}(\alpha, \beta) = \frac{1}{2}(e_{ij}^c(\alpha, \beta) + e_{ij}^c(\beta, \alpha))$$

Finalmente $e_{ij}^{sym}(\alpha, \beta)$ son los parámetros de acoplamiento encontrados. Dado que tenemos 441 (21^2), o 400 si descontamos el estado de hueco (*gap*), parámetros de acoplamiento entre cada par de posiciones, hemos de traducir estos parámetros en una sola medida que colapse los acoplamientos en *scores* entre posiciones. Aplicando la norma de Frobenius [109], obtenemos para cada par de columnas $i\ y\ j$

$$S_{crudo}(i, j) = ||e_{ij}^{sym}||^2 = \sqrt{\sum_{k,l \neq gap}^{q} e_{ij}^{sym}(\alpha, \beta)^2}$$

Se descarta la contribución de los huecos que se ha probado ser beneficioso para la predicción de contactos [202]. Nos referiremos a S_{crudo} como *score* crudo. Sobre este el *score* crudo se puede aplicar la corrección APC

$$S_{APC}(i, j) = \ S_{crudo}(i, j) - APC(i, j) = S_{crudo}(i, j) - \frac{S_{crudo}(:, j) S_{crudo}(i, :)}{S_{crudo}(:, :)}$$

donde $S_{crudo}(:,j)$ es el *score* crudo promedio de la posición j con el resto de posiciones del segundo dominio, $S_{crudo}(i,:)$ es el *score* crudo promedio de la posición i con el resto de posiciones del primer dominio y $S_{crudo}(:,:)$ es el *score* crudo promedio para todos los pares de posiciones [144]. Nos referimos a S_{APC} como *score* APC. Sobre él, se puede aplicar una estandarización mediante la mediana de la desviación absoluta

$$Z_{APC}(i,j) = (S_{APC}(i,j) - \widetilde{M})/1.4826\, MAD, \text{ con}$$

$$MAD = mediana(||(S_{APC}(i,j) - \widetilde{M}||)$$

donde $\widetilde{M}$ es la mediana de la distribución de los S_{APC}.

Nos referimos a este Z_{APC} como *z-score* APC.

3.5 Estimación del número de interacciones humanas con homólogos en procariotas

Estimamos el número de interacciones humanas con suficientes secuencias homólogas en especies procariotas para obtener una idea más clara del potencial que puede tener proyectar la predicción de contactos en eucariotas. De entre los posibles interactomas de especies eucariotas, elegimos trabajar con el interactoma humano ya que es el interactoma mejor estudiado y caracterizado. Para ello, obtuvimos una lista de interacciones proteína a proteína con suficientes secuencias homólogas en procariotas. Determinamos cuantas de estas interacciones tienen o no información estructural, ya sea información experimental o por medio del modelado por homología. Más abajo se describen los detalles de este análisis. A continuación, se resumen los principales pasos. Para cada par de proteína que interaccionan en el interactoma:

- Buscamos dominios Pfam dentro de ellas
- Encontramos la mejor plantilla de la interacción, esto es, un PDB conteniendo ambos interactores y cuya identidad de secuencia es la mayor posible.
- Determinamos el número de secuencias procariotas no redundantes al 80% de identidad en secuencia para cada uno de los dominios Pfam.
- Estimamos el subconjunto de interacciones donde ambos interactores tiene dominios Pfam con suficientes secuencias homólogas en procariotas, pero sin información estructural fiable.

Partimos de una lista de 214 806 interacciones proteína a proteína heterodiméricas, extraídas de BioGRID [203], implicando 16 053 proteínas humanas distintas. En primer lugar, buscamos dominios Pfam en las isoformas principales, obtenidas mediante APPRIS [204] (versión 2016_06.v17; conjunto de genes de Gencode24/Ensembl84), utilizando en programa hmmsearch (con parámetro "--cut_ga") y todos los perfiles HMM disponibles en Pfam. Detectamos alguna región homóloga cubriendo al menos un 80% del perfil HMM en 5 854 de los 14 831 perfiles de Pfam. Realizamos una búsqueda con los mismos criterios sobre las

secuencias de PDB (por medio del fichero pdb_seqres.txt, descargado en septiembre 2016) y calculamos la identidad en secuencia entre las secuencias humanas y aquellas del PDB tras alinear mediante hmmalign (parámetros “--trim” y “--allcol”).

Clasificamos las 214 806 interacciones humanas heterodiméricas en función de la información estructural disponibles para ellas. Para ello, consideramos todos los PDBs que contuvieran al menos un dominio Pfam de cada interactor como potencial plantillas estructural, o *template,* para la interacción. Para cada combinación de PDB y parejas de dominios Pfam, seleccionamos el mínimo porcentaje de identidad [154] de los dos porcentajes de identidad provenientes de los dos interactores. De todas las combinaciones de PDBs y parejas de dominios Pfam, seleccionamos la mejor plantilla estructural posible, esto es, aquel con el porcentaje de identidad más alto. Clasificamos las interacciones humanas conforme al porcentaje de identidad en secuencia con su mejor plantilla (Figura A2A del Anexo I):

- 3 789 interacciones resueltas (porcentaje de identidad ≥98%).
- 13 253 interacciones con plantillas fiables (porcentaje de identidad <98% y ≥ 30%).
- 14 515 interacciones con plantilla no fiables (porcentaje de identidad <30%).
- 165 812 interacciones sin posible plantilla.

A continuación, comprobamos para qué interacciones entre proteínas existen muchas secuencias homólogas en procariotas. Para ello, buscamos los 5 854 dominios Pfam encontrados en proteína humanas contra una base de datos no redundante (quitando redundancia al 80% de identidad en secuencia mediante cd-hit) de los 15 271 genomas procariotas. Para cada interacción, seleccionamos el par de dominios con la mejor plantilla posible como se ha explicado en el párrafo anterior. En caso de no existir plantilla posible, seleccionamos la pareja de dominios con más secuencias en procariotas (de nuevo considerando el mínimo de los dos interactores). Detectamos 158 790 interacciones donde para ambos interactores encontramos al menos un homólogo, 51 461 teniendo 100, 36 213 teniendo 500 y 22 958 teniendo 1 000.

En resumen, encontramos 36 213 interacciones, de un total de 214 806, con al menos 500 secuencias homólogas no redundantes en procariotas. Esto sugiere que hasta en un 17% del interactoma humano podría ser aplicable el método que proponemos. De estas 36 213 interacciones, encontramos 619 complejos con estructura conocida y 3 887 sin estructura conocida pero abordables mediante el modelado por homología (Figura A2B del Anexo I). Para las 31 707 interacciones restantes, el análisis de la coevolución en procariotas y su proyección en eucariotas podría ayudar a modelar convenientemente estas interacciones (Figura A2B del Anexo I. Esto supone un 15% del interactoma humano. Para 5 310 de estos 31 707 es posible encontrar plantillas no fiables, donde nuestra aproximación podría ayudar a identificar plantillas fiables. Para 23 433 (un 89%) de los 26 397 de las interacciones sin plantillas, es posible encontrar buenas plantillas para modelar los interactores por separado. En estos casos, se podría combinar nuestra aproximación con metodologías de acoplamiento de proteínas (*protein-protein docking*).

3.6 Estimación de la divergencia entre procariotas y eucariotas

A continuación, se detalla la estimación de la divergencia entre las proteínas en especies procariotas y eucariotas en secuencia y la divergencia estructural en las interfaces para todos los casos (195 parejas de dominios Pfam; 152 intraproteína y 43 interproteína) donde hemos encontrado homologías entre procariotas y eucariotas y hay disponible, al menos, una estructura de la interacción entre dominios tanto para alguna especie procariota como para alguna especie eucariota.

3.6.1 ESTIMACIÓN DE LA DIVERGENCIA EN SECUENCIA

Una vez seleccionado el complejo representativo (sección 3.2) para cada caso, alineamos las secuencias de los dos dominios tanto en procariotas como en eucariotas usando los perfiles de Pfam correspondientes y el programa hmmalign (parámetros "--trim" y "--allcol"). Tras hallar el porcentaje de identidad mediante estos alineamientos para los dos dominios, tomamos como estimación de la divergencia en secuencia el menor de estos dos porcentajes de identidad. El racional de usar el menor de los dos posibles es que el dominio con el menor porcentaje de identidad tenderá a ser un mejor estimador de la divergencia de la interacción [154].

Como medida complementaria, extrajimos la distancia evolutiva de las secuencias (hojas en el árbol) de los complejos representativos en procariotas y eucariotas en los árboles filogenéticos (sección 3.6.1). Para cada caso, seleccionamos la mayor distancia evolutiva de los dominios en interacción, esto es, la mayor distancia evolutiva de las dos distancias evolutivas que se calculan para cada caso (una para cada uno de los dominios Pfam que intervienen en la interacción).

3.6.2 ESTIMACIÓN DE LA DIVERGENCIA ESTRUCTURAL

Gracias a la clasificación de estructuras como procariotas o eucariotas (sección 3.3) y la definición de interfaces completas o representativas (sección 3.2), tenemos, para cada caso, una interfaz completa (o representativa) para procariotas y otra para eucariotas. Para estimar el grado de divergencia entre estas dos interfaces completas (o representativas), definimos como conservación estructural el porcentaje de pares de posiciones en contacto en la interfaz completa (o representativa) en procariotas cuyas posiciones correspondientes en eucariotas (las posiciones en eucariotas con las que alinean, a veces referidas como posiciones homólogas) están también en contacto en la interfaz completa (o representativa) en eucariotas.

3.7 Estimación de la calidad de los alineamientos

Estimamos la calidad del alineamiento de las secuencias eucariotas para comprobar si podía afectar a la proyección de contactos desde procariotas a eucariotas. Para ello, utilizamos la selección de complejos representativos (sección 3.2) y consideramos las posiciones que en eucariotas alinean con las posiciones de las interfaces completas en procariotas. Partiendo de estas posiciones, recuperamos la calidad estimada para cada posición del alineamiento (número entre 0 y 1 para cada posición estimando la adecuación de cada aminoácido en su posición, información proporcionada por HMMER de acuerdo con la probabilidad posterior por posición), y consideramos los huecos (*gaps*) como posiciones con calidad igual a 0. Obtuvimos la calidad media estimada por posición para los dos dominios y seleccionamos el mínimo de estos dos promedios. Alternativamente, para evitar depender de información estructural de la interfaz, calculamos una estimación equivalente utilizando solo las posiciones que han coevolucionado. Si la calidad media por posición es inferior a 0.8, decimos que se trata de un alineamiento de baja calidad (Figura A8 del Anexo I).

3.8 Reconstrucción de árboles filogenéticos

Para reconstruir árboles filogenéticos recuperamos un conjunto de 89 proteomas de referencia incluyendo tanto especies eucariotas como procariotas (listadas más abajo). Este conjunto de especies de referencia fue tomado de un estudio de proteomas completamente secuenciados para análisis filogenéticos (www.phylomedb.org/phylome_514 y www.phylomedb.org/phylome_28) en PhylomeDB [205]. Añadimos 4 especies para las cuales SIFTS tiene anotadas 500 cadenas de PDB (*T. thermophilus HB8*, *Staphylococcus aureus*, *Sus scrofa*, y *Oryctolagus cuniculus*). Usando el programa hmmsearch (parámetro "--cut_ga"), buscamos en estos proteomas los 274 Pfam de dominios en interacción pertenecientes a 195 pares de dominios Pfam para los cuales se pueden encontrar estructura de la interacción tanto en procariotas como en eucariotas. Alineamos las regiones homólogas en estos proteomas y la secuencia del PDB seleccionado como representativo (sección 3.2) al correspondiente perfil HMM de Pfam descartando todas las posiciones no pertenecientes al perfil mediante el programa hmmalign (parámetros "--trim" y "--allcol"). Inferimos árboles de máxima verosimilitud desde estos alineamientos usando el programa IQ-TREE [206] (parámetros "-m TESTNEWONLY -b 100 -nt 2"; version 1.4.4) con un *boostrapping* de tamaño 100 (remuestreando con reemplazamiento 100 veces). 7 casos fueron descartados porque uno de los interactores no tiene un número suficiente de secuencias en los proteomas de referencia (menos de 4 secuencias).

Los nombres de las especies utilizadas son: *Anopheles gambiae, Aquifex aeolicus VF5, Arabidopsis lyrata, Arabidopsis thaliana, Aspergillus fumigatus A1163, Bacillus subtilis, Bacteroides thetaiotaomicron VPI-5482, Bos taurus, Brachypodium distachyon, Bradyrhizobium japonicum, Branchiostoma floridae, Caenorhabditis elegans, Candida albicans, Canis familiaris, Chlamydia trachomatis A/HAR-13, Chlamydomonas reinhardtii, Chloroflexus aurantiacus J-10-fl, Ciona intestinalis, Cryptococcus neoformans var. neoformans JEC21, Cucumis sativus, Danio rerio, Deinococcus radiodurans R1, Dictyoglomus turgidum DSM 6724, Dictyostelium discoideum, Drosophila melanogaster, Escherichia coli*

K-12, Fusobacterium nucleatum subsp. nucleatum ATCC 25586, Gallus gallus, Geobacter sulfurreducens PCA, Giardia lamblia ATCC 50803, Gloeobacter violaceus PCC 7421, Glycine max, Halobacterium sp. NRC-1, Homo sapiens, Ixodes scapularis, Korarchaeum cryptofilum (strain OPF8), Leishmania major, Leptospira interrogans serovar Lai str. 56601, Macaca mulatta, Medicago truncatula, Methanocaldococcus jannaschii DSM 2661, Methanosarcina acetivorans C2A, Micromonas pusilla CCMP1545, Mimulus guttatus, Monodelphis domestica, Monosiga brevicollis, Mus musculus, Mycobacterium tuberculosis, Nematostella vectensis, Neurospora crassa, Ornithorhynchus anatinus, Oryctolagus cuniculus, Oryza sativa subsp. indica, Oryza sativa subsp. japonica, Ostreococcus lucimarinus (strain CCE9901), Ostreococcus tauri, Pan troglodytes, Phaeosphaeria nodorum SN15, Physcomitrella patens subsp. patens, Physcomitrella patens subsp. patens, Plasmodium falciparum 3D7, Populus trichocarpa, Pseudomonas aeruginosa PAO1, Rattus norvegicus, Rhodopirellula baltica SH 1, Ricinus communis, Saccharomyces cerevisiae, Schistosoma mansoni, Schizosaccharomyces pombe (strain 972/ATCC 24843), Sclerotinia sclerotiorum 1980 UF-70, Selaginella moellendorffii, Sorghum bicolor, Staphylococcus aureus, Streptomyces coelicolor A3 (2), Sulfolobus solfataricus P2,Sus scrofa, Synechocystis sp. PCC 6803 substr. Kazusa, Takifugu rubripes, Thalassiosira pseudonana, Thermococcus kodakarensis KOD1, Thermodesulfovibrio yellowstonii DSM 11347, Thermotoga maritima MSB8, Thermus thermophilus HB8, Trichomonas vaginalis, Ustilago maydis, Vitis vinifera, Xenopus tropicalis, Yarrowia lipolytica, and Zea mays.

3.9 Influencia de la entropía en la relación entre coevolución y conservación estructural

Para considerar la posible influencia de la entropía en la relación entre coevolución y conservación estructural, calculamos la entropía (medida de conservación en secuencia) de las posiciones de los alineamientos emparejados. La Figura A4A del Anexo I muestra las distribuciones de entropía para las posiciones en interfaces y para las posiciones que han coevolucionado. Como es de esperar, las posiciones que han coevolucionado no se corresponden con posiciones altamente conservadas en secuencias (baja entropía), dado que una cierta variabilidad es necesaria para observar covariación. Las posiciones que han coevolucionado se encuentran enriquecidas en entropías intermedias. A continuación, detallamos el análisis llevada a cabo para descartar que estos sesgos expliquen la relación entre coevolución y conservación estructural que observamos (ver sección 4.1.3 de resultados).

Creamos un conjunto de contactos corregido seleccionando pares de posiciones (en contacto) con una probabilidad en función únicamente de la entropía de las posiciones. Elegimos esta probabilidad de forma que el conjunto corregido resultante tenga la misma distribución de entropía que el conjunto de posiciones que han coevolucionado. Para ello, computamos la probabilidad conjunta $P(H_a, H_b)$ de observar, para un par de residuos en contacto, dos valores de entropía H_a, H_b en las columnas correspondientes del alineamiento. También computamos la probabilidad conjunta $P_C(H_a, H_b)$ restringida a posiciones que han coevolucionado. La entropía de cada posición se calcula

$$H[a] = -\sum_{a} \pi(a) \log_{21} \pi(a)$$

donde π es la distribución de aminoácidos (20 aminoácidos naturales más el estado de *gap*) estimada para la posición. Para evitar estados no observados π fue estimada con la aproximación de Laplace, o *pseudocounts* (añadiendo 21 observaciones, una observación para cada aminoácido), $\pi(a) = (n_a + 1) \,/\, (N + 21)$, donde n_a es la frecuencia empírica del aminoácido a y N el número de secuencias en el alineamiento. La probabilidad conjunta fue calculada a intervalos de tamaño 0.1 (el rango es entre 0 y 1 dado que la base del logaritmo es 21). Para cada par de residuos en contacto $\{x, y\}$, computamos el peso $w(x, y) = P_C\big(H[\pi_x], H[\pi_y]\big) \,/\, P(H[\pi_x], H[\pi_y])$, donde $H[\pi_x]$ y $H[\pi_y]$ son los valores de entropía correspondientes a las posiciones x e y del alineamiento. Estos pesos definen un conjunto corregido de contactos que por construcción tienen la misma distribución de entropías que la de las posiciones que han coevolucionado, $P_C(H_a, H_b)$. Los promedios sobre los contactos se calculan

$$\langle A \rangle = \frac{\sum_{\{x,y\}} A(x,y) w(x,y)}{\sum_{\{x,y\}} w(x,y)}$$

Si la entropía fuera suficiente para explicar la relación entre coevolución y conservación estructural, los contactos conservados (estructuralmente entre procariotas y eucariotas) debería ser tan frecuente como en el conjunto de posiciones que han coevolucionado. Como se muestra en la Figura A4B del Anexo I, es prácticamente idéntica a la de las posiciones en la interfaz. De hecho, el enriquecimiento de las posiciones que han coevolucionado (interproteína) sigue siendo altamente significativo cuando se compara contra en conjunto corregido por la entropía (*p-valor* < 10^{-10}, test de Fisher de una cola tanto para interproteína como intraproteína), descartando el efecto de la entropía en la relación entre coevolución y conservación estructural.

3.10 Evaluación del rendimiento de las predicciones

Realizamos dos medidas diferenciadas del rendimiento de las predicciones, dependiendo si se consideran las predicciones de todos los casos conjuntamente (precisión) o si se considera cada caso de forma separada (proporción de casos predichos correctamente). Está última cuantificación permite analizar las distintas aproximaciones en términos de aplicabilidad, esto es, cual es el rendimiento con respecto al número de casos que podemos predecir correctamente. Una mejora en la precisión se suele traducir en una mayor aplicabilidad, pero esto no tiene por qué ser necesariamente así. Una mejora de la precisión puede estar asociada a una mejora solo en determinados casos en detrimentos de otros, lo que podría reducir su aplicabilidad.

3.10.1 PRECISIÓN GLOBAL CONSIDERANDO TODOS LOS CASOS CONJUNTAMENTE

La evaluación del rendimiento de las predicciones de las diferentes aproximaciones se realiza, siguiendo la norma en el campo [144,145], cuantificando la precisión de las predicciones. La precisión se define como

$$Precisión = \frac{verdaderos\ positivos}{falsos\ positivos + verdaderos\ positivos}$$

La precisión mide la proporción de predicciones correctas con respecto al número total de predicciones consideradas. Por ello, evaluamos la precisión con respecto a un número variable de predicciones. Mostramos la precisión con respecto al número de predicciones de dos formas diferentes, pero fundamentalmente equivalentes, ya sea cuantificando el número de predicciones directamente (ver p. ej. Figura 4.12A) o con respecto a un umbral del correspondiente *score* (ver p. ej. Figura 4.2A). Es importante recalcar que la correspondencia entre ambas variables es inequívoca, cada número de predicciones se corresponde con un valor del umbral del *score* en cuestión.

El hecho de priorizar la precisión sobre la exhaustividad (*recall*) se debe a que un número reducido pero fiable de predicciones es suficiente para generar buenos modelos tridimensionales [144,145].

3.10.2 PROPORCIÓN DE CASOS PREDICHOS CORRECTAMENTE

Dado un parámetro de precisión objetivo (por ejemplo, precisión ≥ 0.8) y un umbral del *score* dado, se comprueba para cada caso si existen pares de posiciones con un *score* por encima del umbral, y decimos que es un caso predicho para ese umbral. Si la precisión de estas predicciones de contacto es igual o superior a la precisión objetivo, decimos que este caso ha sido correctamente predicho. Como medida de la aplicabilidad, utilizamos la proporción de casos predichos correctamente (PCPC)

$$PCPC = \frac{Casos\ predichos\ correctamente}{Casos\ predichos}$$

Para cada aproximación, empezamos desde el *score* más alto e iteramos de forma decreciente sobre el *score* computando la proporción de casos predichos correctamente.

3.11 Estimación de la señal filogenética

Para cada alineamiento emparejado, calculamos la distancia entre cada par de secuencias usando la matriz de sustitución BLOSUM62 [207]. En el caso de alineamientos con más de 5 000 secuencias, se tomó una muestra aleatoria de 5 000 secuencias para limitar la memoria y el tiempo de computo. Sobre la matriz cuadrada de distancias (con tamaño igual al número de secuencias en el alineamiento emparejado de partida), aplicamos el análisis de componentes principales (PCA) usando la librería de R FactoMineR [208].

Tomamos la variabilidad explicada por la primera componente principal como estimación de la señal filogenética. Para la visualización de grupos de secuencias, usamos el algoritmo de partición en torno a medianas (*Partitioning Around Medoids* o PAM), considerando desde 2 hasta 20 grupos en el espacio definido por las dos primeras componentes principales del PCA, y seleccionando aquel agrupamiento con el mayor *silhouette score* (que proporciona una medida de cuan similares son las secuencias dentro de su grupo con respecto a las de otros grupos) promedio.

3.12 Estimación de la significancia estadística de la primera predicción correcta

En esta sección estimamos cual es la significancia estadística de observar que en 64 casos de 350 casos con alineamientos pequeños (con pocas secuencias) la predicción con un mayor *score* APC es una predicción correcta, esto es, un contacto físico. Podemos obtener una estimación simple de la significancia estadística de esta observación mediante una prueba binomial. Para todos los casos, calculamos cual es la probabilidad de que la primera predicción sea correcta al azar:

$$P(A) = \frac{Número\ de\ contactos}{Número\ de\ predicciones}$$

Donde el número de predicciones es N*M, siendo N el número de posiciones del primer dominios y M el del segundo y el número de contactos es el número de predicciones que son contactos en la interfaz. La mayor probabilidad observada es 0.146 que corresponde al caso en el que proporcionalmente hay un mayor número de contactos con respecto al número total de predicciones. Si tomamos esta probabilidad para los todos los casos y considerando la observación de 64 éxitos en 350 repeticiones, el *p-valor* de una cola asociado a la prueba binomial es 0.033. La probabilidad 0.146 es atípicamente alta, 327 de los 350 casos tienen una probabilidad menor a 0.05. Si asumimos una probabilidad de éxito del 0.05 y considerando la observación de 54 éxitos en 327 repeticiones, el *p-valor* de una sola cola de la prueba binomial es $1.59*10^{-14}$. Como estamos asumiendo unas probabilidades de éxito muy mayores que las reales en la mayoría de los casos, estos *p-valores* representan cotas superiores de los *p-valores* reales. Los *p-valores* reales no son computables de forma sencilla, pero es esperable que sean mucho menores.

3.13 Disponibilidad de datos y código fuente

Los datos del primer estudio (sección 4.1) se encuentran disponibles en http://cointerfaces.bioinfo.cnio.es/

El código fuente del protocolo desarrollado para la detección de señales coevolutivas se encuentra disponible en https://github.com/juan-rodriguez-rivas/cointerfaces

El código fuente con la implementación de la corrección MEND se encuentra disponible en https://github.com/juan-rodriguez-rivas/mend

4. RESULTADOS

La sección de resultados se encuentra dividida en dos subsecciones, la primera dedicada al estudio de la relación entre coevolución y la conservación estructural de las superficies de interacción entre proteínas, y la segunda a la mejora de las predicciones de contacto entre proteínas basada en coevolución por medio de la detección y corrección de la distribución de fondo (*background*) específica de cada caso.

4.1 Coevolución y conservación estructural en interfaces

Tanto la estructura de las proteínas como la de los complejos se encuentra altamente conservada en comparación a las secuencias de aminoácidos (ver secciones 1.3.2 y 1.3.5). En esta sección estudiaremos la relación entre coevolución y la conservación estructural de la interfaz. Específicamente, analizaremos, por un lado, si las interfaces con mayores señales de coevolución son más similares estructuralmente comparando estructuras en procariotas y eucariotas. Por otro, estudiaremos si los pares de residuos en contacto físico y con mayor señal de coevolución tienden a corresponderse con contactos estructuralmente conservados.

Presentaremos los resultados asociados a esta sección en el siguiente orden: i) análisis del conjunto de datos obtenido, ii) las señales de coevolución detectadas y su relación con propiedades estructurales, iii) la relación entre conservación estructural y coevolución, iv) la proyección de contactos a largas distancias evolutivas, v) un análisis de interacciones entre proteínas concretas en detalle y de carácter ilustrativo.

4.1.1 Análisis del conjunto de datos obtenido

Para investigar la relevancia de la coevolución en la conservación estructural en las interfaces de proteínas entre homólogos muy divergentes, construimos un conjunto de datos de interfaces entre dominios que integra tanto información estructural como de coevolución. En resumen (ver Materiales y Métodos para más detalles), partimos de un conjunto de pares de dominios Pfam no redundantes para los cuales existe información estructural de la interacción mediante la base de datos 3did [182]. Construimos alineamientos emparejados y detectamos señales de coevolución a partir de 15 271 genomas completos de especies procariotas (provenientes de Ensembl bacteria) gracias a la implementación de un protocolo específico diseñado para esta tarea. Integramos la información coevolutiva con la información estructural disponible en PDB [186], y analizamos estos datos obteniendo los resultados que detallamos a continuación.

Como resultado de la aplicación de nuestra metodología, obtuvimos 559 casos (pares de dominios Pfam) no redundantes (secciones 3.1 y 3.4 de Materiales y Métodos). Es importante recalcar que todos los alineamientos emparejados correspondientes a estos 559 casos tienen al menos 500 secuencias (no redundantes al 80%, sección 3.4.2 de Materiales y Métodos), debido al requerimiento de DCA de un gran

número de secuencias no redundantes (Introducción, sección 1.3.4.1). La distribución del número de secuencias por caso se muestra en la Figura A1 del Anexo I.

Dado que estas interacciones pueden ser tanto intraproteína o interproteína dependiendo de si ambos dominios se suelen encontrar en la misma proteína o en proteínas distintas (Figura 4.1A), cuantificamos el número de veces que ocurre una situación o la otra en cada uno de los alineamientos emparejados. La proporción $P_{inter}/(P_{intra}+P_{inter})$, donde P_{inter} es el número de parejas de dominios en la misma proteína y P_{intra} en proteínas distintas, es bimodal con máximos en torno a 0 y 1, por lo que las parejas de dominios Pfam tiene una tendencia muy clara a ser intraproteína o interproteína con pocos casos intermedios (Figura 4.1B). Estos casos intermedios se deben a eventos de fusión y de fisión que, como se explicó en la introducción (sección 1.2), es la información utilizada por un conjunto de métodos para predecir interacciones entre proteínas.

Clasificamos cada estructura tridimensional como procariota o eucariota (Materiales y Métodos, sección 3.3), obteniendo distintos números de casos dependiendo de la información estructural disponible en procariotas y eucariotas (Figura 4.1C). Cabe destacar la obtención de 152 casos intraproteínas y 43 casos interproteína donde tenemos información estructural tanto en especies procariotas como en procariotas (Figura 4.1C). Estos casos son especialmente relevantes en nuestro estudio ya que nos permiten comparar las señales de coevolución con la conservación o divergencia estructural en las interfaces.

En muchos casos, podemos encontrar varias, o incluso muchas, estructuras de interacción entre dominios perteneciente a la misma pareja de dominios Pfam, ya sea en la misma o distintas especies. En relación con la variabilidad conformacional existente en estas estructuras, usamos dos estrategias distintas para definir el conjunto de contactos que conforman cada interfaz entre dominios (Materiales y Métodos, sección 3.2). Por un lado, definimos una *interfaz completa* incluyendo todos los contactos interdominio que pudieran encontrarse en todos los complejos homólogos donde aparezcan la pareja de dominios Pfam. Esta definición incorpora información proveniente de distintos escenarios biológicos (p. ej. conformaciones alternativas) y metodológicos. Por otro lado, seleccionamos una única estructura de un complejo representativo, seleccionado por la calidad de su alineamiento (sección 3.2), y extrajimos los contactos en su interfaz, a la que nos referiremos como *interfaz representativa*. Extrajimos ambos tipos de interfaz para todas las parejas de dominios Pfam y de forma diferenciada para estructuras procariotas y eucariotas. Ambos tipos de interfaz fueron considerados en todos los análisis. Para mejorar la legibilidad, utilizaremos la interfaz completa como la de referencia, indicando los resultados relativos a la interfaz representativa cuando sea necesario. Por defecto, utilizamos un umbral de distancia de 8 angstrom entre átomos pesados para definir contactos, que es la medida más habitual en el campo [144,145], aunque otros umbrales son utilizados de forma complementaria a lo largo del estudio.

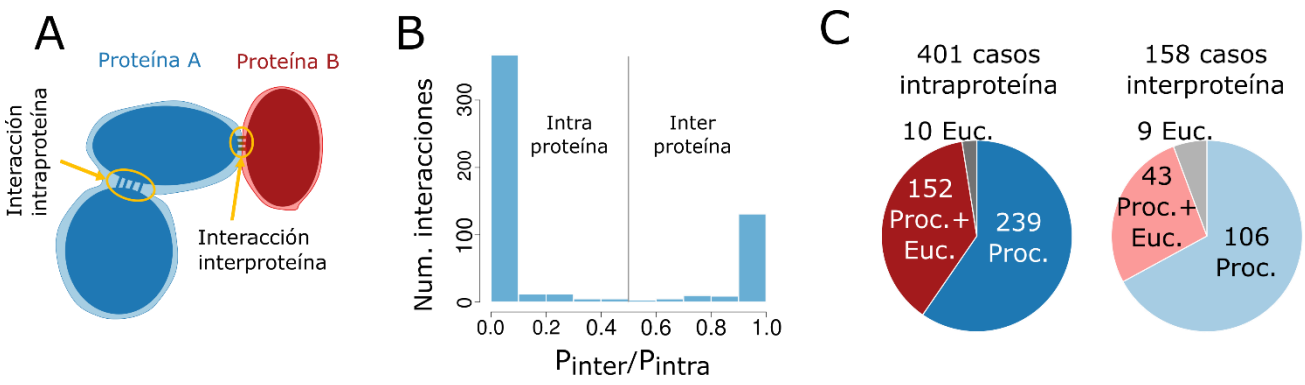


Figura 4.1 Resumen esquemático del conjunto de datos obtenido (adaptada de [117]). A) Los dos tipos de interacciones entre dominios que es posible encontrar: en casos intraproteína los dos dominios están codificados dentro de la misma proteína; en casos interproteína se encuentran en proteínas distintas. B) Histograma de la proporción de emparejamientos interproteína respecto del total en el alineamiento emparejado, esto es, P_{inter} /(P_{intra} + P_{inter}), donde P_{intra} es el número de parejas de dominios en la misma proteína y P_{inter} en proteínas distintas en los emparejamientos de los alineamientos. C) Composición de los 559 casos del conjunto de datos de acuerdo con la clasificación interproteína e intraproteína y la disponibilidad de estructuras 3D en procariotas y eucariotas. Figura adaptada de [117].

4.1.2 Señales de coevolución en procariotas y contactos físicos

Para cada uno de los 559 pares de dominios Pfam, computamos los *z-scores* APC para cada par de aminoácidos posibles entre los dos dominios (Materiales y Métodos, sección 3.4.3). Este valor cuantifica la influencia mutua en términos coevolutivos que se han ejercido dos posiciones, cada una perteneciente a uno de los dominios. Cabe recordar que los *z-scores* APC se obtienen tras computar el modelo DCA, aplicar la corrección APC a los *scores* crudos resultantes y, finalmente, estandarizar mediante la desviación absoluta mediana (Materiales y Métodos, sección 3.4.3). Tras añadir la información estructural (sección 3.2 de Materiales y Métodos), para cada par de aminoácidos interdominio tenemos tanto la información coevolutiva como la estructural.

La integración de información coevolutiva y estructural permite evaluar la relación entre el *z-score* APC y los contactos físicos en interfaces (Materiales y Métodos, sección 3.10.1). La Figura 4.2 (paneles A y B) muestra la relación existente entre ambas medidas en casos con estructura en procariotas. Como era de esperar [144,145], los *z-scores* APC más altos están claramente asociados con contactos físicos. Como umbral de referencia elegimos un *z-score* = 8, representando un compromiso entre una fuerte asociación con contactos y la recuperación de un número elevado de pares de posiciones. Para valores superiores del *z-score* APC diremos que el par de posiciones *ha coevolucionado* para simplificar el lenguaje, aunque ha de entenderse que, en realidad, significa que se ha detectado una fuerte covariación específica entre las dos posiciones probablemente debida a coevolución (ver sección 1.3.4.1). La predicción de contactos para cada par de dominios Pfam estará compuesta por los pares de posiciones que tengan un *z-score* APC superior a este valor. Si un determinado par de dominios Pfam tiene al menos un par de posiciones con un *z-score* APC ≥ 8, diremos que este caso ha sido predicho.

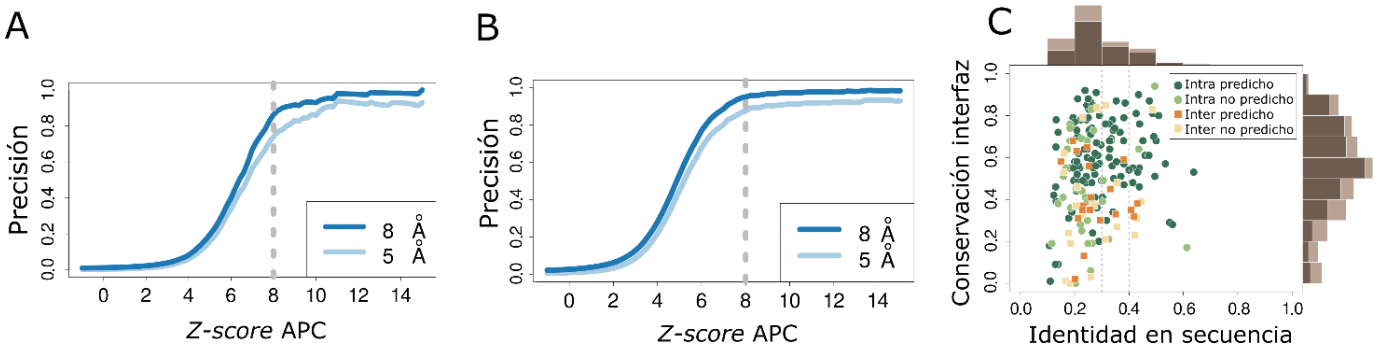


Figura 4.2 Precisión de las predicciones de contactos en procariotas y conservación en estructura y secuencia (adaptada de [117]). Precisión de la predicción de contactos (la proporción de predicciones correctas con respecto a las predicciones consideradas) en función del umbral del *z-score* APC para 149 casos interproteína (A) y 391 intraproteína (B), todos ellos con estructura en procariotas. Para cada valor del *z-score* APC, se consideran como predicciones positivas todos los pares de posiciones con un *z-score* APC superior a ese valor. La línea punteada vertical en gris indica el valor de referencia elegido para un *z-score* APC de 8. Se utilizan dos umbrales para la definición de contacto físico: 8 angstroms (azul oscuro) y 5 angstroms (azul claro). C) Conservación en estructura de la interfaz e identidad en secuencia de los 43 casos interproteína y 152 intraproteína con estructura en procariotas y eucariotas. La conservación de la interfaz se calcula como la proporción de contactos en las interfaces en procariotas que también están en contacto en eucariotas (Materiales y Métodos, sección 3.6.2). La identidad en secuencia indica la proporción de aminoácidos idénticos en el alineamiento entre los complejos representativos en procariotas y eucariotas (Materiales y Métodos, sección 3.6.1). Una interacción entre dominios se clasifica como predicha si se ha detectado coevolución entre al menos un par de posiciones interdominio. Los histogramas marginales representan el número de casos en gris claro y los casos predichos en gris oscuro.

Para medir la conservación estructural de la interfaz entre complejos procariotas y eucariotas, obtuvimos la proporción de contactos conservados con respecto del tamaño total de la interfaz en procariotas. El número de contactos conservados se corresponde con el número de pares de posiciones que están en contacto en procariotas y, a su vez, el par de posiciones correspondientes en eucariotas según el alineamiento están también en contacto en las interfaces de eucariotas. Es decir, pares de posiciones que están en contacto tanto en las interfaces procariotas como en las interfaces eucariotas. Es posible cuantificar la conservación de la interfaz para los 152 casos intraproteína y 43 interproteína donde tenemos información estructural tanto para procariotas como para eucariotas. En la Figura 4.2C se muestra la divergencia estructural con respecto a la divergencia en secuencia, constatando la existencia de importantes divergencias. De hecho, la mayor parte de casos se encuentran para identidades en secuencia inferiores al 30%, donde no es posible utilizar el modelado por homología de forma fiable debido a que las divergencias estructurales son comunes en este escenario [154].

En definitiva, el conjunto de datos utilizado está compuesto por 559 parejas de dominios Pfam, 401 intraproteína y 158 interproteína. Entre éstos, tenemos 152 casos intraproteína y 43 interproteína con estructura en procariotas y eucariotas. Este conjunto permite realizar una primera evaluación sistemática de la relación entre coevolución y conservación estructural en interfaces.

4.1.3 Coevolución y conservación estructural

En este conjunto de datos detectamos fuertes señales de coevolución, *z-scores* APC superiores a 8, en 20 de los 43 casos interproteína y 121 de los 152 casos intraproteína. La menor proporción de casos predichos (sin señales fuertes de coevolución) y baja conservación estructural (ver marginal derecho de la Figura 4.2C) podría indicar que la coevolución está asociada a interacciones con una mayor conservación estructural. Para analizar esta observación en más detalle, estudiamos la relación existente entre la conservación estructural de la interfaz y el grado de coevolución detectada en cada caso. Para ello, calculamos la media de los 5 *z-scores* APC más altos por caso, medida que estima la cantidad de coevolución detectada en cada caso y que denominaremos *acoplamiento en la interfaz*. Como se observa en la Figura 4.3A, el acoplamiento en la interfaz se corresponde con un valor mínimo de la conservación estructural de la interfaz, esto es, cuanto mayor sea el acoplamiento en la interfaz, mayor es el nivel mínimo de conservación estructural observado. Por tanto, la coevolución detectada en la interfaz está asociada a un mínimo nivel de conservación estructural.

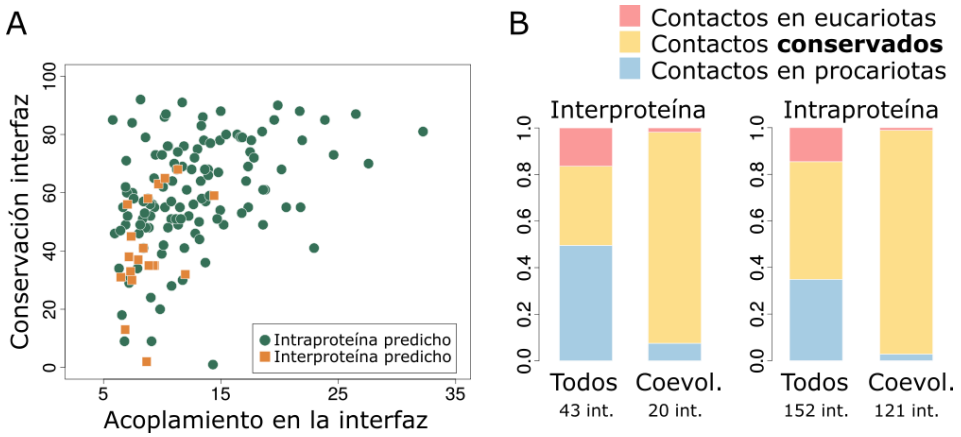


Figura 4.3 Conservación estructural y coevolución (adaptada de [117]). A) Relación entre la conservación estructural de la interfaz y el acoplamiento en la interfaz, medida que estima la cantidad de coevolución detectada en cada caso, para 20 casos interproteína y 121 intraproteína predichos y con estructura en procariotas y eucariotas. B) Proporción de contactos conservados considerando todos los contactos o solo el subconjunto de contactos en posiciones que han coevolucionado. En azul, contactos en procariotas que no están en contacto en eucariotas; en rojo, contactos en eucariotas que no están en contacto es procariotas; en amarillo, contactos conservados, en contacto tanto en procariotas como en eucariotas.

A continuación, en lugar de la conservación estructural de la interfaz en su conjunto, estudiamos la coevolución y conservación en pares de posiciones concretos. Podemos distinguir 3 situaciones distintas: pares de posiciones que solo están en contactos en procariotas, aquellas que están en contacto solo en eucariotas y, finalmente, las que están en contacto en ambas. Nos referiremos a esta última categoría como contacto conservados. Si nos fijamos en procariotas, de los 56 pares de posiciones que han coevolucionado en interproteína, 52 están en contacto físico. En intraproteína, 1070 están en contacto de los 1107 que han coevolucionado. La gran mayoría de estos pares de posiciones están también en contacto en las interfaces de eucariotas, es decir, se trata de contactos conservados (Figura 4.3B). Esta observación contrasta con lo que esperable al azar ya que solo un 37% de los todos contactos interproteína en procariotas o eucariotas

están conservados, mientras que el 91% (48 de 53) están conservados cuando han coevolucionado. Además, 3 de los 4 pares aparentemente no conservados se encuentran relativamente cercanos, a menos de 10 angstroms de distancia. Esta relación se mantiene para complejos particularmente divergentes (Figura A3, Anexo I; <30% de identidad en secuencia), con 23 contactos conservados de 25. Se observa la misma tendencia en intraproteína, la conservación de contactos cambia desde 50% para todos los contactos al 96% cuando detectamos coevolución, con 1039 contactos conservados de 1 082. De nuevo, encontramos esta relación también en complejos divergentes (Figura A3, Anexo I), con 583 contactos conservados de 615.

Para medir la significancia estadística de la relación entre coevolución y conservación de contactos usamos la prueba exacta de Fisher, con categorías contacto conservado o no conservado y par de posiciones que ha coevolucionado o no ha coevolucionado. El número de posiciones que ha coevolucionado y que corresponde a un contacto conservado es extremadamente improbable al azar, con *p-valores* (*p-values*) inferiores a 10^{-10} para la prueba de una cola (*one-tailed test*), tanto para interproteína como para intraproteína. Este resultado también se observa en complejos muy divergentes (<30% de identidad en secuencia; 29 casos interproteína y 100 intraproteína), con *p-valores* inferiores a 10^{-6}.

Como prueba adicional, tomamos los 10 pares con mayor *z-score* APC en cada caso en vez de un umbral de 8. En este caso los *p-valores* son inferiores a 10^{-24}. Comprobamos también sistemáticamente cuanto afecta la definición de contacto y de coevolución al resultado. Queda patente que el resultado es robusto a la distancia usada como umbral de contacto y al umbral del *z-score* APC (Figura 4.4). Se puede observar *p-valores* aún mucho menores para umbrales del *z-score* APC más bajos. Esto no es sorprendente, ya que al ser menos estricto con la definición de coevolución se está incrementando el tamaño de muestra lo que podría llevar, por sí solo, a una mayor significancia estadística. Sin embargo, cabe notar que esto significa que nuestra medida de coevolución es informativa tanto sobre contactos físicos como la conservación estructural de éstos para valores menores de corte del *z-score* APC. Esto no conlleva necesariamente que sean predicciones de contactos lo suficientemente buenas como para que puedan ser útiles en la práctica, aunque es probable que así sea si se encuentra una forma adecuada de aprovechar esta información.

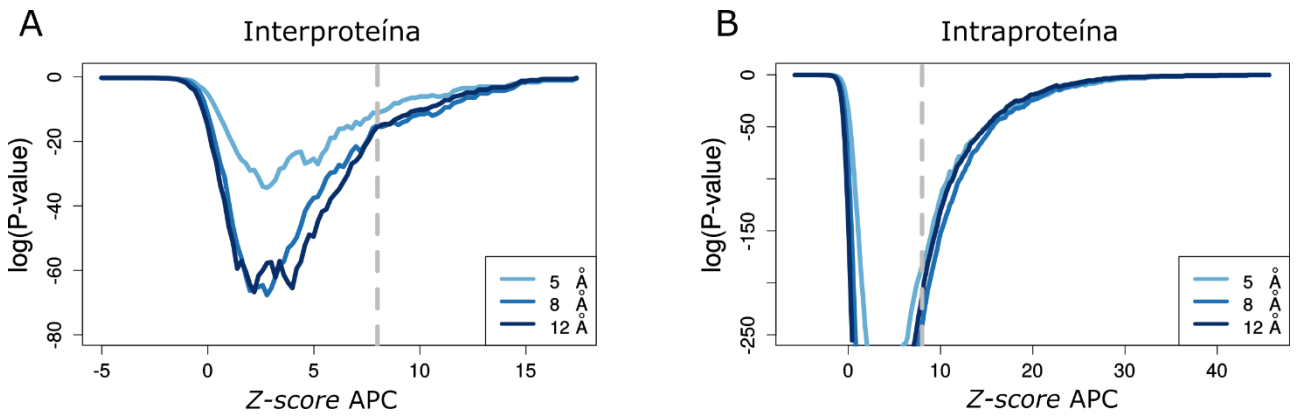


Figura 4.4 Significancia estadística y robustez de la relación entre conservación estructural y coevolución (adaptada de [117]). A) En el eje de ordenadas, el *p-valor* en escala logarítmica para los contactos conservados que han coevolucionado en la prueba exacta de Fisher para las categorías contacto conservado/no conservado y ha coevolucionado/no ha coevolucionado. En el eje de abscisas, el valor del *z-score* APC utilizado como umbral para definir que pares de posiciones han coevolucionado. La línea punteada vertical en gris indica el valor de referencia elegido para un *z-score* APC de 8. Las distintas tonalidades en azul muestran la definición de contacto utilizada, pudiendo ser 5, 8 y 12 angstroms de distancia. Valores inferiores a ~10^{-250} no son alcanzables por la limitación de precisión numérica.

La conservación estructural de los contactos esperable es dependiente de la conservación en secuencia de las posiciones consideradas. Por lo tanto, comprobamos como podría afectar este factor a nuestros resultados. En primer lugar, comparamos las distribuciones de entropías, como medida de conservación en secuencia, de todas las posiciones en interfaces con respecto a aquellas posiciones que han coevolucionado. A pesar de ser distribuciones relativamente similares en general, se observa algunas diferencias, en particular para posiciones con baja entropía, esto es, conservadas en secuencia (Figura A4A, Anexo I). Mientras que éstas son habituales en interfaces, prácticamente no aparecen en posiciones que han coevolucionado. Esto es esperable, dado que para que exista covariación ha de existir, obviamente, variación. Por otro lado, las posiciones que han coevolucionado se encuentran ligeramente enriquecidas con respecto a lo esperable en posiciones de conservación intermedia, con entropías en el intervalo (0.3, 0.7). Tras considerar la distribución de entropías (Materiales y Métodos, sección 3.9), se obtienen *p-valores* inferiores 10^{-10}, tanto para interproteína como intraproteína, para la prueba de Fisher de una cola del número de contactos conservados que han coevolucionado. De hecho, la probabilidad de conservación de contactos corregida por la distribución de entropías permanece prácticamente inalterada con respecto a la original (Figura A4B, Anexo I). Por lo tanto, la asociación entre coevolución y conservación de contactos no puede ser explicada por la distribución de entropías de las posiciones que han coevolucionado.

Dado que la identidad en secuencia puede ser una medida menos fiable en casos de divergencia extrema, calculamos las distancias evolutivas utilizando árboles para obtener una mejor cuantificación de la divergencia en secuencia y comprobar el posible efecto sobre las observaciones presentadas anteriormente. Para ello, tomamos las secuencias de los complejos representativos y reconstruimos sus árboles filogenéticos a partir de un conjunto representativo de genomas (Materiales y Métodos, sección 3.8). Calculamos la distancia evolutiva midiendo la distancia en el árbol entre el complejo representativo en procariotas y su homólogo en eucariotas (Materiales y Métodos, sección 3.6.1). La identidad en secuencia y la distancia

evolutiva correlacionan fuertemente (correlación de Spearman de 0.8, Figura A5E del Anexo I), y mantienen tendencias muy similares con respecto a la conservación estructural de la interfaz (Figura A5E).

Los resultados presentados en esta sección demuestran que los contactos físicos asociados a posiciones que han coevolucionado están altamente conservados estructuralmente, a pesar de las notables divergencias evolutivas existentes. Las señales de coevolución, por tanto, permiten identificar contactos físicos y, en general, regiones estructuralmente conservadas a largas escalas evolutivas, reforzando la idea de que la conservación estructural está asociada a un proceso coevolutivo. Además, esta propiedad puede permitir la predicción de contactos para complejos eucariotas mediante las señales encontradas en procariotas, como analizamos en la siguiente sección.

4.1.4 Predicción de contactos en eucariotas mediante señales en procariotas

Como se discutió anteriormente (sección 1.3.6), la predicción de contactos con alta precisión mediante covariación no es, actualmente, factible de forma sistemática en eucariotas. La alta precisión en la predicción de contactos en procariotas, unidad a la conservación estructural de los contactos predichos, permite proponer una estrategia para extrapolar predicciones de procariotas a eucariotas cuando exista homología. Es decir, proyectamos las señales de coevolución entre pares de posiciones en procariotas a las posiciones en eucariotas con las que alinean. Todo ello a pesar de las enormes distancias evolutivas que las separan y el bajo parecido de secuencia. En esta sección, evaluamos en detalle la calidad de las predicciones de contactos en eucariotas obtenidas de esta forma.

En primer lugar, analizamos la predicción de contactos para casos con información estructural tanto en procariotas como en eucariotas. Encontramos 19 interacciones interproteína y 120 intraproteína con predicciones de contacto por encima de umbral del *z-score* APC y cuyas predicciones tiene distancias definidas (ambas posiciones están mapeadas a residuos resueltos en alguna estructura de PDB). La gran mayoría de estos casos tiene una alta precisión tanto en eucariotas como en procariotas (Figura A6, Anexo I). Solo un caso de los 19 interproteína está predicho con una precisión menor a 0.6, por 6 de los 120 intraproteína. En eucariotas, estos números son solo ligeramente superiores con 2 casos de los 19 interproteína y 11 de los 120 para intraproteína. Utilizando interfaces representativas existen algunos pocos casos más con precisión menor que 0.6 distribuidos de manera uniforme en procariotas y eucariotas, sugiriendo que esta menor precisión no está relacionada con la proyección desde procariotas a eucariotas (Figura A6, Anexo I). La mayor parte de falsos positivos ocurre en casos muy divergentes en secuencia y con una baja conservación estructural (Figura A5A-D, Anexo I), que podría conllevar alineamientos de baja calidad en eucariotas. Por ello, evaluamos el impacto de la calidad de los alineamientos en las proyecciones desde procariotas a eucariotas mediante una estimación de la calidad media del alineamiento por posición (Materiales y Métodos, sección 3.7). De hecho, la mayor parte de casos con baja precisión en la predicción

de contactos en eucariotas, pero alta en procariotas, se corresponden con alineamientos de baja calidad tanto para interfaces completas como representativas (Figura A7, Anexo I).

Finalmente cuantificamos la precisión en eucariotas de las predicciones de contactos en todos los casos con información estructural en eucariotas (Tabla A1, Anexo I). Detectamos 62 pares de posiciones que han coevolucionado en 22 casos interproteína y 1 140 pares en 124 casos intraproteína. La precisión de estas predicciones de contacto es muy alta tanto en interproteína, con una precisión de 0.81 (Figura 4.5A), como en intraproteína, con una precisión de 0.95 (Figura 4.5B). La precisión en procariotas es solo algo más alta, con 0.86 y 0.95 respectivamente (Figura 4.5C y D). Repetimos este análisis considerando la calidad de los alineamientos y, en línea con la discusión previa, es posible identificar aquellos casos donde la precisión de las predicciones en eucariotas se pueda ver comprometida por la calidad del alineamiento (Figura A7E y F y Tabla A1 del Anexo I).

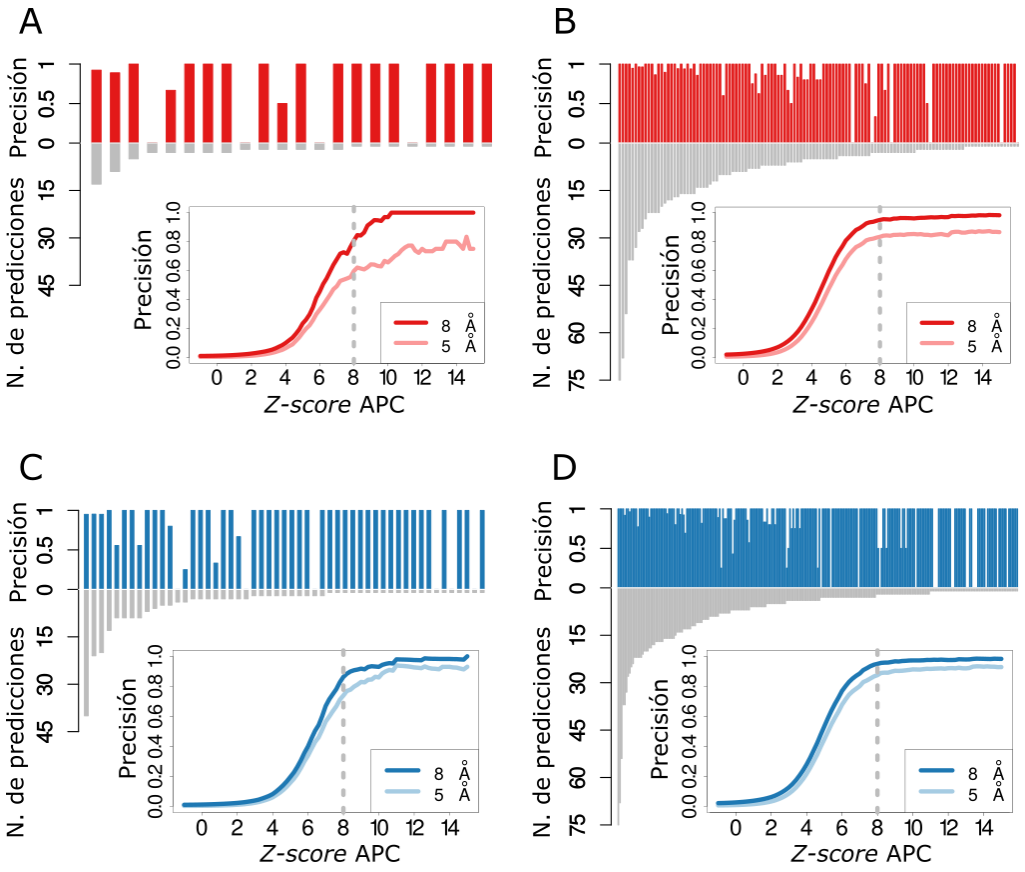


Figura 4.5 Precisión y número de predicciones para cada uno de los 22 casos interproteína (A) y 124 intraproteína (B) de interacciones entre dominios predichas en eucariotas (adaptada de [6]). Para procariotas se muestra la misma información para 53 casos interproteína (C) y 245 interproteína (D). Se consideran como predicciones todos los pares de posiciones con un *z-score* APC superior a 8. El umbral de distancia para la definición de contacto es de 8 angstroms. En las gráficas interiores se muestra la precisión de las predicciones en función del umbral del *z-score* APC utilizando un umbral de contacto de 8 angstroms (color oscuro) o de 5 angstroms (color claro). La línea punteada vertical en gris indica el valor de referencia elegido para un *z-score* APC de 8.

Cabe mencionar que el número de predicciones de contacto por caso es bajo. En intraproteína, hay 7.13 predicciones por caso en promedio para procariotas y 9.19 en eucariotas. En interproteína, la media es de 4.26 predicciones por caso en procariotas y 2.8 en eucariotas. Aunque es cierto que un número pequeño predicciones de contactos fiables son suficientes para generar buenos modelos tridimensionales de las interacciones [144,145], es probable que la calidad de los modelos resultantes tiendan a ser peores en promedio con un menor número de predicciones debido a una mayor incertidumbre en el modo de la interacción, en especial cuando solo se tienen una o dos predicciones. En este sentido, un 40% de los casos interproteína predichos en procariotas tienen solo 1 predicciones de contacto y un 19% tienen solo 2. En eucariotas, hay 36% casos interproteína predichos que tienen solo 1 predicciones y 27% que tienen solo 2.

Los resultados presentados en esta sección demuestran que es posible obtener predicciones de contacto fiables en eucariotas a partir de señales de coevolución en procariotas. Esta aproximación tiene una limitación importante, ya que su dependencia de homologías con procariotas la restringe a complejos evolutivamente muy antiguos. Sin embargo, estimamos que hasta un 17% del interactoma humano ambos interactores tienen homología y suficientes secuencias en procariotas (Figura A2, Anexo I; Materiales y Métodos). Y para la mayor parte de estas interacciones (un 15% del interactoma humano), no hay información estructural disponible (Figura A2, Anexo I). Por lo que, a pesar de las limitaciones, esta aproximación podría ser de utilidad en un número sustancial de interacciones en eucariotas.

4.1.5 Selección de ejemplos ilustrativos

A continuación, revisaremos algunas interacciones entre proteínas concretas ilustrativas sobre los resultados de nuestra aproximación.

El complejo piruvato deshidrogenasa, responsable de la catálisis del piruvato en acetyl-CoA y CO_2, es el complejo con un mayor acoplamiento en la interfaz de todo el conjunto de datos, esto es, es el complejo donde detectamos las señales más fuertes de coevolución. Su componente E1 forma un homodímero de heterodímeros de sus subunidades α y β [209]. Los contactos detectados por nuestro protocolo están distribuidos por toda la interfaz entre las dos subunidades y están bien conservados en la interfaz eucariota. De los 13 pares de posiciones que han coevolucionado, 10 están en contacto en la interfaz entre las subunidades del complejo deshidrogenasa de alfa-cetoácidos de cadena ramificada en *Thermus thermophilus* (Figura 4.6A). Los 10 pares de posiciones correspondientes en el complejo piruvato deshidrogenasa en humanos se encuentran también en contacto (Figura 4.6B), por lo que son contactos conservados. 2 de estos 3 aparentemente falso positivos corresponden, en realidad, a contactos en la interfaz del homodímero. Estos resultados muestran que la coevolución ha sido clave en la conservación de la estructura cuaternaria durante la evolución del componente E1 del piruvato deshidrogenasa.

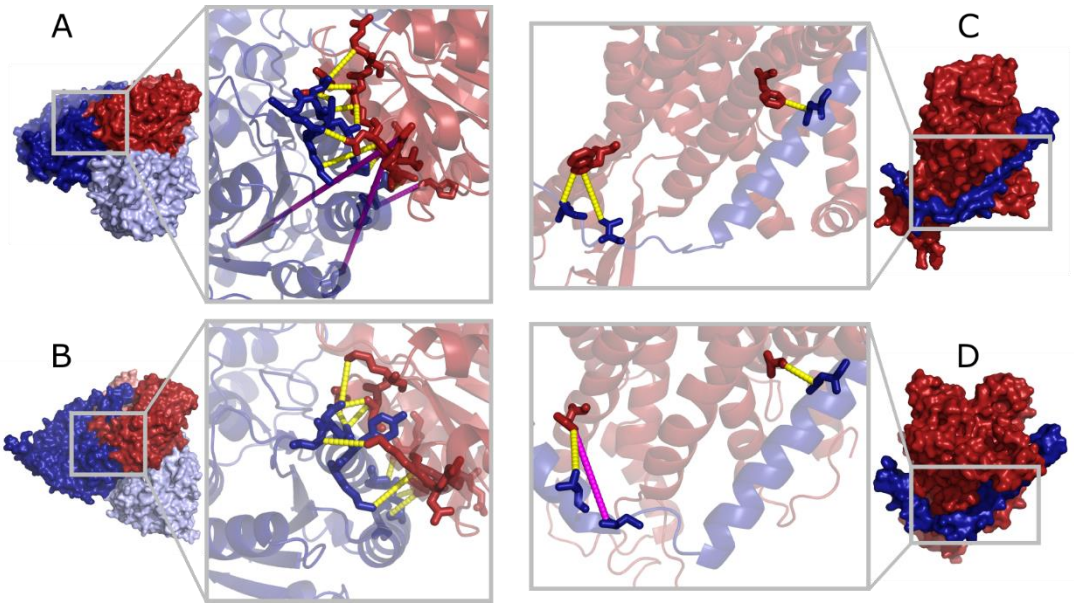


Figura 4.6 Ejemplos de contactos conservados. (A y B) Estructura del componente E1 del complejo deshidrogenasa de alfa-cetoácidos de cadena ramificada en *Thermus thermophilus* (panel A, código PDB: 1UMB [210]) y del complejo piruvato deshidrogenasa en humanos (panel B, código PDB: 3EXI [209]). (C y D) Se muestra las proteínas SecY (rojo) y SecE (azul) del complejo de transporte de proteínas a través de membrana SecYEG en *Thermus thermophilus* (panel C, código PDB: 2ZJS [211]) y las subunidades α (rojo) y γ (azul) del complejo homólogo Sec61 en eucariotas (panel D, código PDB: 4CG7 [212]). En las regiones ampliadas se pueden ver los aminoácidos de las posiciones que han coevolucionado como bastones (*sticks*) y los pares predichos como contactos conectados con líneas amarillas si están en contacto o en magenta en caso contrario.

El complejo traslocón, uno de los principales mecanismos del transporte de proteínas a través de membranas, es un buen ejemplo de la conservación del modo de interacción en casos de divergencia extrema. Las subunidades α y γ del complejo Sec61 en mamíferos son homólogos de las proteínas bacterianas SecY y SecE, respectivamente. A pesar de la baja identidad en secuencia entre estas proteínas en *Thermus thermophilus* y en *Canis lupus* (18.8% SecY/Sec61α y 10.5% SecE/Sec61γ), y la gran divergencia estructural entre los dominios, 2 de los 3 contactos que han coevolucionado se han conservado (Figura 4.6C y D). De hecho, 7 de los 9 pares de posiciones con mayor *z-score* APC están en contacto en *Thermus thermopilus* y estructuralmente conservados en *Canis lupus*. La resolución del complejo en *Canis lupus* (6.8 angstroms, obtenida mediante criomicroscopía electrónica) introduce alguna incertidumbre sobre la definición de la interfaz. En cualquier caso, nuestras predicciones apoyan la disposicional espacial global de la interacción observada en esta estructura experimental y subraya el potencial de la combinación de nuestra aproximación con experimentos de criomicroscopía electrónica.

Entre los 20 casos de interfaces interproteína con fuertes señales de coevolución e información estructural en procariotas y eucariotas, solo detectamos un caso donde fuertes señales de coevolución no están asociadas con una conservación de la interfaz: la interacción entre las subunidades α y β de la fenilalanina ARNt sintetasa (PheRS). La PheRS cataliza la unión del aminoácido fenilalanina con su correspondiente molécula de ARN de transferencia. A pesar de las importantes diferencias entre complejos PheRS en procariotas y procariotas [213], se pueden encontrar varios dominios homólogos tanto en la subunidad α

(dominio con el centro catalítico) como en la subunidad β (dominios B5 y B3/4) entre procariotas y eucariotas (Figura A8A y B, Anexo I). El análisis coevolutivo de la interacción entre el dominio con el centro catalítico y el dominio B3/4 revela dos pares de posiciones que han coevolucionado en la interfaz en *Thermus thermopilus* (Figura A8C, Anexo I). Estas posiciones no alinean en el complejo citosólico en humanos debido a una inserción en la PheRS en *Thermus thermopilus* en comparación con el complejo en humanos, como desvela un análisis estructural (Figura A8C y D, Anexo I). Por lo tanto, no es posible proyectar estas predicciones al complejo en humanos. Sin embargo, mientras que en esta inserción se produce la interacción en *Thermus thermopilus,* en humanos la región de interacción incluye, precisamente, uno de los dos giros donde se encuentra la inserción (Figura A8D). Pese a este cambio relevante en la región de interacción en dominio B3/4, las predicciones son correctas en la interfaz del otro dominio implicado en la interacción (Figura A8E y F, Anexo I). Por otra parte, la subunidad α también interacciones con el dominio B5 de la subunidad β y los 3 contactos que han coevolucionado están completamente conservados en la PheRS en humanos (Figura A8G y H, Anexo I). Este ejemplo ilustra como incluso ante cambios drásticos, como la sustitución completa de una región de interacción en uno de los interactores, el resto de las posiciones que han coevolucionado puede seguir apuntando a interfaces reales.

Entre los casos interproteínas, la peor predicción corresponde a la interacción entre dominios de la topoisomerasa de tipo IIA (girasa o toposimerasa IV en bacterias y toposimerasa II en eucariotas), complejo enzimático encargado de relajar el ADN superenrollado, así como de introducir superenrollamientos tanto negativos como positivos. En la predicción de la interacción entre la subunidad A (identificador de Pfam DNA_topoisoIV) y la subunidad B (segundo dominio, identificador de Pfam DNA_gyraseB) encontramos 4 pares de residuos con un *z-score* APC superior al umbral, siendo todos ellos falsos positivos. Analizamos en más detalle la posible influencia de las relaciones filogenéticas con estos falsos positivos utilizando los árboles reconstruidos para cada uno de los dos dominios Pfam (Materiales y Métodos, sección 3.8). Como muestra la Figura 4.7A y B, los árboles de los dos dominios muestran conjuntos bien diferenciados de secuencias y grupos de secuencias muy próximos (ramas cortas cerca de las hojas). Esto también se observa en los alineamientos emparejados que utilizamos para computar el modelo, como muestra el espacio definido por las primeras componentes de un análisis de componentes principales sobre las distancias entre las secuencias del alineamiento emparejado (Figura 4.7C, sección 3.11 de Materiales y Métodos). Cabe destacar que solo la primera componente es capaz de explicar el 66% de la variabilidad del alineamiento, indicativo de una marcada estructura filogenética. Como discutimos en la introducción (sección 1.3.3.3), las relaciones filogenéticas entre secuencias pueden generar covariaciones que no están relacionadas con la estructura o función de las interacciones entre proteínas. Estos resultados sugieren que los falsos positivos obtenidos en este caso podrían deberse a covariaciones fuertes debidas a la filogenia subyacente.

Figura 4.7 Árboles de los dominios de la topoisomerasa de tipo IIA. Árboles asociados al dominio Pfam DNA_topoisoIV de la subunidad A (A) y al dominio Pfam DNA_gyraseB de la subunidad B (B). C) Distancias entre las secuencias del alineamiento emparejado de la interacción entre el dominio Pfam DNA_topoisoIV y el dominio DNA_gyraseB sobre el espacio definido por las dos primeras componentes principales de un análisis de componentes principales de las distancias entre las secuencias del alineamiento emparejado.

Los métodos coevolutivos también pueden habilitar la reconstrucción total o parcial de complejos multiproteína [144,145]. Observamos una fuerte coevolución en algunas de las interacciones entre proteínas en el complejo NADH deshidrogenasa, complejo que cataliza la transferencia de electrones desde la coenzima NADH a la coenzima Q10 y lleva a cabo la traslocación de protones a través de la membrana. En *Escherichia Coli,* encontramos 3 predicciones por encima del umbral en la interacción entre las proteínas de los genes nuoJ y nuoK, así como 2 posiciones entre los genes nuoK y nuoA (Tabla 4.1 y Figura 4.8). Todas estas predicciones se corresponden con contactos físicos en la estructura (Tabla 4.1 y Figura 4.8). También se observan contactos bien predichos en la parte superior del ranking en la interacción entre nuoN y nuoK, aunque los *z-scores* APC no son suficientemente altos para estar por encima del umbral de *z-score* APC (Tabla 4.1 y Figura 4.8). Este ejemplo muestra que el umbral del *z-score* APC no es idóneo. Por un lado, no se predice la interacción entre nuoN y nuoK (no hay predicciones por encima del umbral) a pesar de contener buenas predicciones, por otro, solo se consideran 3 predicciones de contacto entre nuoJ y nuoK cuando existen muchas otras predicciones correctas en este caso (Tabla 4.1). En la siguiente sección desarrollaremos una nueva metodología, con el propósito de mejorar esta situación, que permite el uso de umbrales o correcciones específicas de cada caso.

Genes	nuoJ nuoK		nuoJ nuoA		nuoN nuoK	
Dominios Pfam	Oxidored_q3 Oxidored_q2		Oxidored_q3 Oxidored_q4		Proton_antipo_M Oxidored_q2	
Índice	Distancia	*z-score* APC	Distancia	*z-score* APC	Distancia	*z-score* APC
1	**3.9**	**15.41**	**4.2**	**8.27**	38.4	6.91
2	**4.4**	**9.01**	**6.4**	**8.10**	**7.7**	5.84
3	**4.3**	**8.79**	12.8	6.66	**6.6**	5.28
4	**3.4***	7.99	48.7	6.12	**4.4**	5.25
5	**4.0**	7.39	19.9	5.90	17.1	5.24
6	**3.6**	7.13	16.6	5.82	**7.4**	5.07
7	**3.4**	6.94	**5.9**	5.13	31.9	5.05
8	**3.4**	6.63	36.11*	5.05	30.7	4.94
9	28.9	6.51	**6.0**	4.96	14.8	4.87
10	**2.7**	6.13	35.0	4.92	36.5	4.78

Tabla 4.1 Predicciones en la NADH deshidrogenasa. Se muestran la distancia y el *z-score* APC para las 10 primeras predicciones de contacto de las interacciones entre las proteínas de los genes nuoJ y nuoK (dominios Pfam Oxidored_q3 y Oxidored_q2); nuoJ y nuoA (Oxidored_q3 y Oxidored_q4); nuoN y nuoK (Proton_antipo_M y Oxidored_q2). Las distancias se han extraído de la estructura 3RKO [214] de la NADH deshidrogena en *Escherichia Coli*. En negrita, se remarcan las distancias por debajo del umbral de contacto (≤8 Å) y los *z-scores* APC por encima del umbral de referencia (≥8). *Las distancias anotadas con asterisco se han extraído de la estructura 4HEA [215] debido a que alguna de las posiciones a esta predicción no se puede mapear a la secuencia de la estructura 3RKO, por lo que no se puede asignar una distancia entre el par de posiciones.

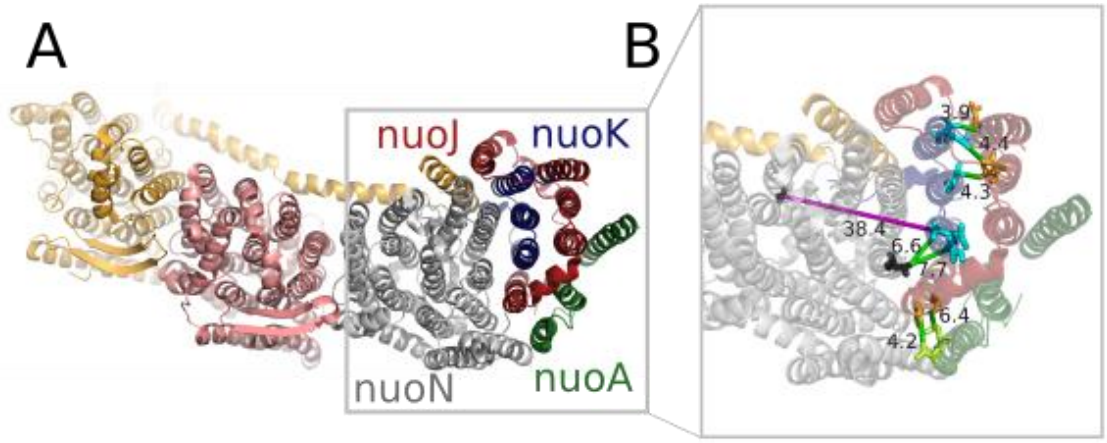


Figura 4.8 Predicciones en la NADH deshidrogenasa. A) Estructura de la parte membranal de la NADH deshidrogenasa en *Escherichia Coli* (código PDB 3RKO). B) Se muestran como bastones los residuos que han coevolucionado (Tabla 4.1) en la interacción entre nuoJ (rojo) y nuoK (azul), así como en la interacción entre nuoA (verde) y nuoJ (rojo). También se muestran como bastones los residuos pertenecientes a las 3 primeras predicciones de contacto en la interacción entre nuoK y NuoN (Tabla 4.1). Las predicciones de contacto están unidas por una línea verde cuando la distancia es inferior al umbral de contacto (8 Å) y magenta en caso contrario. Para cada predicción de contacto, aparece anotada la distancia en angstroms.

4.2 Estimación y corrección de la distribución de fondo

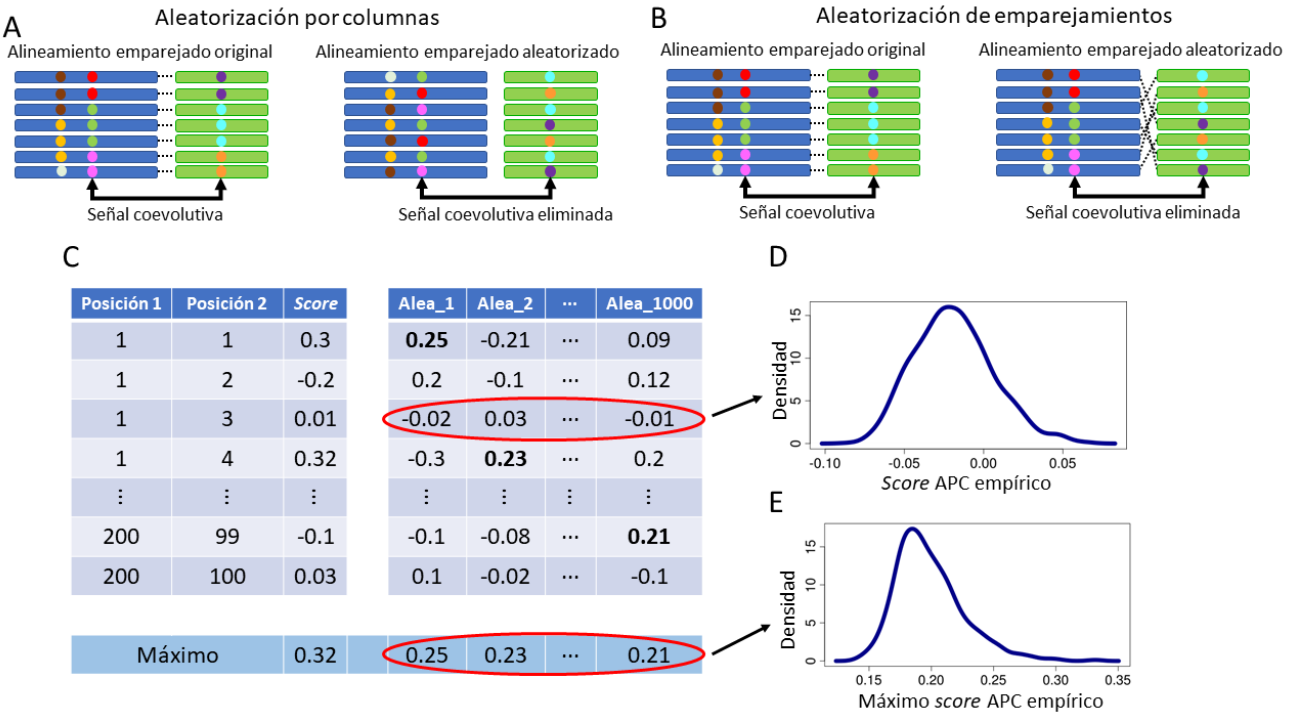
En esta sección estudiaremos como afectan diversos factores a las predicciones de contacto entre proteínas, proponiendo mejoras metodológicas y midiendo su impacto en la calidad de las predicciones de contactos. Los resultados de esta sección se presentarán en el siguiente orden: i) introducción a la aleatorizaciones utilizados con el objetivo de estimar una distribución de fondo (*background*) específica para cada caso y su relación con la corrección APC, ii) aplicación de una corrección específica basada en la estimación del *background* estimado mediante las aleatorizaciones y comparativa de la calidad de las predicciones resultantes, iii) análisis de la influencia del número de secuencias y vi) un estudio más detallado de una selección de casos concretos como ejemplos ilustrativos.

4.2.1 Aleatorizaciones para la estimación de la distribución de fondo

A modo de recordatorio (ver sección 1.3.4.1 de la Introducción o la sección 3.4.3 de Materiales y Métodos para más detalles), aplicamos el modelo estadístico DCA sobre cada alineamiento emparejado. De los parámetros interdominio de este modelo coevolutivo extraemos los *scores* crudos, como medida de la dependencia estadística entre las posiciones interdominio del alineamiento. A partir de los *scores* crudos aplicamos la corrección APC para obtener los *scores* APC. Como se explicó en la introducción (sección 1.3.3.3), APC es una medida del *background* para cada par posiciones que se sustrae de los *scores* crudos, mejorando sistemáticamente las predicciones. Las distribuciones de *scores* APC se pueden estandarizar mediante la aplicación de la desviación absoluta mediana, para dar lugar a los *z-scores* APC.

Como apuntamos en la introducción (sección 1.3.6), diversos factores como el número de secuencias y de posiciones del alineamiento, la conservación de las posiciones y las relaciones filogenéticas entre las secuencias afectan a la detección de señales coevolutivas entre proteínas. Es esperable que distintas familias de proteínas e interacciones presenten distintas combinaciones de estos factores que a su vez influirán de modo distinto en los resultados de los métodos coevolutivos. De esto se deduce que podría ser beneficioso en el proceso de predicción considerar y corregir el *background,* creado por estos factores, específico de cada caso. La diferencia con la corrección APC radica en que APC es una corrección para cada par de posiciones mientras que en este estudio queremos estimar no solo el *background* para cada par de posiciones, si no el *background* de cada caso en su conjunto (considerando todos los pares de posiciones interdominio). Para estudiar este asunto construimos un conjunto de datos de 490 Interacciones Dominio a Dominio Interproteína (IDDI) de forma similar al anterior estudio (Materiales y Métodos, sección 3.1.1). Las principales diferencias son que incluimos alineamientos pequeños, 40 secuencias como mínimo en vez de 500, para poder estudiar casos con pocas secuencias, y que nos enfocamos solo en casos de interacciones interproteína, más relevantes biológicamente, y con estructura en procariotas para evitar posibles problemas con la calidad de los alineamientos.

Con el objetivo de intentar estimar el *background* de cada caso, realizamos dos tipos de aleatorizaciones para obtener distribuciones empíricas que aproximan a la distribución nula: aleatorización por columnas y aleatorización de emparejamientos (Figura 4.9). Para referirnos sucintamente a estas distribuciones empíricas, utilizaremos el término *distribuciones empíricas nulas*. Como se puede observar en la Figura 4.9A, en la aleatorización por columnas se reordenan aleatoriamente los aminoácidos de cada columna del alineamiento por separado. Cualquier covariación existente en alineamiento original desaparece. Los *scores* obtenidos con estos alineamientos se corresponden con las correlaciones espurias que puedan aparecer al azar según las características del alineamiento. En la aleatorización de emparejamientos (Figura 4.9B), se reordenan aleatoriamente las secuencias del segundo dominio. En estas aleatorizaciones se mantienen todas las señales intradominio al tiempo que desaparecen todas las interdominio, por lo que se incluyen las covariaciones y divergencias evolutivas existentes en el alineamiento de cada dominio. Tras realizar estas aleatorizaciones para los 490 alineamientos emparejados, computamos el modelo coevolutivo y los distintos *scores*, repitiendo el proceso 1 000 veces. En este experimento computamos un total de 980 000 modelos, 490 000 con aleatorizaciones por columnas y otros 490 000 con aleatorizaciones de emparejamientos. Estas aleatorizaciones nos permiten obtener distribuciones empíricas, por lo que nos referiremos a los *scores* obtenidos en estas aleatorizaciones como *scores* empíricos (Figura 4.9C-E). En particular llamaremos *máximos empíricos* a los máximo *scores* APC obtenidos en cada aleatorización (Figura 4.9E).



Posición 1	Posición 2	*Score*	Alea_1	Alea_2	···	Alea_1000
1	1	0.3	**0.25**	-0.21	···	0.09
1	2	-0.2	0.2	-0.1	···	0.12
1	3	0.01	-0.02	0.03	···	-0.01
1	4	0.32	-0.3	**0.23**	···	0.2
⋮	⋮	⋮	⋮	⋮	⋮	⋮
200	99	-0.1	-0.1	-0.08	···	**0.21**
200	100	0.03	0.1	-0.02	···	-0.1
Máximo		0.32	0.25	0.23	···	0.21

Figura 4.9 Esquema de las aleatorizaciones realizadas. A) Aleatorizaciones por columnas. En esta aleatorización los aminoácidos de cada columna se reordenan aleatoriamente. B) Aleatorización de emparejamientos. En este caso, se reordenan aleatoriamente las secuencias del segundo dominio. C) Tabla que muestra un ejemplo de los *scores* APC para dos dominios, un primero de 200 posiciones y un segundo de 100 posiciones. D) Ejemplo de la distribución empírica nula de 1000 *scores* APC para un par de posiciones. Cabe notar que en la comparación con APC se utilizan la distribución empírica nula de *scores* crudos (no *scores* APC). E) Ejemplo de una distribución empírica nula de 1000 *scores* APC máximos, que llamamos máximos empíricos.

Como se explicó en la introducción (sección 1.3.4.1), no se comprende con exactitud el motivo por el cuál APC permite mejorar las predicciones en el contexto de DCA. Para arrojar luz sobre este asunto, comparamos el valor de APC con las distribuciones de los *scores* empíricos crudos para cada par de posiciones. Recordemos que $score_{APC}(i,j) = score_{crudo}(i,j) - APC(i,j)$, donde i es la columna del primer domino y j la del segundo. Encontramos que la media de los 1000 *scores* crudos empíricos es muy similar al valor de APC. La Figura 4.10A muestra un ejemplo para un par de posiciones, mientras que la Figura 4.10B muestra una comparación para todos los pares de posiciones de un IDDI típico. La correlación entre ambas cantidades es muy alta para los 490 casos de todo el conjunto de datos (correlación de Pearson media=0.986; Figura 4.10C). La coherencia entre estas dos medidas obtenidas de forma completamente distinta apunta a una estimación exacta de la media de la distribución nula. Esto proporciona una interpretación clara sobre APC en el contexto de DCA, se trata de una buena estimación de la media de la distribución nula, consistente con la interpretación de APC en el contexto de la información mutua [79].

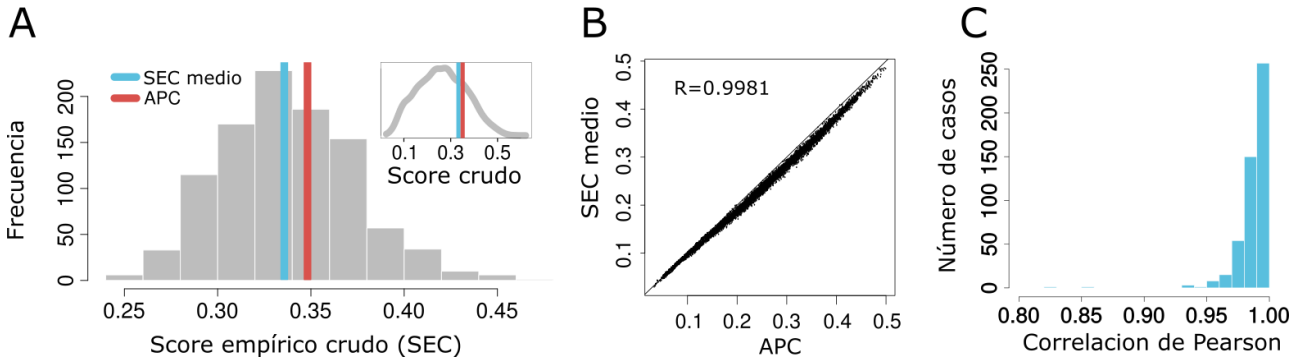


Figura 4.10 Relación entre APC y *scores* empíricos crudos. A) Histograma de los 1000 *Scores* Empíricos Crudos (SECs) en gris, su valor medio en azul y el valor de APC en rojo para el par de posiciones 32 y 12 (siguiendo la numeración de Pfam) de la interacción entre los dominios Rpr2 y UPF0086 (identificadores de Pfam) de la ribonucleoproteína ribonucleasa. Gráfico interno: densidades de la distribución de *scores* crudos para todas las posiciones del modelo coevolutivo original (emparejamientos correctos). B) SEC medio para cada par de posiciones en función del valor de APC, la línea negra diagonal corresponde con la línea de igualdad. La correlación de Pearson es 0.9981. C) Histograma de la correlación de Pearson entre APC y SEC medio para cada uno de los 490 IDDIs bajo estudio.

Considerando que podemos hacer una buena estimación del *background* para cada par de posiciones, razonamos que el máximo de la distribución empírica de *scores* APC puede ser una buena estimación del nivel de ruido esperable en cada caso cuando se ha destruido la señal coevolutiva. De forma que se considera la amplitud de dicha distribución que puede ser indicativa de las diferencias entre casos. Para cada uno de los 490 casos del conjunto de datos, tenemos dos distribuciones de 1000 *scores* APC máximos empíricos (ver ejemplo en Figura 4.9E), una para la aleatorización columnas y otra para la aleatorización de emparejamientos. Los máximos empíricos promedio obtenido con las aleatorizaciones de emparejamientos correlacionan (correlación de Pearson=0.47; *p-valor* de la correlación=$1.2*10^{-28}$) con los máximos *scores* APC originales (Figura 4.11A), esto es, el de los alineamientos emparejados originales sin aleatorizar. Para las aleatorizaciones de emparejamientos, observamos una correlación con los máximos *scores* APC originales muy similar (correlación=0.48, *p-valor*=$1.19*10^{-29}$). Los máximos empíricos promedio de las dos aleatorizaciones son muy parecidos (Figura A9, Anexo I), aunque el de las aleatorizaciones de

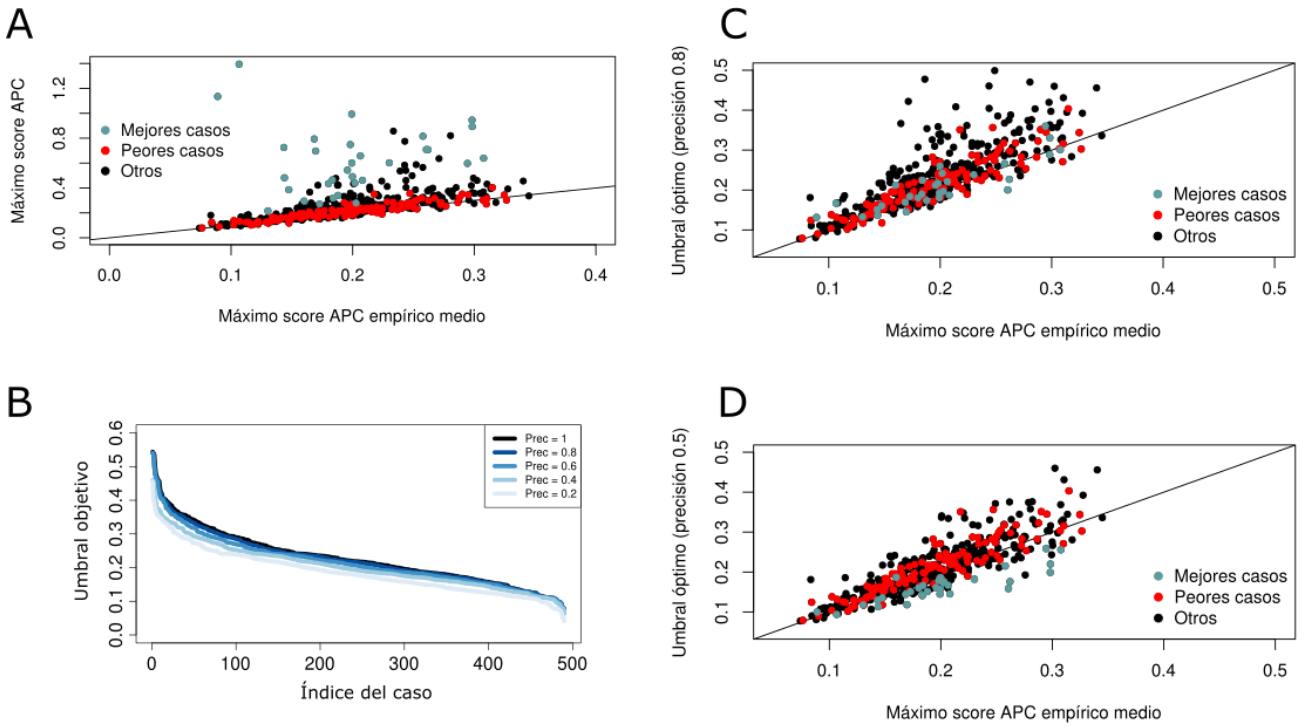


Figura 4.11 Correlaciones entre máximos empíricos y umbrales objetivo. A) En el eje de abscisas se muestra el máximo *score* APC empírico medio obtenido en las aleatorizaciones de emparejamientos. En el de ordenadas, el máximo *score* APC en el alineamiento original. En azul, los mejores casos, definidos como que aquellos cuyas 5 primeras predicciones son todas contactos. En rojo, se muestran los peores casos, definidos como aquellos que no tienen ninguna predicción correcta entre las L/10 primeras predicciones, donde L es el número de posiciones en el alineamiento. B) Los umbrales objetivo, en el eje de ordenadas, para cada uno de los 490 IDDIs, en eje de abscisas del conjunto de datos para diferentes niveles de precisión mostrados en tonos azules. C) El umbral objetivo para una precisión de 0.8 con respecto al máximo *score* APC empírico medio en las aleatorizaciones de emparejamientos. El código de colores es igual al del panel A. D) Idem. para una precisión de 0.5.

emparejamientos es ligeramente inferior de forma sistemática. Esto es probablemente debido a que la presencia de las señales intradominio (filogenéticos o coevolutivas) aumenta los *scores* intradominio y reduce los *scores* interdominio.

Los máximos empíricos promedio correlacionan especialmente bien (correlación=0.91 tanto para la aleatorización por columnas como de emparejamientos, *p-valor*<3.8*10^{-49}) con los máximos *scores* APC originales en los 126 casos peor predichos (pie de Figura 4.11), indicando que las aleatorizaciones reproducen el *score* APC que se obtiene en ausencia de señales de coevolución (Figura 4.11A). De hecho, no solo correlacionan fuertemente, sino que son muy próximos en valor. Por otro lado, los valores atípicos (*outliers*) se corresponden con 31 casos particularmente bien predichos (pie de Figura 4.11) y con muchas secuencias (Figura 4.11A). Los *scores* APC particularmente altos de estos casos son debidos a señales muy fuertes de coevolución, siendo muy distintos de los producidos por el *background,* y están asociados a contactos físicos entre las posiciones que coevolucionan.

La correlación entre máximo empírico promedio y máximo *score* APC sugiere que las aleatorizaciones proporcionan una buena estimación del *background*. Esta estimación puede servir de base para la obtención de un umbral del *score* específico para cada caso. Un umbral de este tipo sería de gran utilidad puesto que, los métodos actuales no son capaces de asignar un valor de confianza a los *scores* APC obtenidos para cada

caso (como se discute en la sección 1.3.6). Esto permitiría estimar un valor de confianza en función a la distancia del *score* APC a un umbral de este tipo. La alta correlación encontradas para los casos peor predichos muestra que las aleatorizaciones producen una buena estimación del *background* ya que el umbral que se debería utilizar ha de ser similar al del máximo *score* APC, habida cuenta de que las predicciones no son correctas en estos casos. Para el resto de los casos, es necesaria una mejor cuantificación del *background* que tenga en cuenta las predicciones que son correctas, ya que el umbral que se debería utilizar debería discriminar las predicciones correctas de las incorrectas.

Para cuantificar el *background*, definimos la medida *umbral objetivo.* El umbral objetivo sirve como estimación del nivel de *background* del *score* APC en cada caso. Se computa utilizando el ranking de predicciones ordenadas por el *score* APC para el caso en cuestión y dependen de una precisión objetivo dada como parámetro. En cada iteración, empezando desde la predicción con menor *score* APC, se evalúa la precisión si se utiliza como umbral el *score* APC de la predicción actual. Esto es, se contabiliza la proporción de predicciones correctas (que conocemos gracias a la información estructural) considerando aquellas con un *score* APC mayor que el *score* APC de la predicción actual. Si la precisión es superior a la precisión objetivo, el umbral objetivo es el *score* APC de la predicción actual. En caso de no alcanzarse, se descarta la predicción actual y se sube una posición en el ranking (tomándose la predicción con el *score* APC más próximo pero superior). En caso de recorrerse el ranking completo y no alcanzarse la precisión objetivo en ningún momento, el umbral objetivo es el mayor *score* APC del ranking. Las predicciones se consideran como positivas cuando el *score* APC es estrictamente mayor que el umbral utilizado. En consecuencia, los umbrales objetivo son idénticos al máximo *score* APC en los casos peor predichos e inferior a éste en los casos donde existen predicciones correctas en la parte superior del ranking.

Calculamos los umbrales objetivo para precisiones de 0.2, 0.4, 0.6, 0.8 y 1. Independientemente de la precisión requerida, los umbrales objetivo muestran una gran dependencia del caso (Figura 4.11B). Esto muestra que en nuestro conjunto de casos, que incluyen muchos con pocas secuencias, el uso de un umbral único para todos los casos [144,145] no está justificado y representa una estrategia inadecuada. En contraste, el máximo empírico promedio correlaciona fuertemente con el umbral objetivo de cada caso, con una correlación de 0.81 (*p-valor*=$3.8*10^{-116}$, precisión objetivo=0.8) tanto para las aleatorizaciones por columnas como la de emparejamientos (Figura 4.11C). Para una precisión objetivo de 0.5, las correlaciones son de 0.84 (*p-valor*=$1.7*10^{-133}$) para la aleatorización por columnas y 0.85 (*p-valor*=$7*10^{-139}$) para la de emparejamientos. De hecho, no solo correlacionan fuertemente con los umbrales objetivo si no que son muy próximos a su valor (Figura 4.11C y D).

La clara correlación entre los umbrales objetivo y la media de máximos empíricos muestra que las aleatorizaciones estiman convenientemente el *background*, lo cual debería permitir realizar correcciones específicas por caso. A continuación, exploraremos alternativas para explotar este hecho y mediremos la potencial mejora en la calidad de las predicciones.

4.2.2 Comparativa de la calidad de las predicciones

Estudiamos distintas estrategias para transformar los *scores* APC originales de forma que se corrija por el *background* específico de cada caso estimado mediante los máximos empíricos. Consideramos 4 estadísticos de las distribuciones de máximos empíricos: media, media más 2 desviaciones estándar, mediana y máximo. Tenemos dos distribuciones de máximos empíricos en cada caso, una proveniente de la aleatorización por columnas y una segunda proveniente de la aleatorización de emparejamientos. Consideramos también dos posibles operaciones para aplicar el estadístico sobre los scores APC originales: división y substracción. En total, consideramos 16 estrategias posibles dependiendo del estadístico utilizado, la aleatorización empleada y la operación aplicada. Para cada una de estas 16 estrategias por separado, transformamos los *scores* APC de los 490 casos dependiendo de la estrategia y de los máximos empíricos obtenidos para cada caso.

Las mejores estrategias son las que se usan la media y la mediana de los máximos empíricos de la aleatorización de emparejamientos (Figura A10A y B, Anexo I). Estas estrategias tienen un rendimiento muy similar, con diferencias numéricas muy pequeñas. Entre ellas, elegimos la estrategia que consiste en dividir los *scores* APC por la media de los máximos empíricos obtenidos mediante las aleatorizaciones de emparejamientos, ya que es la que más casos predice con una precisión de 0.8 o mejor (Figura A10C y D). Este *score*, que llamamos *score* MEND (del inglés *Maxima from Empirical Null Distributions*), o S_{MEND}, se calcula de la siguiente forma

$$S_{MEND}(i,j) = \frac{S_{APC}(i,j)}{\overline{max\_emp}}$$

donde $\overline{max\_emp}$ es la media de los máximos empíricos de la aleatorización de emparejamientos específico de cada caso y $S_{APC}(i,j)$ es el *score* APC entre las posiciones i y j.

Comparamos el *score* MEND con los *scores* crudos, *scores* APC y *z-scores* APC. El *score* MEND mejora ostensiblemente la calidad de las predicciones (Figura 4.12A, sección 3.10.1 de Materiales y Métodos). Predice 400 contactos con una precisión de 0.8, mientras *score* APC lo hace con una precisión de 0.65. El *score* APC solo predice 247 contactos con una precisión de 0.8 lo cual representa una mejora del 62% del *score* MEND con respecto del *score* APC en el número de contactos predicho a esta precisión. Usando una precisión menos restrictiva, el *score* MEND predice 1058 contactos con una precisión de 0.5 por 632 para el *score* APC, un incremento del 67%. Por otra parte, el rendimiento del *z-score* APC es algo inferior al del *score* APC mientras que el de los *scores* crudos es muy bajo (Figura 4.12A).

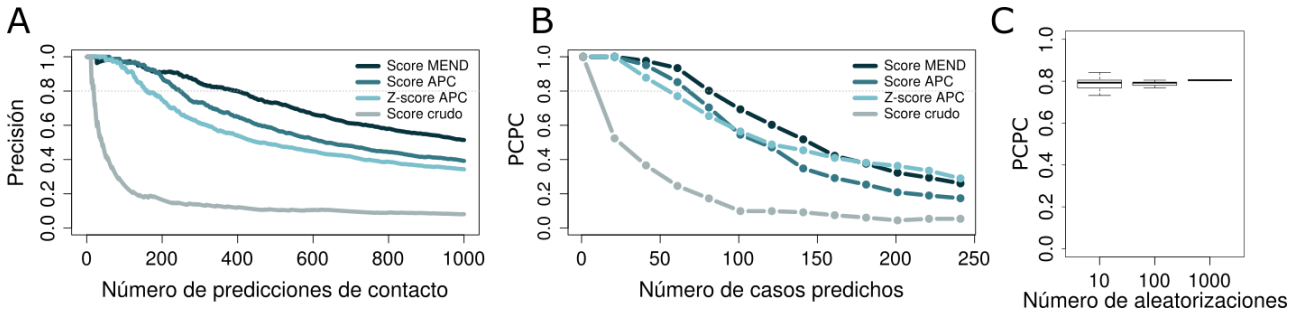


Figura 4.12 Calidad de las predicciones y número de aleatorizaciones necesarias. A) Precisión de las 1 000 primeras predicciones del *score* MEND, el *score* APC, el *z-score* APC y el *score* crudo. Se consideran conjuntamente las predicciones de los 490 casos, ordenadas de decreciente por el correspondiente *score*. B) La proporción de casos predichos correctamente (PCPC, con precisión ≥ 0.8), en función del número de casos predichos para el *score* MEND, el *score* APC, el *z-score* APC y el *score* crudo. C) Proporción de casos predichos correctamente (precisión ≥ 0.8) considerando 82 casos predichos dependiendo del número de aleatorizaciones utilizados para estimar el *score* MEND. La distribución para 10 aleatorizaciones cuenta con 100 estimaciones distintas, para 100 aleatorizaciones 10 estimaciones y para 1000, una.

A continuación, evaluamos el impacto en términos de aplicabilidad, esto es, cual es la mejora si nos fijamos en el número de casos predichos (Materiales y métodos, sección 3.10.2). Diremos que un caso ha sido predicho correctamente si su precisión es igual o mayor a un valor deseado. Utilizaremos precisiones de 0.8 y 0.5 para considerar un criterio estricto y uno más laxo. En términos de aplicabilidad, el *score* MEND es también notablemente superior al resto de *scores* (Figura 4.12B). Utilizando un MEND *score* de 1.496 como umbral, se predicen 82 casos, para los cuales un 80% de los casos está correctamente predicho para una precisión de 0.8 (97 para precisión ≥ 0.5). El umbral análogo para el *score* APC es 0.356 (80% de los casos predichos correctamente), pero solo 69 casos son predichos (71 para precisión ≥ 0.5). Estos 13 casos de diferencia representan un 19% (37% para precisión ≥ 0.5) de incremento con respecto a los *scores* APC en condiciones equivalentes. La Tabla 4.2 presenta una relación de posibles umbrales, donde se observa que la mejora se encuentra entre un 19% y un 37%. Esta tabla también contiene una lista de umbrales en función del balance de precisión y exhaustividad (*recall*) que se desee. En resumen, estos resultados muestran la capacidad del *score* MEND para detectar *a priori*, sin conocer la estructura, un mayor número de casos predichos correctamente.

PCPC	Número de IDDIs			Umbral		Precisión			Número predicciones		
	APC	MEND	Mej.	APC	MEND	APC	MEND	Dif.	APC	MEND	Mej.
Precisión≥0.8											
0.9	53	63	19%	0.3946	1.6169	0.96	0.89	-0.07	137	250	82%
0.8	69	82	19%	0.3563	1.4997	0.9	0.83	-0.07	188	315	67%
0.7	81	100	23%	0.33621	1.4041	0.84	0.81	-0.03	222	387	74%
0.6	96	124	29%	0.31883	1.3189	0.78	0.74	-0.04	258	468	81%
0.5	113	147	30%	0.29817	1.2601	0.69	0.68	-0.01	347	573	65%
0.25	185	247	33%	0.24973	1.1393	0.43	0.55	0.12	852	901	6%
Precisión≥0.5											
0.9	53	64	21%	0.3946	1.6119	0.96	0.89	-0.07	137	252	84%
0.8	71	97	37%	0.35093	1.4143	0.9	0.81	-0.09	194	377	94%
0.7	91	124	36%	0.3229	1.31	0.8	0.74	-0.06	250	477	91%
0.6	110	150	36%	0.30267	1.2584	0.71	0.68	-0.03	325	578	78%
0.5	152	184	21%	0.26921	1.2082	0.53	0.61	0.08	567	689	21%
0.25	281	376	34%	0.21141	1.0239	0.26	0.39	0.13	2274	1613	-29%

Tabla 4.2 Compilación de umbrales y rendimiento dependiendo de la proporción de casos predichos correctamente (PCPC) y usando un criterio más estricto (precisión≥0.8) y uno más exhaustivo (precisión≥0.5) en la definición de caso predicho correctamente. El número de IDDIs es el número de casos totales predichos con el umbral usado. El umbral define las predicciones que se consideran y el número de casos predichos (aquellos con alguna predicción con un *score* superior al umbral). La columna precisión se refiere a la precisión global en la predicción de contactos considerando todas las predicciones de contactos conjuntamente que tengan un *score* superior al umbral. Igualmente, la columna número predicciones se refiere al número global de predicciones de contacto, esto es, el número de predicciones que se encuentran por encima del umbral. Las columnas Mej. representan la mejora de MEND *score* con respecto al *score* APC, midiéndola (N_{MEND}-N_{APC})/N_{APC}*100, donde N_{MEND} es el número de casos o predicciones para el *score* MEND y N_{APC} es el número de casos o predicciones para el *score* APC. La columna Dif. muestra la diferencia entre la precisión del *score* MEND y la del *score* APC.

Como hemos mostrado el *score* MEND supone una mejora relevante sobre el *score* APC. Sin embargo, el hecho de utilizar 1 000 aleatorizaciones podría suponer un impedimento para su aplicación en algunas circunstancias. Por ello, evaluamos su dependencia del número de aleatorizaciones realizadas. Como muestra la Figura 4.12C, la calidad de las predicciones varía muy poco con la utilización de un menor número de aleatorizaciones y la variabilidad de éstas solo es ligeramente mayor con la utilización de un menor número de aleatorizaciones. Tomando solo 10 aleatorizaciones y una proporción de casos predichos correctamente del 80% como referencia (precisión ≥ 0.8), la proporción de casos predichos correctamente queda acotada entre un 73% y 84% en 100 estimaciones distintas. Por tanto, es suficiente usar 10 aleatorizaciones para obtener resultados buenos y robustos. A esto se une el hecho de que los modelos son rápidos de computar (Figura A11, Anexo I), por lo que es posible aplicar esta corrección a larga escala.

Como discutimos en la introducción (sección 1.3.3.3), estos métodos son susceptibles al efecto de las relaciones filogenéticas entre las secuencias, esto es, a que no son observaciones independientes lo cual puede generar covariaciones no relacionadas con la coevolución. Utilizaremos el término señal filogenética para referirnos al efecto que tiene sobre el modelo y los *scores* las relaciones filogenéticas entre las secuencias. Es bien sabido que las técnicas de reducción de dimensiones sobre alineamientos de secuencias

capturan en sus componentes principales la estructura filogenética de las familias de proteínas, ya sea analizando los aminoácidos [82,86] o las distancias entre las secuencias [216]. Por ello, utilizamos la variabilidad explicada por la primera componente principal, aplicando el análisis de componentes principales sobre las distancias entre las secuencias del alineamiento (Materiales y Métodos, sección 3.11), como estimación de la señal filogenética (ver ejemplos en la Figura A12 del Anexo I). Al utilizar la primera componente, nos enfocamos en la segregación más clara entre las secuencias, tendiendo a minimizar la influencia de los eventos más recientes y centrándonos en la señal filogenéticas más influyente de la estructura de la filogenia.

Con el objetivo de comprender mejor el fenómeno analizamos cuales son los factores que más afectan al *background*. Analizamos la relación entre los umbrales objetivo (precisión 0.8) con el número de secuencias, el número de posiciones en el alineamiento y la señal filogenética utilizando correlaciones parciales (mediante el paquete ppcor [217]) para descontar el efecto que unas variables puedan tener sobre otras en su relación con los umbrales objetivo. La correlación parcial más fuerte se produce entre los umbrales objetivo y el número de posiciones (-0.66; n=490; *p-valor*=$3.6*10^{-63}$), mostrando que un mayor número de posiciones se corresponde con un umbral objetivo más bajo. La correlación parcial para el número de secuencias es 0.38 (n=490; *p-valor* =$1.8*10^{-18}$). Sin embargo, se observan dos tendencias diferenciadas dependiendo del número de secuencias. Para menos de 500 secuencias, la correlación parcial es 0.42 (n=350; *p-valor*=$1.4*10^{-16}$) mientras que para más de 500 secuencias es -0.13 (n=140; *p-valor*=0.12). Este cambio de tendencia es aún más claro para alineamientos con 1 000 secuencias o más, donde anticorrelacionan de forma notable (-0.69; n=33; *p-valor*=$1.56*10^{-5}$), observación que discutiremos a continuación.

Con respecto a la señal filogenética, también encontramos que correlaciona con los umbrales objetivo (0.34; n=490; *p-valor*=$1.5*10^{-14}$). Esto es esperable, una mayor señal filogenética tenderá a generar covariaciones más fuertes y, por tanto, a incrementar los umbrales objetivo. De nuevo encontramos dos tendencias con respecto al número de secuencias. Con pocas secuencias, menos de 500, la correlación parcial entre los umbrales objetivo y la señal filogenética es 0.33 (n=340; *p-valor*=$3.8*10^{-10}$). Para un mayor número de secuencias, más de 500, la correlación parcial es más débil (0.19; n=140; *p-valor*=$2.7*10^{-2}$). Con más de 1000 secuencias, esta correlación desaparece (0.04; n=33; *p-valor*=0.85). Por un lado, esto sugiere que DCA podría corregir la señal filogenética. Por otra parte, nos indica una posible explicación para el cambio de tendencia en la correlación entre el número de secuencias y los umbrales objetivo. Con muchas secuencias los *scores* pueden reducirse por una corrección de las covariaciones provenientes de las relaciones filogenéticas. Esto no es, sin embargo, la única interpretación posible. Un mayor número de secuencias también puede estar asociado a un menor impacto de la señal filogenética en el modelo DCA. Al tener más secuencias, la señal filogenética es menos homogénea tendiendo a afectar a menos posiciones de forma global ya que habrá un mayor número de eventos evolutivos cuyo efecto es en promedio más local afectando a distintas particiones del árbol. También es perfectamente posible que se produzca un efecto conjunto de los dos efectos, una corrección por DCA y un menor impacto de la señal filogenética en el modelo DCA por la composición del

alineamiento. En cualquier caso, estas observaciones sobre el menor impacto de la señal filogenética en el *background* para rangos de secuencias mayores a 500 o 1 000 secuencias es consistente con la observación empírica de que en este rango de secuencias es cuando DCA funciona realmente bien [72,117,135,152].

Como es esperable dada la fuerte correlación entre el umbral objetivo y el máximo empírico promedio, las correlaciones parciales entre el máximo empírico promedio y el número de secuencias, posiciones y señal filogenética siguen las mismas tendencias (Tabla A2 del Anexo I). Esto incluye la señal filogenética (0.38 y 0.35 para la aleatorización de emparejamientos y por columnas, respectivamente; n=490; *p-valores*=$2.4*10^{-18}$ y $7.4*10^{-16}$), lo cual implica que los máximos empíricos son capaces de captar, al menos parcialmente, la señal filogenética.

Cabe destacar que los 3 casos con un mayor número de secuencias, más de 10 000 en todos ellos, se encuentran entre aquellos que muestran una correlación más baja entre los valores de APC y la media de los *scores* crudos empíricos (0.82, 0.85 y 0.93; Figura 4.10C). Esto sugiere que, con muchas secuencias, se produce algún tipo de corrección implícita dentro del modelo DCA, una corrección que podría ser estar relacionado con la filogenia subyacente. Por otra parte, 2 de estos 3 casos son los que tienen mejores predicciones, el sistema regulador de dos componentes y los transportadores ABC. En estos casos, el comportamiento de los *scores* crudos es sorprendentemente bueno, similar al del *score* APC (Tabla 4.3). Este hecho está en total contraste con el resto del conjunto de datos, donde el *score* crudo funciona de manera muy pobre (Figura 4.12A y B). Estos resultados sugieren que el modelo DCA podría ser capaz de corregir implícitamente por la señal filogenética sin la necesidad de APC cuando hay suficientes secuencias en el alineamiento. Esto también explicaría la menor correlación entre APC y *scores* empíricos nulos, la diferencia podría venir de la corrección filogenética previa en los *scores* crudos. De nuevo, es posible que esto se pueda explicar, al menos parcialmente, por un cambio en la naturaleza de la señal debido a la composición del alineamiento emparejado o una combinación de estos factores.

	Sistema regulador de dos componentes		Transportadores ABC	
Número de predicciones	PC *score* crudo	PC *score* APC	PC *score* crudo	PC *score* APC
10	10	10	10	10
25	24	24	23	24
50	48	46	31	39
100	72	68	36	45
250	110	98	43	57
500	151	118	51	69

Tabla 4.3 Comparativa de *scores* APC y crudos en dos casos particularmente buenos. Se muestra el número de predicciones correctas (PC) para los dos *scores* dependiendo del número de predicciones para el sistema regulador de dos componentes y los transportadores ABC.

4.2.3 Influencia del número de secuencias

A continuación, nos centramos en un problema clave en la aplicación de los métodos DCA, esto es, la necesidad de grandes alineamientos con muchas secuencias. De hecho, en muchos estudios se ha apuntado que las metodologías DCA no funcionan bien cuando los alineamientos de entrada tienen pocas secuencias [26,72,151]. Para estudiar la influencia del número de secuencias, dividimos nuestro conjunto de casos en dos subgrupos: uno de 140 IDDIs con alineamientos grandes (500 o más secuencias) y otro de 350 casos con alineamientos pequeños (menos de 500 secuencias) [117,135,152]. Como era de esperar, hay muchas predicciones de contacto correctas en el subconjunto de alineamientos grandes (Figura A13, Anexo I). Por el contrario, la precisión cae rápidamente en el subconjunto de alineamientos pequeños (Figura A13, Anexo I), lo cual significa que hay predicciones incorrectas con *scores* APC elevados. Sin embargo, observamos que existen 64 de los 350 IDDIs con alineamientos pequeños para los cuales el par con mayor *score* APC se corresponde con un contacto, algo altamente improbable al azar (Materiales y Métodos, sección 3.12).

De hecho, los casos predichos correctamente con alineamientos pequeños son relativamente comunes. Usando los umbrales objetivo (precisión ≥ 0.8), hay 55 IDDIs con alineamientos grandes correctamente predichos mientras que hay 64 con alineamientos pequeños. Más en concreto, si subdividimos en varios grupos por el número de secuencias, vemos que la probabilidad de que un alineamiento sea correctamente predicho crece linealmente con el número de secuencias (Figura A14 del Anexo I). No es sorprendente que, conforme la cantidad de datos crece, la probabilidad de encontrar predicciones correctas crezca en consecuencia, dado que existe más señal y se obtienen mejores estimaciones de las frecuencias de los aminoácidos. Lo expuesto anteriormente significa que la estrategia comúnmente utilizada de descartar alineamientos pequeños conlleva que perdamos una parte importante de predicciones correctas. Por tanto, el argumento habitual de que DCA no funciona con alineamientos pequeños induce a error. El problema consiste en distinguir las predicciones correctas de las incorrectas ya que con pocas secuencias es más difícil. Esto es probablemente debido a la existencia de más ruido, proveniente de un mayor impacto de la señal filogenética que actúa como un factor de confusión y de una peor estimación de las frecuencias de aminoácidos debido a la limitada cantidad de datos disponible. Esta mayor cantidad de ruido posiblemente hace que los *backgrounds* fluctúen más entre casos y sus *scores* sean menos comparables.

El *score* MEND mejora notablemente la situación. Para ilustrar este hecho, utilizaremos un umbral del *score* APC de 0.3 propuesto en un trabajo anterior como un umbral fiable [135]. Este umbral se comporta razonablemente bien en alineamientos grandes, la proporción de casos predichos correctamente (precisión ≥ 0.8) es de 69% (de 52 IDDIs predichos). Pero es notablemente peor para alineamientos pequeños, la proporción es del 37% (de 59 IDDIs predichos). En el caso del *score* MEND, la proporciones de casos predichos correctamente equivalentes son 73% (de 52 IDDIs predichos) y 59% (de 59 IDDIs predichos), respectivamente. Independientemente del umbral usado, el score MEND se comporta mejor en ambos escenarios, aunque la mejora es más importante con pocas secuencias (Figura 4.13). En cuanto al *z-score* APC, su comportamiento es similar al de *score* APC, aunque algo inferior con pocas secuencias. Pero podría ser útil cuando se busque

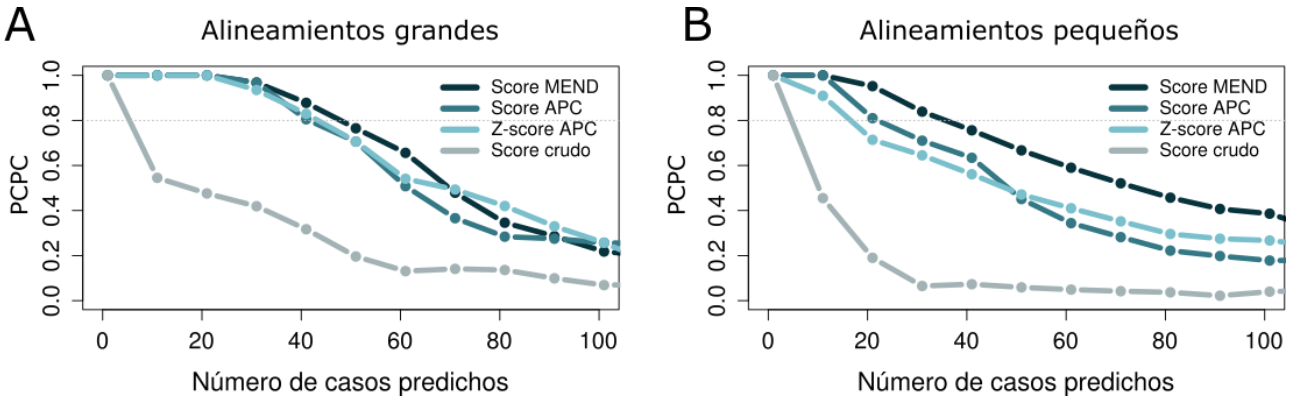


Figura 4.13 Precisión y número de secuencias. A) Proporción de casos predichos correctamente (PCPC, con precisión ≥ 0.8), en función del número de casos predichos para el *score* MEND, el *score* APC, el *z-score* APC y el *score* crudo para los 140 casos con alineamientos grandes (500 o más secuencias). B) Idem. para los 350 casos con alineamientos pequeños (menos de 500 secuencias).

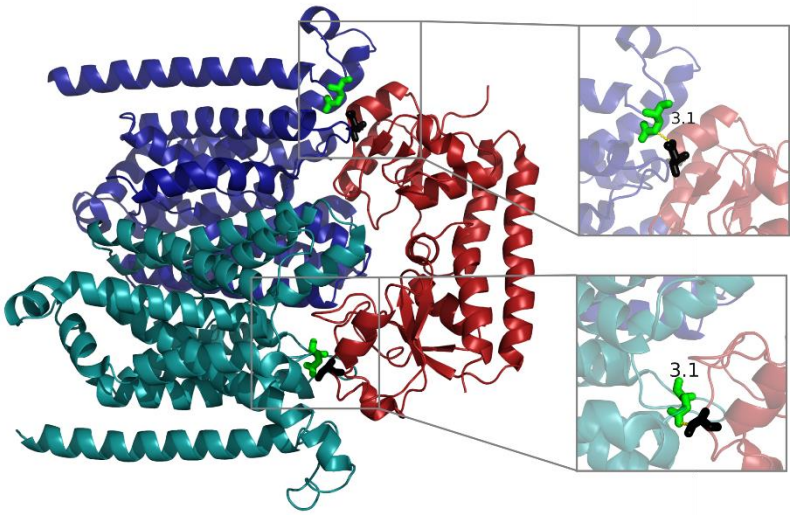
mayor exhaustividad (*recall*) sacrificando precisión. Por tanto, nuestra aproximación muestra una mayor robustez al número de secuencias lo cual permite trabajar más sistemáticamente con alineamientos pequeños y recuperar casos predichos correctamente en ellos, una característica importante en términos de aplicabilidad.

A pesar de la ventaja evidente que supone poder recuperar predicciones correctas de casos con alineamientos pequeños, el número de predicciones de contactos que se pueden recuperar por caso es muy bajo (Figura A14 del Anexo I). Como ya mencionamos anteriormente, aunque es cierto que un número pequeño de predicciones de contacto fiables es suficiente para generar buenos modelos estructurales de la interacción entre proteínas [144,145], la calidad de los modelos resultantes tenderá a ser más baja con pocas predicciones ya que existe una mayor incertidumbre sobre el modo de interacción.

Resumiendo lo observado anteriormente, es posible distinguir 3 escenarios dependiendo del número de secuencias en los alineamientos emparejados: i) con una gran cantidad de secuencias, el comportamiento de los *scores* crudos es comparable al del *score* APC. ii) Para un número de secuencias relativamente alto, el *score* APC se comporta adecuadamente, pero el *score* MEND es ligeramente superior (Figura 4.13A). iii) Con pocas secuencias, la diferencia entre ellos es mayor (Figura 4.13B).

4.2.4 Selección de ejemplos ilustrativos

Un ejemplo de donde se observa la mejora de *score* MEND es el complejo BuCDF, un complejo membranal importador de la vitamina B12. A pesar de tener un alineamiento emparejado bien poblado (2578 secuencias; dominios Pfam FecCD y Peripla_BP_2) y predicciones correctas en la parte superior del ranking, el *score* APC de las predicciones de contacto se encuentra lejos de un umbral fiable (p. ej. 0.3563). En cambio, como muestra la Figura 4.14 si existen dos predicciones correctas que superan un umbral riguroso para el *score* MEND (> 1.4997).



Índice	Dist.	*Score* APC	*Score* MEND
1	**3.1**	0.22	**1.66**
2	**3.14**	0.21	**1.62**
3	**3.37**	0.17	1.29
4	**4.62**	0.15	1.14
5	**2.96**	0.14	1.07

Figura 4.14 Predicción en BtuCDF. En la parte izquierda, se muestra la estructura (código PDB 2QI9) de la interacción entre la proteína BtuF (cadena F, en rojo) con la proteína BtuC (dos cadenas, A y B, es distintos tonos azules) del complejo importador de vitamina B12 BtuCDF. En las regiones ampliadas se muestran las predicciones de contacto por encima del umbral del *score* MEND (>1.4997), con los correspondientes aminoácidos en bastones, en verde para BtuC y en negro para BtuF. Los átomos más cercanos entre los aminoácidos están conectados por una línea amarilla y se muestra la distancia en angstroms que los separa. En la parte derecha, se puede observar las 5 primeras predicciones con su correspondiente distancia en la estructura (columna Dist., en negrita cuando es menor del umbral de distancia), *score* APC (ninguna supera el umbral de 0.3563) y *score* MEND (en negrita cuando superan el umbral de 1.4997).

Es posible encontrar casos predichos a pesar de tener alineamiento emparejados con muy pocas secuencias. En la interacción entre el citocromo f y la proteína de Rieske (dominios Pfam Apocytochr_F_C y CytB6-F_Fe-S) en el complejo del citocromo b_6f, encargado de transferir electrones entre los fotosistemas I y II durante la fotosíntesis, encontramos una fuerte coevolución (Tabla 4.4). Entre las 5 primeras predicciones, se observan 3 predicciones correctas. Este caso representa una predicción sorprendentemente buena considerando que el alineamiento emparejado tiene solo 45 secuencias.

Aunque el *score* MEND es robusto al número de secuencias en el alineamiento, no siempre es capaz de recuperar las predicciones correctas que se encuentran en la parte superior de los rankings. Ejemplo de ello es la interacción entre la flagelina y la chaperona FliS (dominios Pfam Flagellin_C y FliC) cuyo alineamiento emparejado correspondiente cuenta con solo 147 secuencias. A pesar de que las 3 primeras predicciones son contactos físicos, sus *scores* MEND no son lo suficientemente altos como para poder ser detectados *a priori* (Tabla 4.4). Aunque en estos casos las predicciones de contacto puedan ser útiles combinadas con otras informaciones, por sí solas han de ser interpretadas con cautela pues en estos niveles del *score* MEND los falsos positivos son habituales (Tabla 4.2).

Dominios Pfam	Apocytochr_F_C CytB6-F_Fe-S					Flagellin_C FliC				
Índice	Pdb1	Pdb2	Dist.	Score APC	Score MEND	Pdb1	Pdb2	Dist.	Score APC	Score MEND
1	271	27	**6.2**	0.48	1.60	2508	1069	**3.6**	0.26	1.32
2	268	27	**4.3**	0.40	1.33	2493	1081	**2.3**	0.24	1.24
3	176	40	60.1	0.32	1.05	2487	1112	**3.9**	0.21	1.08
4	265	27	**7.4**	0.27	0.91	2499	1022	11.5	0.21	1.08
5	197	38	69.7	0.26	0.90	2493	1069	14.8	0.18	1.02

Tabla 4.4 Predicciones en casos con alineamientos con pocas secuencias. Se muestran las 5 primeras predicciones de contacto entre los dominios Pfam Apocytochr_F_C y CytB6-F_Fe-S (parte izquierda, interacción entre el citocromo f y la proteína de Rieske), así como entre los dominios Pfam Flagellin_C y FliC (parte derecha, interacción entre la flagelina y la chaperona FliS). En las columnas Pdb1 y Pdb2 aparecen la numeración en el PDB correspondiente de las posiciones Pfam mapeadas a la estructura del citocromo b_6f (código PDB: 4OGQ, cadenas C and D para Apocytochr_F_C y CytB6-F_Fe-S, respectivamente) y de la interacción entre la flagelina y la chaperona FliS (código PDB: 1ORY, cadenas B and A, para la flagelina y FliS respectivamente). La columna Dist. muestra la distancia en angstroms en la estructura para cada par de posiciones.

Volviendo al caso de la NADH deshidrogenasa estudiado en la sección 4.1.5, podemos observar la mejora producida por el uso del *score* MEND. Utilizando umbrales equivalentes (Tabla 4.2), el *score* MEND permite predecir más contactos en la interacción entre las proteínas de los genes nuoJ y nuoK, así como predecir contactos en la interacción entre nuoJ y nuoA (siendo el único que tiene predicciones por encima del umbral). Por el contrario, ni siquiera el *score* MEND recupera predicciones en la interacción entre nuoN y nuoK (Tabla 4.5). El hecho de que estas predicciones se encuentren en torno al máximo empírico promedio (*score* MEND cercano a 1) sugiere que estas predicciones podrían ser difíciles de recuperar incluso con formas mejoradas de estimar el *background* y sea necesario mejorar el modelo coevolutivo en sí mismo para poder obtener predicciones correctas y que sean detectables *a priori*. En comparación con lo mostrado en la sección 4.1.5, es necesario usar un umbral más estricto para el *z-score* APC (10.9428 en vez de 8). Esto es una consecuencia de la incorporación de muchos casos con pocas secuencias donde aparecen predicciones con un *z-score* APC alto pero incorrectas, lo cual obliga a utilizar un umbral más exigente del *z-score* APC para seguir manteniendo una precisión alta. El *score* MEND es robusto a este efecto, permitiendo la predicción de más contactos correctamente y en un mayor número de casos en condiciones equivalentes.

Genes	nuoJ nuoK				nuoJ nuoA				nuoN nuoK			
Pfams	Oxidored_q3 Oxidored_q2				Oxidored_q3 Oxidored_q4				Proton_antipo_M Oxidored_q2			
Índice	Dist.	score APC	z-score APC	score MEND	Dist.	score APC	z-score APC	score MEND	Dist.	score APC	z-score APC	score MEND
1	**3.9**	**0.66**	**15.41**	**3.19**	**4.2**	0.33	8.27	**1.73**	38.4	0.23	6.91	1.34
2	**4.4**	**0.38**	9.01	**1.86**	**6.4**	0.33	8.10	**1.69**	**7.6**	0.20	5.84	1.13
3	**4.3**	**0.37**	8.79	**1.81**	12.8	0.27	6.66	1.39	**6.6**	0.18	5.28	1.02
4	**3.4***	0.34	7.99	**1.64**	48.7	0.25	6.12	1.27	**4.4**	0.18	5.25	1.02
5	**4.0**	0.31	7.39	**1.52**	19.9	0.24	5.90	1.23	17.1	0.18	5.24	1.02
6	**3.6**	0.30	7.13	1.46	16.6	0.23	5.82	1.21	**7.4**	0.17	5.07	0.98
7	**3.4**	0.29	6.94	1.43	**5.9**	0.21	5.13	1.06	31.9	0.17	5.05	0.98
8	**3.4**	0.28	6.63	1.36	36.11*	0.20	5.05	1.05	30.7	0.16	4.94	0.96
9	28.9	0.28	6.51	1.34	**6.0**	0.20	4.96	1.03	14.8	0.16	4.87	0.94
10	**2.7**	0.26	6.13	1.26	35.0	0.20	4.92	1.02	36.5	0.16	4.78	0.93

Tabla 4.5 Predicciones en la NADH deshidrogenasa usando diferentes *scores.* Se muestran la distancia (columna Dist.) y los *scores* APC *z-scores* APC y *scores* MEND para las 10 primeras predicciones de contacto de las interacciones entre las proteínas de los genes nuoJ y nuoK (dominios Pfam Oxidored_q3 y Oxidored_q2); nuoJ y nuoA (Oxidored_q3 y Oxidored_q4); nuoN y nuoK (Proton_antipo_M y Oxidored_q2). Las distancias se han extraído de la estructura 3RKO de la NADH deshidrogena en *Escherichia Coli*. En negrita, se muestra las distancias por debajo del umbral de contacto (8 Å) y los *scores* por encima del umbral correspondiente a una proporción del 80% de casos predichos correctamente con precisión≥0.8. Estos umbrales son: 0.3563 para *score* APC, 10.9428 para el *z-score* APC, 1.4997 para el *score* MEND. *Las distancias anotadas con asterisco se han extraído de la estructura 4HEA debido a que alguna de las posiciones correspondientes a la predicción en cuestión no se puede mapear a la secuencia de la estructura 3RKO, por lo que no se puede asignar una distancia entre el par de posiciones

5. DISCUSIÓN

5.1 Coevolución y conservación estructural

Como comentamos en la introducción, es bien sabido el alto grado de conservación estructural de la estructura de las proteínas (sección 1.3.2) y de las interfaces (sección 1.3.5) aun cuando existe una gran divergencia en secuencia. Esto implica que existen diversos conjuntos de interacciones entre residuos compatibles con estructuras e interfaces similares. Cabe preguntarse hasta qué punto esta conservación se debe al mantenimiento de un conjunto, relativamente reducido con respecto al total, de interacciones entre residuos y cuál es la relación de la coevolución con ellos. Diversos trabajos han apuntado la importancia de grupos, relativamente pequeños, de interacciones entre residuos para el mantenimiento de la estabilidad de la estructura, ya sea mediante coevolución [218] o simulaciones moleculares [219]. Y, en el caso de interacciones entre proteínas, se ha mostrado la habitual presencia de *hot regions* y *hot spots* (regiones y residuos de anclaje) [220]. Estos trabajos destacan la desigual responsabilidad de las distintas interacciones entre residuos en la estabilidad de las proteínas y sus interacciones. Aunque es importante recordar que se trata de procesos altamente cooperativos, por lo que es difícil saber hasta qué punto podrían ser reductibles a grupos particulares de interacciones entre residuos. Por otro lado, se ha postulado habitualmente a la coevolución entre residuos como probable causa del mantenimiento de esta conservación estructural de las proteínas [221] y de sus interfaces [222]. En efecto, la observación de coadaptaciones en proteínas e interfaces mediante mutaciones compensatorias [45,223] provee de un mecanismo sobre la conservación de interacciones entre residuos, potencialmente capaz de explicar la conservación estructural de las proteínas e interfaces. Sin embargo, la observación de una relación más global y sistemática de la relación entre coevolución y conservación estructural ha sido elusiva. En este sentido, nuestras observaciones proporcionan una asociación más directa y sistemática de la relación entre coevolución y conservación estructural en interfaces, probablemente extensible a las proteínas según apunta el trabajo de Zea *et al* [224].

Como hemos mostrado en nuestro primer estudio (sección 4.1), la coevolución está asociada a una conservación estructural en la región de interacción entre dominios. Esto se observa tanto a nivel de toda la interfaz como para contactos físicos específicos. En relación con toda la interfaz, la detección de fuertes señales de coevolución se corresponde con un mínimo nivel de conservación de la interfaz en su conjunto (Figura 4.3A). En cuanto a los contactos físicos, observamos que la gran mayoría de contactos físicos con fuertes señales de coevolución en procariotas se corresponden con contactos conservados en las estructuras de procariotas y eucariotas (Figura 4.3B). Este resultado es coherente con la observación de que se detecte mayores señales de coevolución en el núcleo (*core*) de las interfaces dada su mayor tendencia a estar conservados estructuralmente [225]. Mostramos que esta asociación entre coevolución y conservación estructural es robusta a factores como la conservación de las posiciones, la definición de contacto y el umbral coevolución (Figura 4.4). De hecho, esta asociación es aún más fuerte, en términos estadísticos, para

umbrales menores al umbral de referencia, esto es, donde las predicciones son menos fiables pero se consideran un mayor número de observaciones (Figura 4.4).

Posteriores estudios han reforzado y extendido esta observación. Zea *et al.* estudiaron la variabilidad estructural en familias de proteínas y su relación con la coevolución, mostrando que los contactos predichos por DCA están conservados en éstas, en lugar de ser específicos del confórmero [224]. Por otra parte, muy recientemente, Con *et al.* han mostrado también que fuertes señales de coevolución están asociadas a contactos conservados en interfaces entre proteínas a distancias evolutivas relativamente cortas [166]. Este resultado refuerza nuestra observación inicial y la extiende a casos de menor divergencia evolutiva. Estos estudios apuntan a que la relación entre coevolución y contactos estructuralmente conservados es una propiedad general en la evolución de las proteínas y sus regiones de interacción con otras proteínas.

Desde la introducción de DCA y metodologías similares, ha quedado patente la gran capacidad de predicción de contactos que ofrecen los métodos basados en covariación, incluyendo los resultados presentados en esta tesis. Este hecho significa que existe un enriquecimiento, que puede ser muy sutil, hacia determinados pares de aminoácidos en algunos pares de posiciones. Se ha mostrado que las predicciones están asociadas en su inmensa mayoría a contactos físicos, con poca evidencia que apunte a relaciones alostéricas [226]. Algo que quizás ha pasado bastante desapercibido en la comunidad es que estos enriquecimientos deben estar asociados a restricciones estructurales (como contactos físicos) que han debido permanecer conservadas durante el tiempo suficiente como para alcanzar un nivel de señal detectable. En la primera parte de la tesis mostramos que, efectivamente, este es el caso para los contactos físicos. Esta observación es importante tanto por sus consecuencias prácticas en la predicción de contactos a grandes distancias evolutivas, como por lo que revela en cuanto a la importancia de la coevolución como restricción en la evolución de los complejos de proteínas.

También puede ser importante considerar, junto a la coevolución, la conservación de las posiciones, por varias razones. Por un lado, hay restricciones muy importantes sin, o con poca, variación en secuencia que es necesario incluir para tener una perspectiva más completa. Por otro lado, la conservación ayuda a detectar residuos funcionalmente importantes [83,227–229]. Además, la conservación de la posición y la coevolución no son independientes y su análisis conjunto puede ayudar a identificar restricciones importantes que sean difíciles de detectar por separado (ver p. ej. [230]).

El balance entre conservación y cambio puede interpretarse como el resultado de la colisión de fuerzas actuando a niveles distintos, la conservación a nivel de restricciones de estructura y el cambio a nivel de la variación de la secuencia de aminoácidos. Un balance que puede interpretarse en el contexto de la hipótesis de la Reina Roja de Van Valen (sección 1.1.1) para la coevolución entre especies, en el que compiten las restricciones por interacciones con otras especies y con la propia evolución de cada especie. Un escenario en el que la coevolución actuaría como la fuerza que conecta estos dos niveles y compatibiliza la variación en secuencia con la conservación estructural.

5.2 Estimación y corrección de la distribución de fondo

En la segunda parte de la tesis proponemos un método para estimar y corregir la distribución del fondo (*background*) específica de cada caso. En esta parte de la tesis, primero observamos la similitud entre la media de los *scores* crudos empíricos con la corrección APC (Figura 4.10). Esto permite dar una interpretación claro al efecto que tiene la corrección APC en el contexto de DCA, se trata de una estimación de la media de la distribución nula de cada par de posiciones que mejora la calidad de las predicciones mediante su substracción.

Mostramos que la estrategia de un umbral único para todos los casos está lejos de ser óptima, sobre todo, cuando se consideran casos con alineamientos de pocas secuencias (Figura 4.11B). Estudiamos el uso de aleatorizaciones para estimar el *background* esperable para cada caso, esto es, cada par de familias, constatando que los máximos obtenidos en dichas aleatorizaciones proporcionan una buena estimación del *background* de cada caso (Figura 4.11C y D). Para realizar una comparativa sistemática con otros esquemas de *scoring*, traducimos la estimación del *background* en una corrección específica de cada caso. La comparativa del rendimiento revela que se mejora la precisión de las predicciones de contacto (Figura 4.12A y Tabla 4.2) y que es posible predecir bien más casos en condiciones equivalentes (Figura 4.12B y Tabla 4.2), lo que se traduce en una ampliación del ámbito de aplicación. Estas mejoras son más pronunciadas en el caso de alineamientos de pocas secuencias (Figura 4.13). Dado que el número de secuencias es, en la práctica, el factor más limitante en las predicciones de contacto entre proteínas, esto supone un avance relevante ya que permite trabajar con alineamientos pequeños de forma más sistemática.

Es probable que aproximaciones similares basadas en aleatorizaciones puedan ser útiles en el caso de predicciones de contactos intradominio. Sin embargo, hay que tener presente algunas diferencias que podrían dificultar la adaptación de nuestra aproximación a este contexto. En primer lugar, la mejora es menos notable en casos donde hay un número no pequeño de secuencias. Y para dominios el número de secuencias promedio es mucho más alto, dado que no es necesario realizar emparejamientos de secuencias. En segundo lugar, las señales intradominio, coevolutivas o filogenéticas, son más numerosas y fuertes en el caso de dominios con respecto al de interacciones, lo que puede tener un impacto importante en la estimación del *background*. El análisis de su aplicabilidad en el contexto de predicciones intradominio requerirá trabajos adicionales en el futuro.

Hemos estudiado el efecto que tienen factores como el número de secuencias y posiciones en el alineamiento emparejado y la influencia de las relaciones filogenéticas en el *background* estimado en cada caso. Estas observaciones clarifican la influencia de estos factores y pueden ser útiles para la posterior mejora de este tipo métodos. La influencia de las relaciones filogenéticas es de especial interés por lo que la discutiremos en más detalle en la siguiente sección.

5.3 El problema de las relaciones filogenéticas entre secuencias

Como se ha discutido en varios puntos de la tesis (sección 1.3.3.3, 4.1.5 y 4.2.3), las relaciones filogenéticas entre las secuencias suponen un problema importante para los métodos basados en covariaciones ya que es difícil discriminar la covariación debida a la filogenia subyacente de aquella debida a restricciones estructurales o funcionales. En el trabajo de Vorberg *et al.*, los autores proponen una nueva corrección para el *background* de cada par de posiciones [116]. Esta corrección se basa exclusivamente en la conservación de las posiciones y tiene un rendimiento similar, aunque algo inferior, al que tiene la corrección APC. Los autores interpretan que el ruido asociado al *background* de cada par de posiciones es principalmente debido a la conservación de éstas y que la contribución de la conservación de las posiciones es aproximadamente el doble que el de la filogenia [116]. Nuestros resultados siguen esta línea. En primer lugar, el rendimiento de la aleatorización de emparejamientos (que incluye todas las señales intradominio) es similar, aunque algo superior, al de por columnas, a pesar de que la aleatorización por columnas solo considera el número de secuencias, posiciones y la conservación de éstas. En segundo lugar, las aleatorizaciones permiten obtener buenas estimaciones del *background* (tanto de cada par de posiciones como para cada caso en conjunto) pese a que no consideran directamente las señales interdominio (incluidas las filogenéticas). En este sentido, tener en cuenta la conservación de posiciones parece ser suficiente para obtener buenas estimaciones.

Sin embargo, es necesario hacer una puntualización importante. La conservación de las posiciones y cuanto afectan al modelo no es independiente de la filogenia. Volviendo al ejemplo de la sección 1.3.3.3, si tenemos una filogenia con dos grupos de secuencias muy similares dentro de su grupo y muy distintas con respecto al otro grupo, la conservación observada de las posiciones reflejará esta situación y estará más relacionadas con la estructura filogenética de la familia que con restricciones estructurales. Desde este punto de vista, la correlación entre la estimación del *background* mediante los máximos de las aleatorizaciones y la estimación de la señal filogenética podría ser consecuencia de los patrones de conservación de las posiciones. Un mayor número de posiciones con conservaciones parecidas debido a las relaciones filogenéticas entre las secuencias pueden tender a generar, incluso al azar, *scores* más altos.

El éxito de DCA en la predicción de contactos es debida a su capacidad de distinguir correlaciones directas de correlaciones indirectas. Estas correlaciones indirectas han sido típicamente atribuidas a cadenas de contactos que propagan las correlaciones, un argumento basado en las redes de contactos observadas en las estructuras de las proteínas. Sin embargo, la inversión de la matriz de covarianza minimiza el efecto proveniente de las primeras componentes principales [115,231], que sabemos que están asociadas a la señal filogenética. Esto sugiere que estas correlaciones pueden también venir de la filogenia subyacente por lo que el éxito de DCA podría estar relacionado con su capacidad para minimizar la señal filogenética. Hemos observados dos evidencias que apoyan esta interpretación. Por un lado, en los casos con una enorme cantidad de secuencias, los *scores* crudos extraídos del modelo funcionan de forma similar a aquellos obtenidos tras aplicar APC. Por otra parte, el *background* deja de correlacionar con la señal filogenética cuando hay bastantes secuencias. Sin embargo, estas observaciones también podrían ser explicadas por los

cambios en la composición en los alineamientos emparejados y su posible influencia en el impacto que tiene la filogenia en el modelo. En cualquier caso, esto no excluye que las correlaciones indirectas puedan venir, al menos parcialmente, de las redes de contactos. Pero pone de relieve que el éxito de DCA podría estar más relacionado con su capacidad para minimizar la señal filogenética de lo que es actualmente asumido en la comunidad de este campo de investigación.

El tratamiento de la señal filogenética continúa siendo un obstáculo importante para los métodos basados en covariación, dada su capacidad de generar covariación no relacionada con restricciones estructurales. De hecho, probablemente sea la mayor dificultad teórica actualmente, el "elefante en la habitación" en palabras de Erik van Nimwegen [232]. Nuestro trabajo sugiere que la señal filogenética parece tener un mayor impacto en el caso de DCA en alineamientos pequeños. Existen numerosas vías de mejora. Entre ellas, podemos destacar: i) un tratamiento más explícito de la señal filogenética que permita estimar, y contrastar o eliminar en caso de ser posible, su contribución en las covariaciones, ii) la inclusión de variables informativas de la filogenia en métodos de aprendizaje automático (como puede ser la estimación del *background* que proponemos aquí), iii) la búsqueda de patrones diferenciados entre las covariaciones debidas a restricciones estructurales o asociadas a la filogenia que permitan discriminarlos.

5.4 El proceso coevolutivo entre posiciones dentro y entre proteínas

Estudios pioneros en los años 60 y 70 ya observaron la codependencia entre posiciones y su relación con restricciones estructurales [233–235]. Sin embargo, propusieron mecanismos distintos para explicar esta observación: una mutación deletérea en una posición seguida de una mutación compensatoria en la otra [233] o dos cambios neutrales o casi neutrales en las dos posiciones [234,235]. En este sentido, podemos imaginar dos extremos acerca de cómo se produce el proceso coevolutivo entre dos posiciones: desde una mutación muy deletérea en una posición seguida de una mutación compensatoria en la otra a dos cambios neutrales o casi neutrales en las dos posiciones. Ambos extremos, sin embargo, presentan dificultades. En el primer caso, mutaciones muy deletéreas deberían desaparecer rápidamente de la población dificultando la aparición de mutaciones compensatorias. En el segundo caso, mutaciones neutrales o casi neutrales estarán asociadas a presiones selectivas débiles dificultando (aunque no excluyendo) la manifestación de la codependencia entre las posiciones. Cabe mencionar que cambios neutrales pueden estar asociados a una codependencia (que restringe los posibles caminos evolutivos), pero la falta de una presión de selección en un sentido hace más improbable, en principio, que éstas aparezcan y se fijen en la población. Entre estos extremos cabe todo un continuo de posibilidades. Dada la magnitud del proceso (potencialmente todas las proteínas existentes o que hayan existido) y la miríada de funciones y grados de conservación de las familias de proteínas, es probable que todas estas posibilidades hayan sido exploradas en el curso de la evolución. Cabe preguntarse cuál podría ser la forma de esta distribución, algo desconocido en la actualidad. El estudio de Fodor *et al.* [75], donde los métodos locales que consideran la conservación de la posición muestran una

superioridad sobre los que se basan únicamente en la covariación, sugiere que al menos cierta conservación, asociadas a mayores presiones de selección, podría ser más habitual.

Los cambios correlacionados se deben a interdependencias entre distintas posiciones dentro de las proteínas debidas a restricciones evolutivas importantes. Los cambios compensatorios y coadaptaciones forman una parte relevante de estas dependencias, pero es razonable asumir que existe un gradiente de interdependencias y restricciones evolutivas que, además, pueden variar con el tiempo y escenario evolutivo. En base a la alta cooperatividad entre muchas posiciones con respecto al plegamiento, dinámica y función de las proteínas, existe una ingente cantidad de posibilidades que pueden dar lugar patrones sutiles de covariación.

La coevolución entre residuos está íntimamente ligada a la epistasia (el efecto fenotípico no aditivo de dos o más mutaciones), ambas se deben a codependencias entre posiciones. De hecho, algunos autores lo utilizan de forma intercambiable [236]. Sin embargo, mientras que la epistasia (medida por medio de los efectos fenotípicos de las mutaciones de una determinada proteína en un contexto concreto) está relacionada a las codependencias en una proteína, la coevolución se refiere a las codependencias evolutivas. La epistasia es el sustrato sobre el que la (co)evolución actúa. Pero la correspondencia entre coevolución y epistasia no es siempre directa, basta con pensar en que cambios en el entorno pueden hacer que una determinada familia deje de tener algún impacto para el *fitness* (adecuación biológica) del organismo, por lo que la epistasia entre las posiciones dejará de tener influencia sobre la (co)evolución de la proteína. Y factores evolutivos (tamaño poblacional, deriva genética, entre otros) influyen en esta relación. Nuestra comprensión sobre la epistasia en, y entre, proteínas ha mejorado mucho a lo largo del tiempo y, sobre todo, recientemente gracias en gran medida al gran número de genomas secuencias y al desarrollo técnicas como el escaneo profundo de mutaciones (*deep mutational scanning*) que permite obtener medidas de *fitness* para un gran número de variaciones [237]. Está bien establecido que: i) los paisajes adaptativos son rugosos debido a epistasia [237,238] y ésta es ubicua [237,239,240], ii) la epistasia negativa (el *fitness* de los dobles mutantes es más bajo que el de la suma de las dos mutaciones individuales) es más habitual que la positiva (el *fitness* de mutaciones dobles es más alto que el de la suma de las dos mutaciones individuales) [238], iii) el efecto en *fitness* de la sustitución de aminoácidos observados en proteínas homólogas crece con la divergencia [45]. Este último hecho implica que un buen número de codependencias tenderán a manifestarse solo cuando exista suficiente divergencia.

Prácticamente todos los casos donde existe suficiente información en secuencia producen buenas predicciones, y son muchas las familias de dominios e interacciones estudiadas (ver p. ej. [153,166]). Esto significa que el factor limitante es la cantidad de información en secuencia [151], sugiriendo que la coevolución entre residuos podría ser un proceso ubicuo. Aunque es necesario recalcar que no tiene por qué producirse con la misma intensidad en, o entre, todas las familias de proteínas [27]. La coevolución dentro y entre proteínas probablemente sea un proceso ubicuo, pero no es descartable que existan casos específicos donde no sea así o su prevalencia sea marginal con respecto a otros factores. El hecho de la ubicuidad de la

epistasia apoya la ubicuidad de la coevolución. Además, las proteínas son marginalmente estables [241] por lo que efectos relativamente pequeños en estabilidad se pueden traducir en efectos importantes en *fitness* [238]. También hay que considerar que tanto la deriva genética como cambios en el entorno unido a la rugosidad de los paisajes adaptativo pueden motivar la constante búsqueda de nuevas soluciones cumpliendo un elevado número de restricciones cambiantes.

La ubicuidad de estos fenómenos debería ir asociado a cambios en la propensión de los aminoácidos a aparecer en las posiciones de las proteínas, algo para lo que existe evidencia [242–244]. Sin embargo, existe controversia sobre este asunto [245], con otros investigadores apoyando una gran conservación de la propensión de los aminoácidos [246]. Es posible que ambas perspectivas sean compatibles, las propensiones pueden estar conservadas durante largos periodos de tiempo o para algunas regiones o posiciones, pero la divergencia debería tender a cambiarlas paulatinamente. Estos cambios en las propensiones no significan necesariamente que las restricciones subyacentes tengan que cambiar de forma significativa, algo sugerido por nuestras observaciones sobre la conservación de contactos.

En la mayor parte de métodos de predicción de contactos mediante coevolución, como DCA por ejemplo, el análisis de la coevolución se realiza entre pares de posiciones. Sin embargo, cada aminoácido se encuentra en contacto físico con entre 2 y 11 aminoácidos [247]. Es por ello por lo que el análisis en términos de pares sea probablemente una simplificación exitosa. Trabajar en órdenes superiores es notoriamente más difícil. La paulatina divergencia tenderá a modificar qué aminoácidos son aceptables en una posición en función, entre otros factores, de cuales han sido los cambios que se han producido en su entorno tridimensional. Esto puede provocar un sutil enriquecimiento de determinados pares de aminoácidos en pares de posiciones. Algunas, y potencialmente muchas, restricciones estructurales solo se manifestarán cuando se han producido suficientes cambios para promover nuevos emparejamientos. Es necesaria suficiente divergencia. El hecho de que el efecto (negativo) en *fitness* de la sustitución de aminoácidos observados en proteínas homólogas crece con la divergencia apoya esta hipótesis.

Según lo propuesto por Andreas Wagner, la evolución de las proteínas se puede entender como una exploración en el paisaje adaptativo (*fitness landscape*) donde se alternan periodos de exploración neutral (mediante mutaciones neutrales o casi neutrales, mucho más comunes) con puntuales mutaciones adaptativas importantes para la adaptación de la especie [248]. En el contexto de la coevolución, un ejemplo de estas mutaciones adaptativas son los cambios en los cambios en las posiciones determinantes de especificidad, detectables mediante conservación diferencial en subfamilias que han motivado un conjunto de métodos [26]. Estos cambios adaptativos suelen estar asociados a duplicaciones y a una aceleración (ramas más largas en la filogenía) debido a un ajuste fino (*fine-tuning*) con rendimientos decrecientes (*diminishing returns*) [248]. Son comúnmente precedidas por mutaciones que incrementan la estabilidad para dar cabido a la usual desestabilización producidas por esas mutaciones adaptativas asociadas a mejoras funcionales [238]. Por otro lado, los periodos de exploración neutral han de ser compatibles con las restricciones estructurales existentes en cada momento.

Estudios recientes de evolución *in vitro*, donde se generan un gran número de secuencias a partir de una secuencia de partida mediante ciclos de generación de mutaciones aleatorias y selección (p. ej. utilizando antibióticos), han explorado la posibilidad de obtener buenas predicciones de contactos a partir de estas secuencias mediante coevolución y generar modelos tridimensionales a partir de ellas [249,250]. De uno de estos estudios se concluye que es posible obtener buenos modelos y que los contactos predichos están asociados a epistasia positiva [249]. Sin embargo, en el otro estudio no se obtienen buenas predicciones de contactos a pesar de seguir metodologías parecidas y alineamientos similares [250]. La diferencia es probablemente debida a que en este segundo estudio la concentración antibióticos es mucho más alta que en el segundo. Concentraciones altas (equivalentes a presiones selectivas altas) posiblemente dificultan una exploración más amplia del paisaje adaptativo, por lo que la cantidad de interacciones epistáticas exploradas se reduce. Esto podría implicar un balance entre presión selectiva y cuanta coevolución es esperable encontrar a un determinado nivel de divergencia. Estos trabajos se enmarcan en una corriente más extensa y muy importante sobre el efecto que tienen los paisajes adaptativos sobre la evolución de las proteínas y sus interacciones, una corriente apoyada en buena medida en la reciente disponibilidad de paisajes adaptativos experimentales gracias al escaneo profundo de mutaciones.

En base a todo lo anterior es posible construir un modelo conceptual de carácter hipotético acerca de la coevolución entre posiciones. La evolución de las proteínas se produce, generalmente, como una exploración en el paisaje adaptativo donde se alternan periodos de exploración neutral con puntuales mutaciones adaptativas importantes para la adaptación de la especie. La paulatina divergencia, la estabilidad marginal y cambios en las presiones de selección provocan una persistente, pero posiblemente sutil, coevolución entre posiciones para mantener la estabilidad estructural. Esta coevolución tenderá a estar asociada a restricciones, al menos, relativamente importantes, mutaciones neutrales o casi neutrales y epistasia positiva. La manifestación de estas restricciones estructurales puede requerir suficiente divergencia. Esto no significa que no se produzcan mutaciones deletéreas seguidas de mutaciones compensatorias, si no que podrían ser menos frecuentes en comparación con cambios neutrales o casi neutrales en respuesta a la divergencia. Esto también depende del umbral que se utilice para distinguir entre deletérea y casi neutral (ligeramente deletérea). Lo que destacamos es que es probable que mutaciones neutrales o ligeramente deletéreas sean mucho más comunes en este proceso. La asociación entre coevolución y epistasia positiva es interpretable desde dos puntos de vista. Por un lado, puede estar asociada a mutaciones compensatorias que probablemente ocurran en la población de forma relativamente simultánea (en el caso de mutaciones deletéreas). Por otro, puede estar asociada a mutaciones neutrales o casi neutrales que hayan sido habilitadas por substituciones o mutaciones precedentes (en el caso de mutaciones neutrales o casi neutrales). Por otra parte, las mutaciones adaptativas pueden desencadenar una rápida, en términos evolutivos, coevolución en grupo (afectando a un mayor número de posiciones) producto de la exploración de una nueva región funcional del paisaje adaptativo. Estos procesos se pueden ver influidos en gran medida por cambios relativamente frecuentes en el entorno que afectan de forma dinámica al paisaje adaptativo.

5.5 Limitaciones e influencia de otros factores

Hemos mostrado en esta tesis que es posible ampliar el ámbito de aplicación de los métodos de contacto entre proteínas a complejos en eucariotas y a casos con menos secuencias en el alineamiento emparejado. Aun así, el ámbito de aplicación sigue siendo bastante limitado. En primer lugar, porque la ampliación a eucariotas requiere homologías detectables en secuencia con procariotas, por lo que solo es aplicables a complejos muy antiguos evolutivamente. En segundo lugar, porque, aunque es posible trabajar con casos con pocas secuencias, la probabilidad de detectar señales de coevolución decrece con el número de secuencias. Por lo que, pese a los avances, la falta de suficientes secuencias en los alineamientos de entrada continúa siendo un obstáculo muy importante en la aplicación de este tipo de aproximaciones.

La limitación asociada a la falta de secuencia es mucho más grave en el caso de interacciones entre proteínas en eucariotas, algo debido tanto a la pequeña cantidad de genomas completamente secuenciados disponibles como a la mayor dificultad en generar emparejamientos fiables, tanto por la imposibilidad de utilizar la adyacencia genómica como evidencia de interacción, como por la existencia de un elevado número de paralogías. En eucariotas son más comunes las familias de proteínas que han sufrido duplicaciones y también es mayor el número de duplicaciones acontecidas en promedio. Estas expansiones basadas en duplicaciones hacen muy difícil distinguir qué parálogos de una familia interactúan con qué parálogos de la otra.

Cabe también mencionar que la predicción de contactos es mejor para complejos estables, con numerosas y fuertes interacciones entre residuos, que para interacciones lábiles y transitivas asociadas a interacciones más débiles [251]. Aunque es cierto que es posible predecir interacciones lábiles, con el sistema de dos componentes como ejemplo paradigmático [71,252,253], es una tarea mucho más difícil que podría requerir gran cantidad de datos para poder detectar señales más débiles. En términos generales, aunque sea posible, la detección de coevolución para interacciones lábiles está más limitada. En este sentido, algo similar ocurre con los plegamientos incorrectos y las agregaciones indeseadas. Aunque exista coevolución (asociada a una selección negativa) entre posiciones para evitarlos [254], es probablemente demasiado sutil para poder ser captada, al menos en la actualidad.

Otro tipo de casos difícilmente tratables mediante las aproximaciones tratadas aquí son las interacciones entre proteínas con un origen evolutivo reciente, ya que tendrán alineamientos poco poblados. En el caso de complejos de proteínas, sabemos que las interfaces más recientes se van añadiendo al complejo normalmente respetando las ya existentes [255]. Incluso si estas interfaces más recientes aparecen en las mismas proteínas que ya formaban el complejo antes de su incorporación, su señal será mucho más débil ya que en solo una parte del alineamiento habrá covariaciones asociadas a estas recientes restricciones estructurales.

5.6 Perspectivas futuras

La aplicación más evidente de los métodos propuestos en esta tesis es el estudio, o incluso modelado estructural en combinación con métodos de acoplamiento entre proteínas (*protein-protein docking)*, de interacciones entre proteínas en especies eucariotas o con pocas secuencias. Ya sea en casos concretos de especial interés o con información adicional o en escaneos a larga escala partiendo de interactomas experimentales.

Una opción interesante consistiría en utilizar los *scores* MEND en el contexto de interacciones en eucariotas. En primer lugar, el *score* MEND, al ser más robusto al número de secuencias del alineamiento de entrada, permite estudiar pares de familias de proteínas que tengan pocas duplicaciones. Para ello, se pueden utilizar los recientes y mejorados métodos para realizar emparejamientos [171,172,174]. En los casos con más duplicaciones, también se podrían utilizar estos métodos de emparejamiento, pero usando criterios más estrictos dada la robustez de nuestra aproximación al número de secuencias. Esto requeriría la búsqueda de un balance entre la rigurosidad del criterio y el número de secuencias recuperadas, algo que probablemente dependa del caso en cuestión. En este sentido, cabe destacar que, dado que nuestra aproximación se adapta al alineamiento de entrada en mayor medida, se podrían probar distintos niveles de rigurosidad y comparar los *scores* MEND obtenidos para tomar una decisión acerca de cuál es más adecuado en cada caso.

La información sobre coevolución puede servir como medida de conservación estructural, tal y como muestra el hecho de que seamos capaces de detectar contactos conservados tanto en eucariotas como en procariotas utilizando señales de coevolución solamente en procariotas. En este sentido, nuestro trabajo se asemeja a los trabajos desarrollados en el laboratorio liderado por de Ruth Nussinov sobre la conservación estructural de interfaces [156,256]. De estos trabajos se concluye que la identificación de interfaces conservadas estructuralmente, utilizando estructuras y cuando existe divergencia en secuencia, permite extrapolar estas interfaces a homólogos más distantes en secuencia. Esto permite mejorar la predicción de estructura de interacciones entre proteínas, en particular para casos de grandes divergencias. Sin embargo, tiene la desventaja de que es necesaria información estructural de interacciones divergentes, algo que no suele ser común. Estos resultados encajan con nuestras observaciones, con la diferencia que nosotros proponemos inferir conservación estructural de la interfaz a partir de la información de secuencia y la coevolución y no las propias estructuras.

Como hemos mencionado anteriormente en la discusión (sección 5.3), la influencia de las relaciones filogenéticas entre las secuencias sigue siendo un problema importante donde sería esperable que ulteriores desarrollos de métodos mejoren la capacidad de predicción de contactos mediante un mejor tratamiento de esta problemática. En particular el uso de métodos de aprendizaje profundo puede ser especialmente fructífero, aunque la escasez de información estructural de interacciones entre proteínas podría suponer un obstáculo importante.

Hemos observado que existe un claro enriquecimiento de las predicciones de contacto correctas para umbrales del z-*score* APC mucho menores de lo habitualmente utilizado (Figura 4.4). Esto significa que en los rankings hay un enriquecimiento, en muchos casos, de predicciones correctas en la parte superior. Esto probablemente sea una información útil si es convenientemente explotada en la combinación de predicciones de contactos entre proteínas por coevolución y métodos de acoplamiento entre proteínas (*protein-protein docking*). Existen también otras informaciones que no han sido aún explotadas en este contexto, entre las que cabe destacar la tendencia que tienen una posición a coevolucionar con residuos de su propio dominio o con los del otro dominio o proteína. Si consideramos el crecimiento del número de genomas secuencias, la mejora de los métodos de emparejamiento, la mejora presentada aquí o ulteriores mejoras, así como las informaciones adicionales que acabamos de mencionar es razonable pensar que la predicción de contactos por coevolución entre proteínas sumada a los métodos de acoplamiento entre proteínas (cuando existen modelos tridimensionales de los interactores) pueda proveer de modelos tridimensionales fiables para un número relevante y creciente de interacciones entre proteínas. La inclusión de predicciones de contactos fiables debería permitir tener una buena idea aproximada de la región de interacción. Lo cual podría ser refinado mediante el uso de métodos de acoplamiento entre proteínas que consideran la dinámica en la superficie de las proteínas, ya que habitualmente se producen reorganizaciones en la región de interacción al producirse el acoplamiento, para dar lugar a modelos 3D de alta precisión.

La disponibilidad de paisajes adaptativos experimentales puede ayudar enormemente a mejorar nuestra comprensión sobre cómo tiene lugar la coevolución en y entre proteínas. Esto, además de ser muy importante en sí mismo, puede traducirse en un mejor modelado del proceso y una mayor capacidad de detección de la coevolución entre residuos. De forma paralela, los avances conceptuales sobre la coevolución entre especies, como la teoría del mosaico coevolutivo, pueden servir de inspiración para el desarrollo de la coevolución molecular. La teoría de mosaico coevolutivo es capaz de integrar una mayor diversidad de escenarios y factores dando lugar a una perspectiva más amplia del fenómeno coevolutivo entre especies (sección 1.1.1). Entre éstos, se encuentra la dependencia de las interacciones entre especies a factores espaciales y temporales, así como a cambios en el entorno. A este respecto, es razonable pensar que existe margen de progreso en el campo de la coevolución molecular donde no es habitual considerar factores como la importancia funcional de cada familia de proteínas, la intensidad de las interacciones y los cambios en el entorno.

Como reflexión final, me gustaría destacar que el campo de la coevolución molecular, al que esperamos humildemente haber contribuido a desarrollar con esta tesis, ha cruzados unas primeras etapas importantes en términos de fiabilidad y aplicabilidad. Su impacto en el presente en la bioinformática estructural y en la biología estructural es enorme [257]. Queda aún mucho recorrido, oportunidades para avanzar y espacios que explorar. Desde mi perspectiva, la coevolución molecular no solo será relevante en su aplicación a otros campos de la biología si no como parte importante del propio conocimiento biológico.

6. CONCLUSIONES

1. La metodología que hemos desarrollado para la detección de señales coevolutivas entre proteínas, y entre dominios de proteínas, en procariotas predice de forma fiable y sistemática contactos físicos en las regiones de interacción.

2. La coevolución en complejos de proteínas está asociada a una conservación global de la estructura de la interfaz de interacción a largas distancias evolutivas.

3. La coevolución permite identificar contactos en interfaces particularmente conservados aún entre proteínas distantes evolutivamente. Estas observaciones indican que la coevolución debe haber jugado un papel importante en la conservación estructural de un número considerable de interfaces de proteínas.

4. La conservación estructural asociada a la coevolución permite obtener predicciones de contactos fiables en interfaces de eucariotas mediante la proyección de contactos predichos en proteínas de procariotas a sus homólogos en eucariotas, superando las limitaciones en eucariotas para la obtención de alineamientos emparejados suficientemente poblados.

5. Los máximos obtenidos mediante aleatorizaciones de los alineamientos emparejados proporcionan una buena estimación de la distribución de fondo específica de cada par de familias de proteínas en interacción, permitiendo fijar un umbral específico para cada caso.

6. Implementamos un nuevo método que, haciendo uso de los umbrales específicos, mejora de modo sistemático la calidad de las predicciones y extiende el ámbito de aplicación a un mayor número de casos. Esta mejora es especialmente significativa para casos con alineamientos de pocas secuencias.

7. BIBLIOGRAFÍA

1. Thompson JN. *The Coevolutionary Process*. Chicago: University of Chicago Press, 1994.

2. Darwin C. *On the Various Contrivances by Which British and Foreign Orchids Are Fertilised by Insects: And on the Good Effect of Intercrossing*. John Murray, 1862.

3. Darwin C. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. John Murray, 1859.

4. Ehrlich PR, Raven PH. Butterflies and Plants: A Study in Coevolution. *Evolution* 1964;**18**:586.

5. West K, Cohen A, Baron M. Morphology and behaviour of crabs and gastropods from lake Tanganyika, Africa: Implications for lacsutrine predator-prey coevolution. *Evolution* 1991;**45**:589–607.

6. Morand S, Hafner MS, Page RDM, Reed DL. Comparative body size relationships in pocket gophers and their chewing lice. *Biol J Linn Soc* 2000;**70**:239–49.

7. Dobzhansky T. Genetics of Natural Populations. XIX. Origin of Heterosis through Natural Selection in Populations of Drosophila Pseudoobscura. *Genetics* 1950;**35(3)**:288.

8. Wallace B. On Coadaptation in Drosophila. *Am Nat* 1953;**87**:343–58.

9. Ridley M. *Evolution*. 3rd ed. Malden, MA: Blackwell Pub, 2004.

10. Johnson SD, Anderson B. Coevolution Between Food-Rewarding Flowers and Their Pollinators. *Evol Educ Outreach* 2010;**3**:32–9.

11. Pazos F, Valencia A. Protein co-evolution, co-adaptation and interactions. *EMBO J* 2008;**27**:2648–55.

12. Eichler W. Some rules in Ectoparasitism. *J Nat Hist* 1948;**1**:588–598.

13. Hafner MS, Nadler SA. Phylogenetic trees support the coevolution of parasites and their hosts. *Nature* 1988;**332**:258.

14. Fahrenholz H. Ectoparasiten und abstammungslehre. *Zool Anz* 1913;**41**:371–374.

15. Kellogg VL. *New Mallophaga*. Leland Stanford, Jr. University, 1896.

16. Thompson JN. *Relentless Evolution*. University of Chicago Press, 2013.

17. Althoff DM, Segraves KA, Johnson MTJ. Testing for coevolutionary diversification: linking pattern with process. *Trends Ecol Evol* 2014;**29**:82–9.

18. Juan D, Pazos F, Valencia A. Co-evolution and co-adaptation in protein networks. *FEBS Lett* 2008;**582**:1225–30.

19. Van Valen L. A new evolutionary law. *Evol Theory* 1973;**1**:1–30.

20. Cott HB. *Adaptive Coloration in Animals*. Methuen; London, 1940.

21. Fox LR. Defense and Dynamics in Plant-Herbivore Systems. *Am Zool* 1981;**21**:853–64.

22. Thompson JN. *The Geographic Mosaic of Coevolution*. University of Chicago Press, 2005.

23. Gomulkiewicz R, Thompson JN, Holt RD, Nuismer SL, Hochberg ME. Hot Spots, Cold Spots, and the Geographic Mosaic Theory of Coevolution. *Am Nat* 2000;**156**:156–74.

24. Brown JKM, Tellier A. Plant-Parasite Coevolution: Bridging the Gap between Genetics and Ecology. *Annu Rev Phytopathol* 2011;**49**:345–67.

25. Melamed D, Young DL, Miller CR, Fields S. Combining natural sequence variation with high throughput mutational data to reveal protein interaction sites. *PLoS Genet* 2015;**11**:e1004918.

26. de Juan D, Pazos F, Valencia A. Emerging methods in protein co-evolution. *Nat Rev Genet* 2013;**14**:249–61.

27. Juan D. Desarrollo de métodos computacionales basados en co-evolución para la predicción de interacciones entre proteínas. 2015.

28. Thompson JN. ECOLOGY: Mutualistic Webs of Species. *Science* 2006;**312**:372–3.

29. Marcotte E, Pellegrini M, Ng H, Rice D. Detecting protein function and protein-protein interactions from genome sequences. *Science* 1999;**285**:751–4.

30. Gaasterland T, Ragan MA. Microbial Genescapes: Phyletic and Functional Patterns of ORF Distribution among Prokaryotes. *Microb Comp Genomics* 1998;**3**:199–217.

31. Fryxell, Karl. The coevolution of gene family trees. *Trends Genet* 1996;**12**:364–9.

32. Pazos F, Valencia A. Similarity of phylogenetic trees as indicator of protein–protein interaction. *Protein Eng* 2001;**14**:609–14.

33. Pazos F, Ranea JAG, Juan D, Sternberg MJE. Assessing Protein Co-evolution in the Context of the Tree of Life Assists in the Prediction of the Interactome. *J Mol Biol* 2005;**352**:1002–15.

34. Juan D, Pazos F, Valencia A. High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. *Proc Natl Acad Sci U S A* 2008;**105**:934–9.

35. Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA. Protein interaction maps for complete genomes based on gene fusion events. *Nature* 1999;**402**:86–90.

36. Eck RV, Dayhoff MO. Evolution of the Structure of Ferredoxin Based on Living Relics of Primitive Amino Acid Sequences. *Science* 1966;**152(3720)**:363–6.

37. Tang H, Thomas PD. Tools for Predicting the Functional Impact of Nonsynonymous Genetic Variation. *Genetics* 2016;**203**:635–47.

38. Altschuh D, Lesk AM, Bloomer AC, Klug A. Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. *J Mol Biol* 1987;**193**:693–707.

39. Altschuh D, Vernet T, Berti P, Moras D, Nagai K. Coordinated amino acid changes in homologous protein families. *Protein Eng Des Sel* 1988;**2**:193–199.

40. Göbel U, Sander C, Schneider R, Valencia A. Correlated mutations and residue contacts in proteins. *Proteins* 1994;**18**:309–17.

41. Neher E. How frequent are correlated changes in families of protein sequences? *Proc Natl Acad Sci* 1994;**91**:98–102.

42. Shindyalov IN, Kolchanov NA, Sander C. Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Protein Eng* 1994;**7**:349–58.

43. Olmea O, Rost B, Valencia A. Effective use of sequence correlation and conservation in fold recognition11Edited by J. M. Thornton. *J Mol Biol* 1999;**293**:1221–39.

44. Ortiz AR, Kolinski A, Rotkiewicz P, Ilkowski B, Skolnick J. Ab initio folding of proteins using restraints derived from evolutionary information. *Proteins* 1999;**Suppl 3**:177–85.

45. Ivankov DN, Finkelstein AV, Kondrashov F a. A structural perspective of compensatory evolution. *Curr Opin Struct Biol* 2014;**26**:104–12.

46. Talavera D, Lovell SC, Whelan S. Covariation is a poor measure of molecular coevolution. *Mol Biol Evol* 2015;**32**:msv109-.

47. Hakes L, Lovell S. Specificity in protein interactions and its relationship with sequence diversity and coevolution. *Proc Natl Acad Sci* 2007;**104**:7999–8004.

48. Englander SW, Mayne L. The nature of protein folding pathways. *Proc Natl Acad Sci* 2014;**111**:15873–80.

49. Anfinsen CB. Principles that govern the folding of protein chains. *Science* 1973;**181**:223–230.

50. Venclovas C, Zemla A, Fidelis K, Moult J. Assessment of progress over the CASP experiments. *Proteins Struct Funct Bioinforma* 2003;**53**:585–95.

51. Moult J. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Curr Opin Struct Biol* 2005;**15**:285–9.

52. Bonneau R, Tsai J, Ruczinski I, Chivian D, Rohl C, Strauss CE, Baker D. Rosetta in CASP4: progress in ab initio protein structure prediction. *Proteins* 2001;**Suppl 5**:119–26.

53. Schaarschmidt J, Monastyrskyy B, Kryshtafovych A, Bonvin AMJJ. Assessment of contact predictions in CASP12: Co-evolution and deep learning coming of age. *Proteins Struct Funct Bioinforma* 2018;**86**:51–66.

54. Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A. Critical assessment of methods of protein structure prediction (CASP)—Round XII. *Proteins Struct Funct Bioinforma* 2018;**86**:7–15.

55. Rost B. Twilight zone of protein sequence alignments. *Protein Eng* 1999;**12**:85–94.

56. Baker D, Sali A. Protein structure prediction and structural genomics. *Science* 2001.

57. Martı-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F. Comparative Protein Structure Modeling of Genes and Genomes. 2000:37.

58. Sanchez R, Sali A. Large-scale protein structure modeling of the Saccharomyces cerevisiae genome. *Proc Natl Acad Sci* 1998;**95**:13597–602.

59. Koehl P, Levitt M. A brighter future for protein structure prediction. 1999;**6**:5.

60. Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A. Critical assessment of methods of protein structure prediction: Progress and new directions in round XI: Progress in CASP XI. *Proteins Struct Funct Bioinforma* 2016;**84**:4–14.

61. Kryshtafovych A, Monastyrskyy B, Fidelis K, Moult J, Schwede T, Tramontano A. Evaluation of the template-based modeling in CASP12. *Proteins Struct Funct Bioinforma* 2018;**86**:321–34.

62. Abriata LA, Tamò GE, Monastyrskyy B, Kryshtafovych A, Dal Peraro M. Assessment of hard target modeling in CASP12 reveals an emerging role of alignment-based contact prediction methods. *Proteins Struct Funct Bioinforma* 2018;**86**:97–112.

63. Monastyrskyy B, D'Andrea D, Fidelis K, Tramontano A, Kryshtafovych A. New encouraging developments in contact prediction: Assessment of the CASP11 results. *Proteins Struct Funct Bioinforma* 2015:1–14.

64. Vendruscolo M, Kussell E, Domany E. Recovery of protein structure from contact maps. *Fold Des* 1997;**2**:295–306.

65. Bohr J, Bohr H, Brunak S, Cotterill RM, Fredholm H, Lautrup B, Petersen SB. Protein structures from distance inequalities. *J Mol Biol* 1993;**231**:861–869.

66. Zhang Y, Arakaki AK, Skolnick J. TASSER: An automated method for the prediction of protein tertiary structures in CASP6. *Proteins Struct Funct Bioinforma* 2005;**61**:91–8.

67. Ezkurdia I, Graña O, Izarzugaza JMG, Tress ML. Assessment of domain boundary predictions and the prediction of intramolecular contacts in CASP8. *Proteins Struct Funct Bioinforma* 2009;**77**:196–209.

68. Ortiz AR, Kolinski A, Skolnick J. Combined multiple sequence reduced protein model approach to predict the terciary structure of small proteins. :12.

69. Ortiz AR, Kolinski A, Skolnick J. Fold assembly of small proteins using Monte Carlo simulations driven by restraints derived from multiple sequence alignments. *J Mol Biol* 1998;**277**:419–48.

70. Ortiz AR, Kolinski A, Skolnick J. Nativelike topology assembly of small proteins using predicted restraints in Monte Carlo folding simulations. *Proc Natl Acad Sci* 1998;**95**:1020–5.

71. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T. Identification of direct residue contacts in protein-protein interaction by message passing. *Proc Natl Acad Sci U S A* 2009;**106**:67–72.

72. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, Weigt M. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci U S A* 2011;**108**:E1293–301.

73. Marks DS, Colwell LJ, Sheridan R, Hopf T a, Pagnani A, Zecchina R, Sander C. Protein 3D structure computed from evolutionary sequence variation. *PloS One* 2011;**6**:e28766–e28766.

74. Hopf TA, Colwell LJ, Sheridan R, Rost B, Sander C, Marks DS. Three-dimensional structures of membrane proteins from genomic sequencing. *Cell* 2012;**149**:1607–21.

75. Fodor A a, Aldrich RW. Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. *Proteins* 2004;**56**:211–21.

76. Olmea O, Valencia A. Improving contact predictions by the combination of correlated mutations and other sources of sequence information. *Fold Des* 1997;**2**:S25-32.

77. Fares MA. A Novel Method for Detecting Intramolecular Coevolution: Adding a Further Dimension to Selective Constraints Analyses. *Genetics* 2006;**173**:9–23.

78. Korber BT, Farber RM, Wolpert DH, Lapedes a S. Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. *Proc Natl Acad Sci U S A* 1993;**90**:7176–80.

79. Dunn SD, Wahl LM, Gloor GB. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinforma Oxf Engl* 2008;**24**:333–40.

80. Tillier ERM, Lui TWH. Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments. *Bioinformatics* 2003;**19**:750–5.

81. Wollenberg KR, Atchley WR. Separation of phylogenetic and functional associations in biological sequences by using the parametric bootstrap. :4.

82. Casari G, Sander C, Valencia A. A method to predict functional residues in proteins. *Nat Struct Biol* 1995;**2**:171–8.

83. Lichtarge O, Bourne HR, Cohen FE. An Evolutionary Trace Method Defines Binding Surfaces Common to Protein Families. *J Mol Biol* 1996;**257**:342–58.

84. Buslje CM, Santos J, Delfino JM, Nielsen M. Correction for phylogeny, small number of observations and data redundancy improves the identification of coevolving amino acid pairs using mutual information. *Bioinformatics* 2009;**25**:1125–31.

85. Ortiz AR, Skolnick J. Sequence Evolution and the Mechanism of Protein Folding. *Biophys J* 2000;**79**:1787–99.

86. 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**:1995–2000.

87. Fodor A a, Aldrich RW. Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. *Proteins* 2004;**56**:211–21.

88. Yeang CH, Haussler D. Detecting coevolution in and among protein domains. *PLoS Comput Biol* 2007;**3**:2122–34.

89. Dutheil J, Pupko T, Jean-Marie A, Galtier N. A Model-Based Approach for Detecting Coevolving Positions in a Molecule. *Mol Biol Evol* 2005;**22**:1919–28.

90. Pollock DD, Taylor WR, Goldman N. Coevolving protein residues: maximum likelihood identification and relationship to structure1. *J Mol Biol* 1999;**287**:187–198.

91. Barker D, Pagel M. Predicting Functional Gene Links from Phylogenetic-Statistical Analyses of Whole Genomes. *PLoS Comput Biol* 2005;**1**:e3.

92. Pollock DD, Taylor WR. Effectiveness of correlation analysis in identifying protein residues undergoing correlated evolution. *Protein Eng Des Sel* 1997;**10**:647–57.

93. Tuffery P, Darlu P. Exploring a Phylogenetic Approach for the Detection of Correlated Substitutions in Proteins. *Mol Biol Evol* 2000;**17**:1753–9.

94. Fariselli P, Olmea O, Valencia A, Casadio R. Prediction of contact maps with neural networks and correlated mutations. *Protein Eng Des Sel* 2001;**14**:835–43.

95. Hamilton N, Burrage K, Ragan MA, Huber T. Protein contact prediction using patterns of correlation. *Proteins Struct Funct Bioinforma* 2004;**56**:679–84.

96. Fariselli P, Casadio R. A neural network based predictor of residue contacts in proteins. *Protein Eng Des Sel* 1999;**12**:15–21.

97. Cheng J, Baldi P. Improved residue contact prediction using support vector machines and a large feature set. *BMC Bioinformatics* 2007;**8**:113–113.

98. Li Y, Fang Y, Fang J. Predicting residue–residue contacts using random forest models. *Bioinformatics* 2011;**27**:3379–84.

99. Punta M, Rost B. PROFcon: novel prediction of long-range contacts. *Bioinformatics* 2005;**21**:2960–8.

100. Sborgi L, Verma A, Piana S, Lindorff-Larsen K, Cerminara M, Santiveri CM, Shaw DE, de Alba E, Muñoz V. Interaction Networks in Protein Folding via Atomic-Resolution Experiments and Long-Time-Scale Molecular Dynamics Simulations. *J Am Chem Soc* 2015;**137**:6506–16.

101. Stanley HE. Phase transition and critical phenomena. :5.

102. Giraud BG, Heumann JM, Lapedes a S. Superadditive correlation. Preprint. *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Top* 1999;**59**:4983–91.

103. Lapedes AS, Giraud BG, And LL, Stormo GD. Correlated Mutations in Models of Protein Sequences: Phylogenetic and Structural Effects. *Stat Mol Biol* 1999;**33**:236–56.

104. Lapedes A, Giraud B, Jarzynski C. Using Sequence Alignments to Predict Protein Structure and Stability With High Accuracy. *arXiv* 2002:1207.2484-1207.2484.

105. Burger L, van Nimwegen E. Disentangling direct from indirect co-evolution of residues in protein alignments. *PLoS Comput Biol* 2010;**6**:e1000633–e1000633.

106. Balakrishnan S, Kamisetty H, Carbonell JG, Lee SI, Langmead CJ. Learning generative models for protein fold families. *Proteins* 2011;**79**:1061–78.

107. Jaynes E. Information theory and statistical mechanics. *Phys Rev* 1957.

108. Jaynes E. Information theory and statistical mechanics. II.

109. Ekeberg M, Lövkvist C, Lan Y, Weigt M, Aurell E. Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. *Phys Rev E - Stat Nonlinear Soft Matter Phys* 2013;**87**:1–16.

110. Ekeberg M, Hartonen T, Aurell E. Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences. *J Comput Phys* 2014;**276**:341–56.

111. Barton JP, De Leonardis E, Coucke A, Cocco S. ACE: adaptive cluster expansion for maximum entropy graphical model inference. *Bioinformatics* 2016;**32**:3089–97.

112. Sutto L, Marsili S, Valencia A, Gervasio FL. From residue coevolution to protein conformational ensembles and functional dynamics. *Proc Natl Acad Sci U S A* 2015;**112**:13567–72.

113. Figliuzzi M, Barrat-Charlaix P, Weigt M. How Pairwise Coevolutionary Models Capture the Collective Residue Variability in Proteins? *Mol Biol Evol* 2018;**35**:1018–27.

114. Cocco S, Feinauer C, Figliuzzi M, Monasson R, Weigt M. Inverse Statistical Physics of Protein Sequences: A Key Issues Review. 2017, DOI: 10.1088/1361-6633/aa9965.

115. Jones DT, Buchan DW, Cozzetto D, Pontil M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinforma Oxf Engl* 2012;**28**:184–90.

116. Vorberg S, Seemayer S, Söding J. Synthetic protein alignments by CCMgen quantify noise in residue-residue contact prediction. Wallner B (ed.). *PLOS Comput Biol* 2018;**14**:e1006526.

117. 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**:15018–23.

118. Skwark MJ, Abdel-Rehim A, Elofsson A. PconsC: Combination of direct information methods and alignments improves contact prediction. *Bioinforma Oxf Engl* 2013:1–2.

119. Jones DT, Singh T, Kosciolek T, Tetchner S. MetaPSICOV: Combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics* 2015;**31**:999–1006.

120. Wang S, Sun S, Li Z, Zhang R, Xu J. Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. *PLOS Comput Biol* 2017:34.

121. Tang Y, Huang YJ, Hopf T a, Sander C, Marks DS, Montelione GT. Protein structure determination by combining sparse NMR data with evolutionary couplings. *Nat Methods* 2015;**12**:751–4.

122. Meyer MJ, Beltrán JF, Liang S, Fragoza R, Rumack A, Liang J, Wei X, Yu H. Interactome INSIDER: a structural interactome browser for genomic studies. *Nat Methods* 2018;**15**:107–114.

123. Di Lena P, Nagata K, Baldi P. Deep architectures for protein contact map prediction. *Bioinforma Oxf Engl* 2012;**28**:2449–57.

124. Jones DT, Kandathil SM. High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features. Valencia A (ed.). *Bioinformatics* 2018;**34**:3308–15.

125. Kukic P, Mirabello C, Tradigo G, Walsh I, Veltri P, Pollastri G. Toward an accurate prediction of inter-residue distances in proteins using 2D recursive neural networks. *BMC Bioinformatics* 2014;**15**, DOI: 10.1186/1471-2105-15-6.

126. Walsh I, Baù D, Martin AJ, Mooney C, Vullo A, Pollastri G. Ab initio and template-based prediction of multi-class distance maps by two-dimensional recursive neural networks. *BMC Struct Biol* 2009;**9**:5.

127. Ji S. DeepCDpred: Inter-residue distance and contact prediction for improved prediction of protein structure. :15.

128. Xu J. Distance-based Protein Folding Powered by Deep Learning. :16.

129. Kassem MM, Christoffersen LB, Cavalli A, Lindorff-Larsen K. Enhancing coevolution-based contact prediction by imposing structural self-consistency of the contacts. *Sci Rep* 2018;**8**, DOI: 10.1038/s41598-018-29357-y.

130. Jacob E, Unger R, Horovitz A. Codon-level information improves predictions of inter-residue contacts in proteins by correlated mutation analysis. :14.

131. dos Santos RN, Ferrari AJR, de Jesus HCR, Gozzo FC, Morcos F, Martínez L. Enhancing protein fold determination by exploring the complementary information of chemical cross-linking and coevolutionary signals. Valencia A (ed.). *Bioinformatics* 2018;**34**:2201–8.

132. AlQuraishi M. AlphaFold at CASP13. Valencia A (ed.). *Bioinformatics* 2019, DOI: 10.1093/bioinformatics/btz422.

133. Morcos F, Jana B, Hwa T, Onuchic JN. Coevolutionary signals across protein lineages help capture multiple protein conformations. *Proc Natl Acad Sci U S A* 2013;**110**:20533–8.

134. Sfriso P, Duran-Frigola M, Mosca R, Emperador A, Aloy P, Orozco M. Residues Coevolution Guides the Systematic Identification of Alternative Functional Conformations in Proteins. *Structure* 2016;**24**, DOI: 10.1016/j.str.2015.10.025.

135. Uguzzoni G, John Lovis S, Oteri F, Schug A, Szurmant H, Weigt M. Large-scale identification of coevolution signals across homo-oligomeric protein interfaces by direct coupling analysis. *Proc Natl Acad Sci* 2017;**114**:E2662–71.

136. dos Santos RN, Morcos F, Jana B, Andricopulo AD, Onuchic JN. Dimeric interactions and complex formation using direct coevolutionary couplings. *Sci Rep* 2015;**5**:13652–13652.

137. Morcos F, Schafer NP, Cheng RR, Onuchic JN, Wolynes PG. Coevolutionary information, protein folding landscapes, and the thermodynamics of natural selection. *Proc Natl Acad Sci U S A* 2014:1–6.

138. Noel JK, Morcos F, Onuchic JN. Sequence co-evolutionary information is a natural partner to minimally-frustrated models of biomolecular dynamics. *F1000Research* 2016;**5**:106.

139. Wozniak PP, Vriend G, Kotulska M. Correlated mutations select misfolded from properly folded proteins. Valencia A (ed.). *Bioinformatics* 2017;**33**:1497–504.

140. Toth-Petroczy A, Palmedo P, Ingraham J, Hopf TA, Berger B, Sander C, Marks DS. Structured States of Disordered Proteins from Genomic Sequences. *Cell* 2016;**167**:158–170.e12.

141. Weinreb C, Riesselman AJ, Ingraham JB, Gross T, Sander C, Marks DS. 3D RNA and Functional Interactions from Evolutionary Couplings. *Cell* 2016;**165**:963–75.

142. De Leonardis E, Lutz B, Ratz S, Cocco S, Monasson R, Schug A, Weigt M. Direct-Coupling Analysis of nucleotide coevolution facilitates RNA secondary and tertiary structure prediction. *Nucleic Acids Res* 2015:gkv932–gkv932.

143. Taylor WR, Hamilton RS. Exploring RNA conformational space under sparse distance restraints. *Sci Rep* 2017;**7**, DOI: 10.1038/srep44074.

144. Ovchinnikov S, Kamisetty H, Baker D. Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *eLife* 2014;**3**:e02030–e02030.

145. Hopf TA, Schärfe CPI, Rodrigues JPGLM, Green AG, Kohlbacher O, Sander C, Bonvin AMJJ, Marks DS. Sequence co-evolution gives 3D contacts and structures of protein complexes. *eLife* 2014;**3**:e03430–e03430.

146. Figliuzzi M, Jacquier H, Schug A, Tenaillon O, Weigt M. Coevolutionary landscape inference and the context-dependence of mutations in beta-lactamase TEM-1. *Mol Biol Evol* 2015;**33**:msv211–msv211.

147. Hopf TA, Ingraham JB, Poelwijk FJ, Schärfe CPI, Springer M, Sander C, Marks DS. Mutation effects predicted from sequence co-variation. *Nat Biotechnol* 2017;**2017**, DOI: 10.1038/nbt.3769.

148. Riesselman AJ, Ingraham JB, Marks DS. Deep generative models of genetic variation capture the effects of mutations. *Nat Methods* 2018, DOI: 10.1038/s41592-018-0138-4.

149. Figliuzzi M, Barrat-Charlaix P, Weigt M. how pairwise coevolutionary models capture the collective residue variability in proteins. *Mol Biol Evol* 2017:1–24.

150. Haldane A, Flynn WF, He P, Levy RM. Coevolutionary Landscape of Kinase Family Proteins: Sequence Probabilities and Functional Motifs. *Biophys J* 2018;**114**:21–31.

151. Kamisetty H, Ovchinnikov S, Baker D. Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proc Natl Acad Sci U S A* 2013;**110**:15674–9.

152. Avila-Herrera A, Pollard KS. Coevolutionary analyses require phylogenetically deep alignments and better null models to accurately detect inter-protein contacts within and between species. *BMC Bioinformatics* 2015;**16**:268–268.

153. Ovchinnikov S, Park H, Varghese N, Huang P-S, Pavlopoulos GA, Kim DE, Kamisetty H, Kyrpides NC, Baker D. Protein structure determination using metagenome sequence data. *Science* 2017;**355**:294–8.

154. Aloy P, Ceulemans H, Stark A, Russell RB. The Relationship Between Sequence and Interaction Divergence in Proteins. *J Mol Biol* 2003;**332**:989–98.

155. Mosca R, Céol A, Aloy P. Interactome3D: adding structural details to protein networks. *Nat Methods* 2013;**10**:47–53.

156. Tuncbag N, Gursoy A, Nussinov R, Keskin O. Predicting protein-protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using PRISM. *Nat Protoc* 2011;**6**:1341–54.

157. Negroni J, Mosca R, Aloy P. Assessing the Applicability of Template-Based Protein Docking in the Twilight Zone. *Structure* 2014;**22**:1356–62.

158. Hwang H, Vreven T, Janin J, Weng Z. Protein-protein docking benchmark version 4.0. *Proteins Struct Funct Bioinforma* 2010;**78**:3111–4.

159. Smith GR, Sternberg MJE. Prediction of protein–protein interactions by docking methods. *Curr Opin Struct Biol* 2002;**12**:28–35.

160. Eisenstein M, Katchalski-Katzir E. On proteins, grids, correlations, and docking. *C R Biol* 2004;**327**:409–20.

161. Ruvinsky AM, Kirys T, Tuzikov AV, Vakser IA. Side-Chain Conformational Changes upon Protein–Protein Association. *J Mol Biol* 2011;**408**:356–65.

162. Anishchenko I, Kundrotas PJ, Tuzikov AV, Vakser IA. Protein models: The Grand Challenge of protein docking: Protein Models for Docking. *Proteins Struct Funct Bioinforma* 2014;**82**:278–87.

163. Russel D, Lasker K, Webb B, Vel??zquez-Muriel J, Tjioe E, Schneidman-Duhovny D, Peterson B, Sali A. Putting the pieces together: Integrative modeling platform software for structure determination of macromolecular assemblies. *PLoS Biol* 2012;**10**:e1001244–e1001244.

164. Van Zundert GCP, Rodrigues JPGLM, Trellet M, Schmitz C, Kastritis PL, Karaca E, Melquiond ASJ, Van Dijk M, De Vries SJ, Bonvin AMJJ. The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *J Mol Biol* 2016;**428**:720–5.

165. Herman D, Ochoa D, Juan D, Lopez D, Valencia A, Pazos F. Selection of organisms for the co-evolution-based study of protein interactions. *BMC Bioinformatics* 2011;**12**:363–363.

166. Cong Q, Anishchenko I, Ovchinnikov S, Baker D. Protein interaction networks revealed by proteome coevolution. 2019:6.

167. Iserte J, Simonetti FL, Zea DJ, Teppa E, Marino-Buslje C. I-COMS: Interprotein-COrrelated Mutations Server. *Nucleic Acids Res* 2015:1–6.

168. Zhou T, Wang S, Xu J. Deep learning reveals many more inter-protein residue-residue contacts than direct coupling analysis. *bioRxiv* 2018, DOI: 10.1101/240754.

169. Izarzugaza JMG, Juan D, Pons C, Ranea JAG, Valencia A, Pazos F. TSEMA: interactive prediction of protein pairings between interacting families. *Nucleic Acids Res* 2006;**34**:W315–9.

170. Izarzugaza JMG, Juan D, Pons C, Pazos F, Valencia A. Enhancing the prediction of protein pairings between interacting families using orthology information. *BMC Bioinformatics* 2008;**9**:35–35.

171. Gueudré T, Baldassi C, Zamparo M, Weigt M, Pagnani A. Simultaneous identification of specifically interacting paralogs and inter-protein contacts by Direct-Coupling Analysis. *Proc Natl Acad Sci* 2016:10.1073/pnas.1607570113-10.1073/pnas.1607570113.

172. Bitbol A-F, Dwyer RS, Colwell LJ, Wingreen NS. Inferring interaction partners from protein sequences. *Proc Natl Acad Sci* 2016:10.1073/pnas.1606762113-10.1073/pnas.1606762113.

173. Correa Marrero M, Immink RGH, de Ridder D, van Dijk ADJ. Improved inference of intermolecular contacts through protein–protein interaction prediction using coevolutionary analysis. Hancock J (ed.). *Bioinformatics* 2019;**35**:2036–42.

174. Bitbol A-F. Inferring interaction partners from protein sequences using mutual information. *PLOS Comput Biol* 2018;**14**:1–24.

175. Pazos F, Helmer-Citterich M, Ausiello G, Valencia A. Correlated mutations contain information about protein-protein interaction. *J Mol Biol* 1997;**271**:511–23.

176. Halperin I, Wolfson H, Nussinov R. Correlated mutations: Advances and limitations. A study on fusion proteins and on the Cohesin-Dockerin families. *Proteins Struct Funct Bioinforma* 2006;**63**:832–45.

177. Khan S, Guo TW, Misra S. A coevolution-guided model for the rotor of the bacterial flagellar motor. *Sci Rep* 2018;**8**:1–13.

178. Tamir S, Rotem-Bamberger S, Katz C, Morcos F, Hailey KL, Zuris J a, Wang C, Conlan AR, Lipper CH, Paddock ML, Mittler R, Onuchic JN, Jennings P a, Friedler A, Nechushtai R. Integrated strategy reveals the protein interface between cancer targets Bcl-2 and NAF-1. *Proc Natl Acad Sci U S A* 2014;**111**:5177–82.

179. Noivirt O, Eisenstein M, Horovitz A. Detection and reduction of evolutionary noise in correlated mutation analysis. *Protein Eng Des Sel* 2005;**18**:247–53.

180. Mao W, Kaya C, Dutta A, Horovitz A, Bahar I. Comparative study of the effectiveness and limitations of current methods for detecting sequence coevolution. *Bioinformatics* 2015;**31**:1929–37.

181. Bursteinas B, Britto R, Bely B, Auchincloss A, Rivoire C, Redaschi N, O'Donovan C, Martin MJ. Minimizing proteome redundancy in the UniProt Knowledgebase. *Database* 2016;**2016**:baw139.

182. Mosca R, Céol A, Stein A, Olivella R, Aloy P. 3did: a catalog of domain-based interactions of known three-dimensional structure. *Nucleic Acids Res* 2013;**42**:D374–9.

183. Stein A, Céol A, Aloy P. 3Did: Identification and Classification of Domain-Based Interactions of Known Three-Dimensional Structure. *Nucleic Acids Res* 2011;**39**:D718-23.

184. Stein A, Russell RB, Aloy P. 3Did: Interacting Protein Domains of Known Three-Dimensional Structure. *Nucleic Acids Res* 2005;**33**:D413-7.

185. Stein A, Panjkovich A, Aloy P. 3did Update: domain-domain and peptide-mediated interactions of known 3D structure. *Nucleic Acids Res* 2009;**37**:D300-4.

186. Berman H, Henrick K, Nakamura H. Announcing the worldwide Protein Data Bank. *Nat Struct Biol* 2003;**10**:980–980.

187. Yellaboina S, Tasneem A, Zaykin DV, Raghavachari B, Jothi R. DOMINE: a comprehensive collection of known and predicted domain-domain interactions. *Nucleic Acids Res* 2011;**39**:D730–5.

188. Segura J, Sorzano COS, Cuenca-Alba J, Aloy P, Carazo JM. Using neighborhood cohesiveness to infer interactions between protein domains. *Bioinformatics* 2015;**31**:2545–52.

189. Sengupta U, Ukil S, Dimitrova N, Agrawal S. Identification of altered regulatory pathways in diabetes type II and complications through expression networks. *2009 IEEE Int Workshop Genomic Signal Process Stat*.

190. Chen YF, Xia Y. Convergent perturbation of the human domain-resolved interactome by viruses and mutations inducing similar disease phenotypes. *PLOS Comput Biol* 2019;**15**:e1006762.

191. García-Pérez CA, Guo X, Navarro JG, Aguilar DAG, Lara-Ramírez EE. Proteome-wide analysis of human motif-domain interactions mapped on influenza a virus. *BMC Bioinformatics* 2018;**19**, DOI: 10.1186/s12859-018-2237-8.

192. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer ELL, Tate J, Punta M. Pfam: The protein families database. *Nucleic Acids Res* 2014;**42**:D222–30.

193. Eddy SR. HMMER User's Guide. *Hmmer Man* 2010:0–93.

194. Kersey PJ, Allen JE, Christensen M *et al.* Ensembl Genomes 2013: Scaling up access to genome-wide data. *Nucleic Acids Res* 2014;**42**:D546–52.

195. Velankar S, Dana JM, Jacobsen J, Van Ginkel G, Gane PJ, Luo J, Oldfield TJ, O'Donovan C, Martin MJ, Kleywegt GJ. SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. *Nucleic Acids Res* 2012;**41**:D483–9.

196. Sayers EW, Barrett T, Benson DA *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2011;**37**:D5–15.

197. Bairoch A, Apweiler R. The SWISS-PROT protein sequence data bank and its new supplement TREMBL. *Nucleic Acids Res* 1997;**25**:31–6.

198. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* 1997;**25**:3389–402.

199. Li W, Jaroszewski L, Godzik A. Tolerating some redundancy significantly speeds up clustering of large protein databases. *Bioinforma Oxf Engl* 2002;**18**:77–82.

200. Remmert M, Hauser A. HH-suite for sensitive sequence searching based on HMM-HMM alignment. 2012:951–60.

201. Ravikumar P, Wainwright MJ, Lafferty JD. High-dimensional Ising model selection using ℓ1-regularized logistic regression. *Ann Stat* 2010;**38**:1287–319.

202. Feinauer C, Skwark MJ, Pagnani A, Aurell E. Improving contact prediction along three dimensions. *PLoS Comput Biol* 2014;**10**:e1003847–e1003847.

203. Chatr-Aryamontri A, Breitkreutz BJ, Oughtred R *et al.* The BioGRID interaction database: 2015 update. *Nucleic Acids Res* 2015;**43**:D470–8.

204. Rodriguez JM, Carro A, Valencia A, Tress ML. APPRIS WebServer and WebServices. *Nucleic Acids Res* 2015;**43**:W455–9.

205. Huerta-Cepas J, Capella-Gutiérrez S, Pryszcz LP, Marcet-Houben M, Gabaldón T. PhylomeDB v4: Zooming into the plurality of evolutionary histories of a genome. *Nucleic Acids Res* 2014;**42**:897–902.

206. Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* 2015;**32**:268–74.

207. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* 1992;**89**:10915–9.

208. Lê S, Josse J, Husson F. FactoMineR: An R Package for Multivariate Analysis. *J Stat Softw* 2008;**25**, DOI: 10.18637/jss.v025.i01.

209. Kato M, Wynn RM, Chuang JL, Tso SC, Machius M, Li J, Chuang DT. Structural basis for inactivation of the human pyruvate dehydrogenase complex by phosphorylation: role of disordered phosphorylation loops. *Structure* 2008;**16**:1849–59.

210. Nakai T, Nakagawa N, Maoka N, Masui R, Kuramitsu S, Kamiya N. Ligand-induced conformational changes and a reaction intermediate in branched-chain 2-oxo acid dehydrogenase (E1) from Thermus thermophilus HB8, as revealed by x-ray crystallography. *J Mol Biol* 2004;**337**:1011–33.

211. Tsukazaki T, Mori H, Fukai S, Ishitani R, Mori T, Dohmae N, Perederina A, Sugita Y, Vassylyev DG, Ito K, Nureki O. Conformational transition of Sec machinery inferred from bacterial SecYE structures. *Nature* 2008;**455**:988–91.

212. Gogala M, Becker T, Beatrix B, Armache J-P, Barrio-Garcia C, Berninghausen O, Beckmann R. Structures of the Sec61 complex engaged in nascent peptide translocation or membrane insertion. *Nature* 2014;**506**:107–10.

213. Finarov I, Moor N, Kessler N, Klipcan L, Safro MG. Structure of human cytosolic phenylalanyl-tRNA synthetase: evidence for kingdom-specific design of the active sites and tRNA binding patterns. *Structure* 2010;**18**:343–53.

214. Efremov RG, Sazanov LA. Structure of the membrane domain of respiratory complex I. *Nature* 2011;**476**:414–20.

215. Baradaran R, Berrisford JM, Minhas GS, Sazanov LA. Crystal structure of the entire respiratory complex I. *Nature* 2013;**494**:443–8.

216. Pelé J, Bécu J-M, Abdi H, Chabbert M. Bios2mds: an R package for comparing orthologous protein families by metric multidimensional scaling. *BMC Bioinformatics* 2012;**13**:133.

217. Kim S. ppcor: An R Package for a Fast Calculation to Semi-partial Correlation Coefficients. *Commun Stat Appl Methods* 2015;**22**:665–74.

218. Halabi N, Rivoire O, Leibler S, Ranganathan R. Protein sectors: evolutionary units of three-dimensional structure. *Cell* 2009;**138**:774–86.

219. Galdadas I, Lovera S, Pérez-Hernández G, Barnes MD, Healy J, Afsharikho H, Woodford N, Bonomo RA, Gervasio FL, Haider S. Defining the architecture of KPC-2 Carbapenemase: identifying allosteric networks to fight antibiotics resistance. *Sci Rep* 2018;**8**:12916.

220. Keskin O, Gursoy A, Ma B, Nussinov R. Principles of Protein − Protein Interactions : What are the Preferred Ways For Proteins To Interact ? 2008;**108**:1225–44.

221. Shim Choi S, Li W, Lahn BT. Robust signals of coevolution of interacting residues in mammalian proteomes identified by phylogeny-aided structural analysis. *Nat Genet* 2005;**37**:1367–71.

222. Andreani J, Guerois R. Evolution of protein interactions: From interactomes to interfaces. *Arch Biochem Biophys* 2014;**554**:65–75.

223. Aakre CD, Herrou J, Phung TN, Perchuk BS, Crosson S, Laub MT. Evolving New Protein-Protein Interaction Specificity through Promiscuous Intermediates. *Cell* 2015;**163**:594–606.

224. Zea DJ, Monzon AM, Parisi G, Marino-Buslje C. How is structural divergence related to evolutionary information? *Mol Phylogenet Evol* 2018;**127**:859–66.

225. Teppa E, Zea DJ, Marino-Buslje C. Protein-protein interactions leave evolutionary footprints: High molecular coevolution at the core of interfaces. *Protein Sci* 2017;**26**:2438–44.

226. Anishchenko I, Ovchinnikov S, Kamisetty H, Baker D. Origins of coevolution between residues distant in protein 3D structures. *Proc Natl Acad Sci* 2017:201702664–201702664.

227. Capra JA, Singh M. Predicting functionally important residues from sequence conservation. *Bioinformatics* 2007;**23**:1875–82.

228. Karlin S, Brocchieri L. Evolutionary conservation of RecA genes in relation to protein structure and function. *J Bacteriol* 1996;**178**:1881–94.

229. Valdar WS, Thornton JM. Protein-protein interfaces: analysis of amino acid conservation in homodimers. *Proteins* 2001;**42**:108–24.

230. Lockless SW, Ranganathan R. Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families. 1999;**286**:295–9.

231. Qin C, Colwell LJ. Power law tails in phylogenetic systems. *Proc Natl Acad Sci U S A* 2018;**2017**:201711913.

232. van Nimwegen E. Inferring Contacting Residues within and between Proteins: What Do the Probabilities Mean? *PLOS Comput Biol* 2016;**12**:e1004726–e1004726.

233. Yanofsky C, Horn V, Thorpe D. Protein Structure Relationships Revealed by Mutational Analysis. *Science* 1964;**146**:1593–4.

234. Fitch WM, Markowitz E. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochem Genet* 1970;**4**:579–93.

235. Fitch WM. Rate of change of concomitantly variable codons. *J Mol Evol* 1971;**1**:84–96.

236. Goldstein RA, Pollard ST, Shah SD, Pollock DD. Nonadaptive Amino Acid Convergence Rates Decrease over Time. *Mol Biol Evol* 2015;**32**:1373–81.

237. de Visser JAGMM, Krug J. Empirical fitness landscapes and the predictability of evolution. *Nat Rev Genet* 2014;**15**:480–90.

238. Starr TN, Thornton JW. Epistasis in protein evolution. *Protein Sci* 2016;**22**:1204–18.

239. Podgornaia AI, Laub MT. Pervasive degeneracy and epistasis in a protein-protein interface. *Science* 2015;**347**:673–7.

240. Breen MS, Kemena C, Vlasov PK, Notredame C, Kondrashov F a. Epistasis as the primary factor in molecular evolution. *Nature* 2012;**490**:535–8.

241. Taverna DM, Goldstein RA. Why are proteins marginally stable? *Proteins Struct Funct Genet* 2002;**46**:105–9.

242. Pollock DD, Thiltgen G, Goldstein RA. Amino acid coevolution induces an evolutionary Stokes shift. *Proc Natl Acad Sci* 2012;**109**:E1352–9.

243. Usmanova DR, Ferretti L, Povolotskaya IS, Vlasov PK, Kondrashov FA. A Model of Substitution Trajectories in Sequence Space and Long-Term Protein Evolution. *Mol Biol Evol* 2015;**32**:542–54.

244. McCandlish DM, Shah P, Plotkin JB. Epistasis and the dynamics of reversion in molecular evolution. *Genetics* 2016;**203**:1335–51.

245. Pollock DD, Goldstein RA. Strong evidence for protein epistasis, weak evidence against it. *Proc Natl Acad Sci* 2014;**111**:E1450–E1450.

246. Ashenberg O, Gong LI, Bloom JD. Mutational effects on stability are largely conserved during protein evolution. *Proc Natl Acad Sci* 2013;**110**:21071–6.

247. Fariselli P, Casadio R. Prediction of the Number of Residue Contacts in Proteins. :6.

248. Wagner A. Neutralism and selectionism: a network-based reconciliation. *Nat Rev Genet* 2008;**9**:965–74.

249. Rollins NJ, Brock KP, Poelwijk FJ, Stiffler MA, Gauthier NP, Sander C, Marks DS. Inferring protein 3D structure from deep mutation scans. *Nat Genet* 2019;**51**:1170–6.

250. Fantini M, Lisi S, De Los Rios P, Cattaneo A, Pastore A. Protein structure without structure determination: direct coupling analysis based on in vitro evolution. *bioRxiv* 2019, DOI: 10.1101/582056.

251. Mintseris J, Weng Z. Structure, function, and evolution of transient and obligate protein-protein interactions. *Proc Natl Acad Sci U S A* 2005;**102**:10930–5.

252. Lunt B, Szurmant H, Procaccini A, Hoch J a, Hwa T, Weigt M. Inference of direct residue contacts in two-component signaling. 2010;**471**, DOI: 10.1016/S0076-6879(10)71002-8.

253. Cheng RR, Morcos F, Levine H, Onuchic JN. Toward rationally redesigning bacterial two-component signaling systems using coevolutionary information. *Proc Natl Acad Sci U S A* 2014;**111**:E563–71.

254. Stefani M. Protein misfolding and aggregation: new examples in medicine and biology of the dark side of the protein world. *Biochim Biophys Acta BBA - Mol Basis Dis* 2004;**1739**:5–25.

255. Levy ED, Boeri Erba E, Robinson CV, Teichmann SA. Assembly reflects evolution of protein complexes. *Nature* 2008;**453**:1262–5.

256. Ma B, Elkayam T, Wolfson H, Nussinov R. Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proc Natl Acad Sci U S A* 2003;**100**:5772–7.

257. de Oliveira S, Deane C. Co-evolution techniques are reshaping the way we do structural bioinformatics. *F1000Research* 2017;**6**:1224–1224.

ANEXO I. MATERIAL ADICIONAL

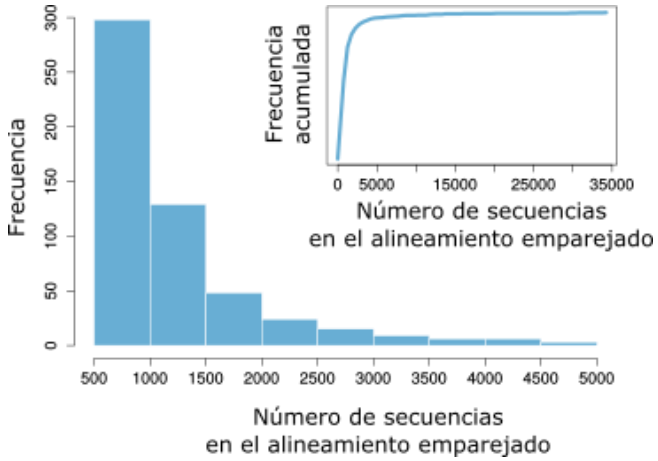


Figura A1. Distribución de la frecuencia del número de secuencias no redundantes (menos de un 80% de identidad en secuencia) en los alineamientos emparejados para un rango de 500 a 5 000 secuencias. En el gráfico interno, se muestra la distribución acumulada de la frecuencia para el rango completo del número de secuencias no redundantes en los alineamientos emparejados. Figura adaptada de [117].

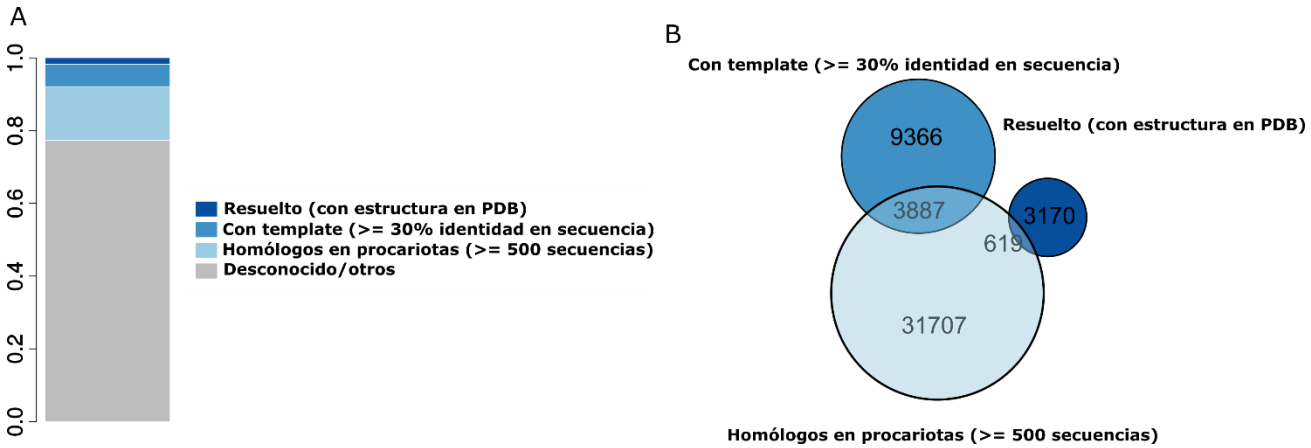


Figura A2. A) Proporción de pares de proteínas en el interactoma humano (214 806 interacciones heterodiméricas obtenidas desde BIOGRID) conteniendo interacciones entre dominios interproteína con estructura resuelta (azul oscuro), con *templates* disponibles con una identidad de secuencia superior al 30% (azul), con más de 500 secuencias homólogas no redundantes (80% de identidad en secuencia como máximo) en procariotas y sin información estructural ni template fiable (azul claro). El resto de interacciones se muestra en gris. B) Diagrama de Venn mostrando el solapamiento de los 3 conjuntos de interacciones descritos en el panel A. Figura adaptada de [117].

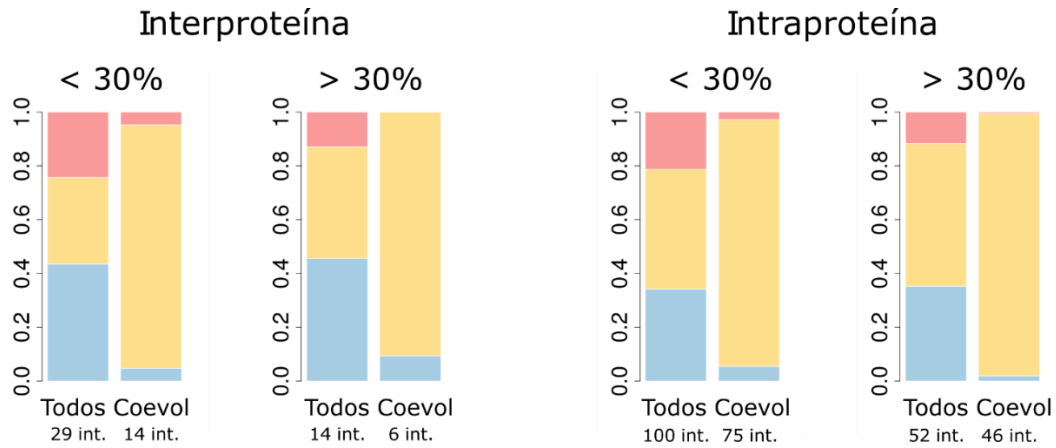


Figura A3. Proporción de contactos encontrados solo en procariotas, solo en eucariotas o conservados para todos los contactos y para el subconjunto de contactos en posiciones que han coevolucionado. Los casos han sido divididos en dos grupos dependiendo de si el porcentaje de identidad en secuencia es superior o inferior al 30% entre los complejos representativos en procariotas o eucariotas.

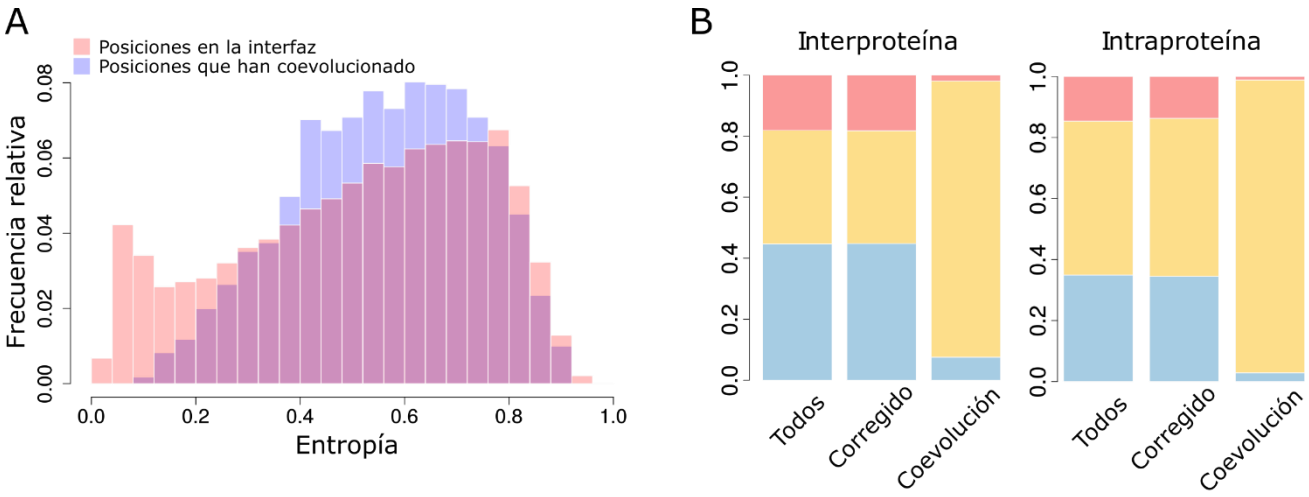


Figura A4. A) Distribución de la entropía de las posiciones en los alineamientos relativos a dos conjuntos de posiciones: posiciones en la interfaz (contactos interdominio, rojo) y posiciones que han coevolucionado (azul). La entropía es calculada para cada posición a partir de los alineamientos emparejados. B Proporción de contactos conservados considerando todos los contactos (Todos), solo el subconjunto de contactos en posiciones que han coevolucionado (coevolución) o para todos los contactos corregido de forma que tiene la misma distribución de entropía que el subconjunto que ha coevolucionado (Corregido). En azul, contactos en procariotas que no están en contacto en eucariotas; en rojo, contactos en eucariotas que no están en contacto es procariotas; en amarillo, contactos conservados, en contacto tanto en procariotas como en eucariotas. Figura adaptada de [117].

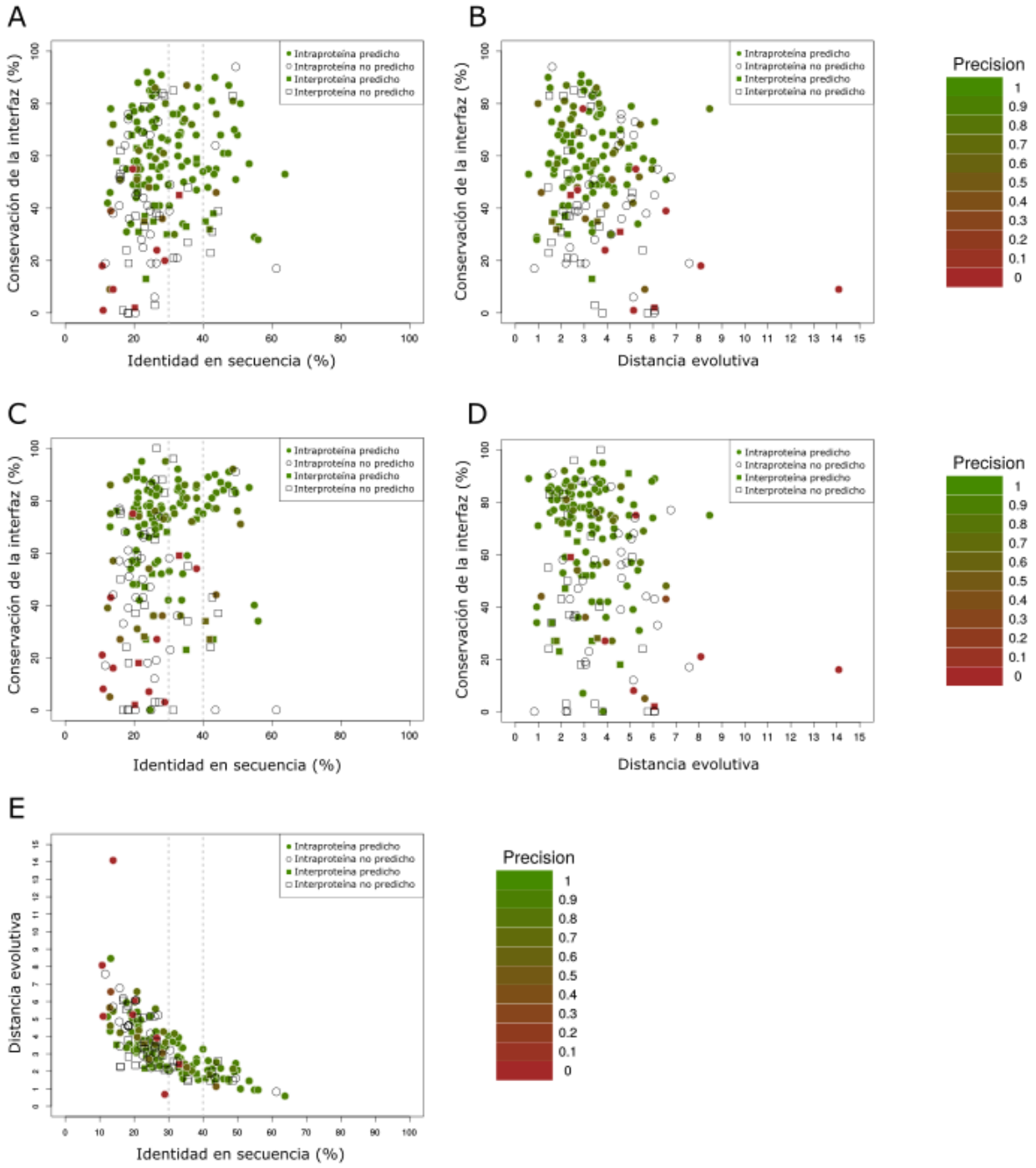


Figura A5. Comparación entre distancias evolutivas e identidades en secuencia como medidas de divergencia en secuencia y la conservación estructural de la interfaz. Cada punto representa una interacción dominio a dominio (interproteína como cuadradros, intraproteína como círculos) con estructura en procariotas y eucariotas, coloreado por según la precisión en la predicción de contactos (en blanco cuando no hay predicciones). Utilizando interfaces completas para la medida de conservación de la interfaz y la precisión, se muestra conservación de la interfaz en función de la identidad en secuencia (A) y conservación de la interfaz en función de la distancia evolutiva (B). Utilizando interfaces representativas para la medida de conservación de la interfaz y la precisión, se muestra, conservación de la interfaz en función de la identidad en secuencia (C) y conservación de la interfaz en función de la distancia evolutiva (D). Comparativa entre identidad en secuencia y distancia evolutiva (E). Figura adaptada de [117].

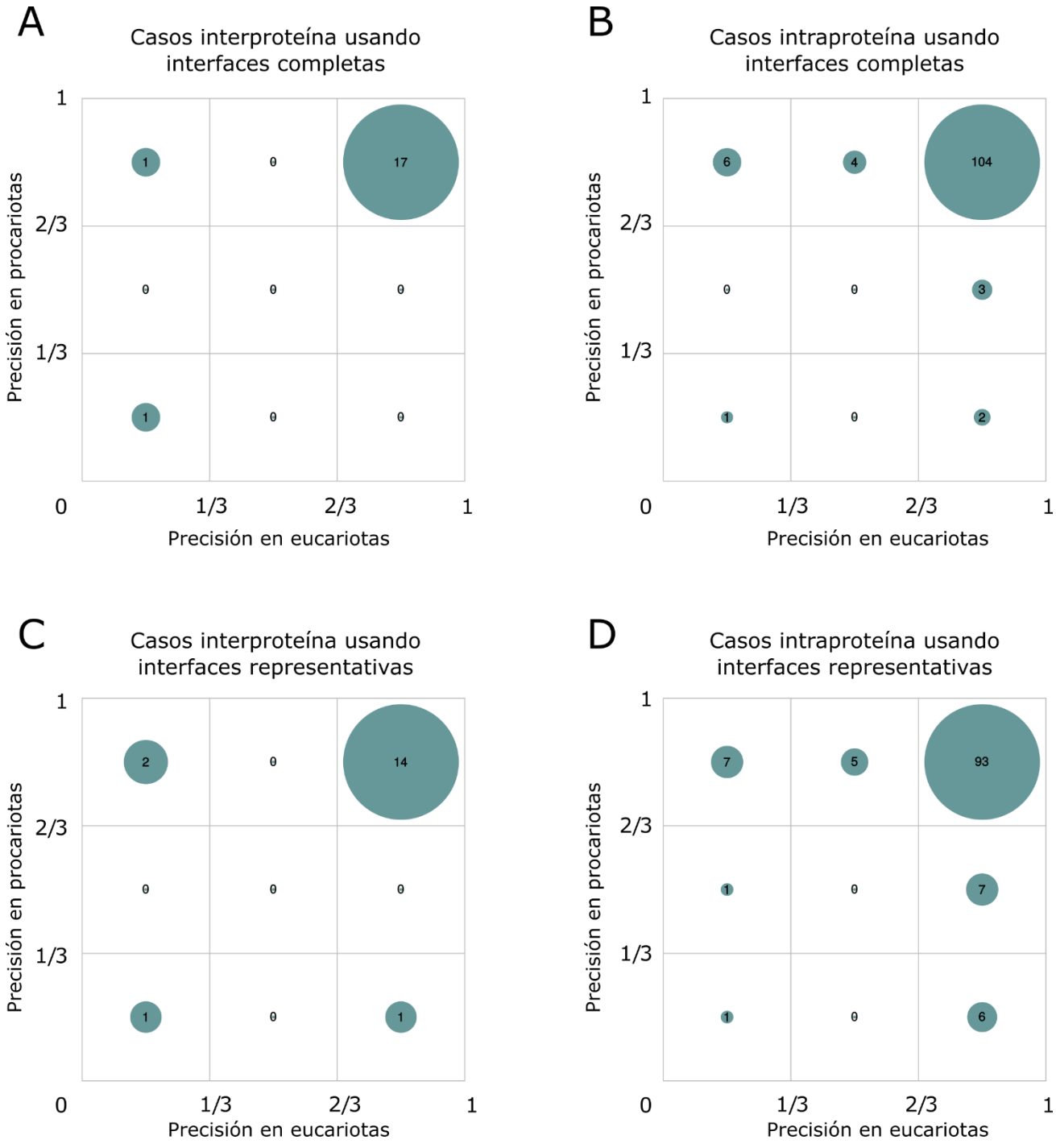


Figura A6. Comparación de la precisión de contactos entre interfaces procariotas y eucariotas. A) Frecuencia relativa de la precisión en casos interproteína en tres intervalos de precisión usando interfaces completas. B) Frecuencia relativa de la precisión en casos intraproteína en tres intervalos de precisión usando interfaces completas. C) Frecuencia relativa de la precisión en casos interproteína en tres intervalos de precisión usando interfaces representativas. D) Frecuencia relativa de la precisión en casos intraproteína en tres intervalos de precisión usando interfaces representativas. Figura adaptada de [117].

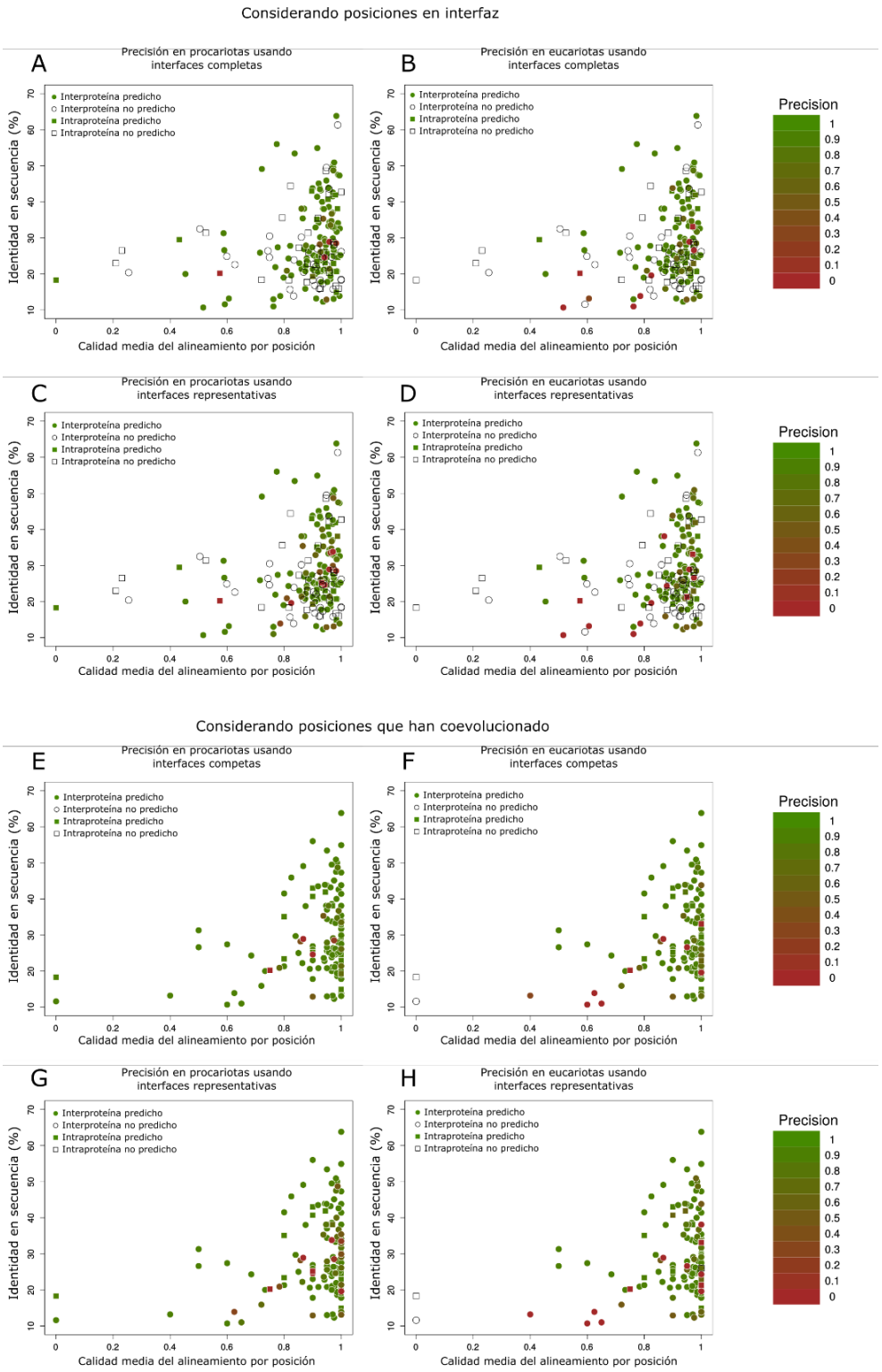


Figura A7. Influencia de la calidad del alineamiento en la transferencia de predicciones desde procariotas a eucariotas. Cada punto representa una interacción dominio a dominio (interproteína como cuadradros, intraproteína como círculos) con estructura en procariotas y eucariotas, coloreado por según la precisión en la predicción de contactos (en blanco cuando no hay predicciones). La calidad media del alineamiento por posición se obtiene mediante la media de la exactitud del alineamiento en cada posición estimados por HMMER y se corresponden con la probabilidad a posteriori del aminoácido presente en cada posición. (A-D) Para la calidad del alineamiento se consideran las posiciones en la interfaz en procariotas y las posiciones homólogas de éstas en eucariotas utilizando para medir la precisión interfaces completas (A y B) o interfaces representativas (C y D). (E-H) Para la calidad del alineamiento se consideran las posiciones que han coevolucionado en procariotas y las posiciones homólogas de éstas en eucariotas utilizando para medir la precisión interfaces completas (E y F) o interfaces representativas (G y H). Nótese que utilizando las posiciones que han coevolucionado no es necesario tener ninguna información estructural por lo que es siempre aplicable. Figura adaptada de [117].

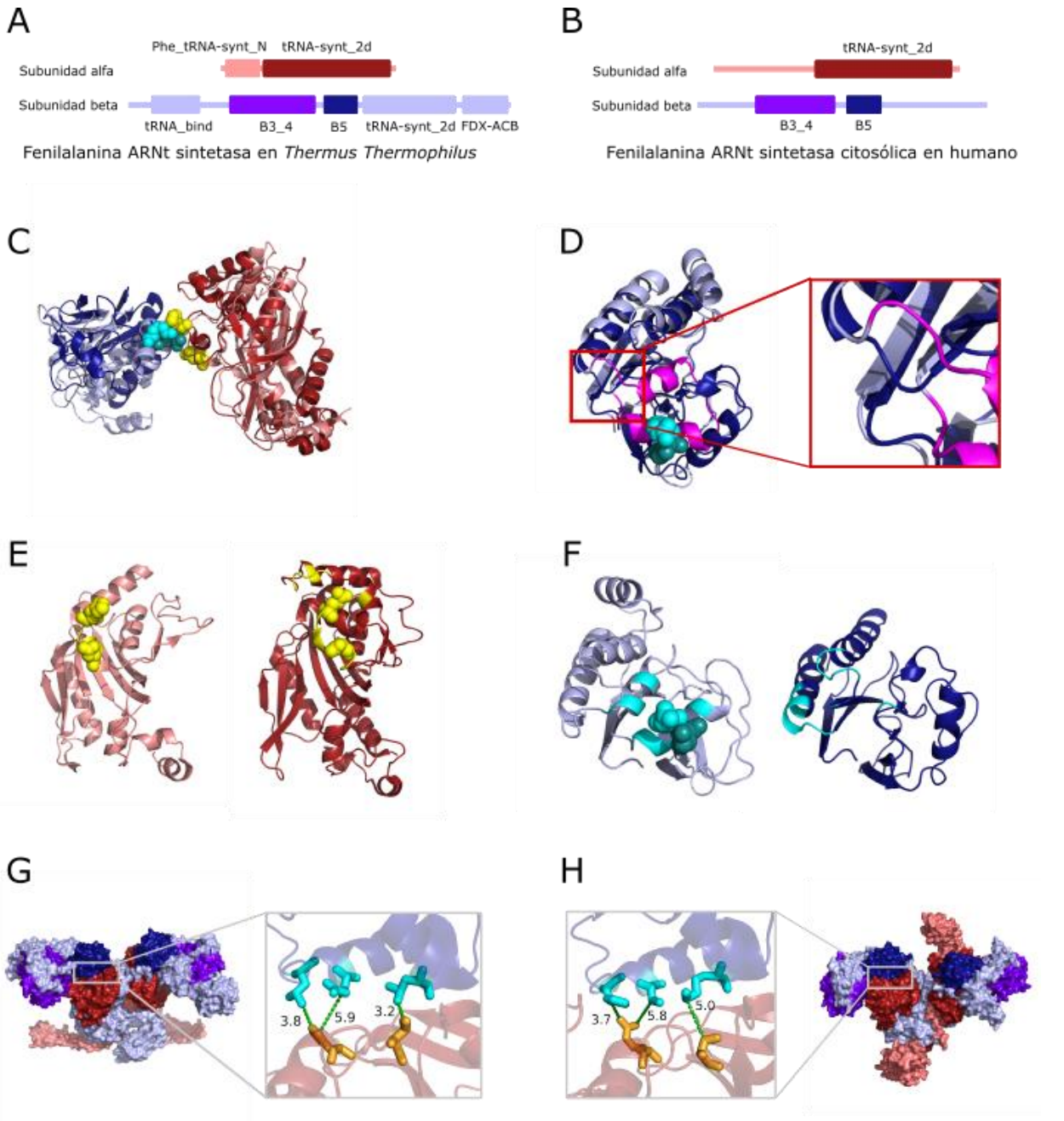


Figura A8. (A y B) Dominios Pfam encontrados en in *T. thermophilus* (A) y en humano (B) de la fenilalanina ARNt sintetasa. Los dominios Pfam que se pueden encontrar tanto en los complejos procariotas como en los eucariotas se muestran en colores oscuros, con el dominio con el centro catalítico (código Pfam: tRNA-synt_2d) en la subunidad α en rojo y los dominios B5 (código Pfam: B5) y B3/4 (código Pfam: B3_4) en violeta y azul oscuro respectivamente. C) Superposición de las estructuras de la PheRS (código PDB: 1EIY) en *T. thermophilus* y humano donde los dominios con el centro catalítico de ambos complejos, en rojo claro y oscuro, ha sido estructuralmente alineado. Los dominios B3/4 se muestran en azul claro y oscuro. Los aminoácidos de los pares de aminoácidos que han coevolucionado se muestran como esferas en cian en los dominios B3/4 y en amarillo en los dominios con el centro catalítico. D) Alineamiento estructural de los dominios B3/4 en azul claro y oscuro, las posiciones que han coevolucionado en esferas en cian y la inserción en procariotas en magenta. La región ampliada muestra donde ocurre la inserción en detalle. E) El dominio con el centro catalítico en *T. thermophilus* (deracha) y en humano (izquierda) con los residuos en la interfaz en amarillo y los que han coevolucionado en esferas amarillas. Pese a los cambios en el dominio B3/4, las predicciones de contacto en dominio con el centro catalítico se encuentran en la interfaz de éste. (G y H) Los 3 pares de posiciones que han coevolucionado en la interfaz entre el dominio con el centro catalítico y el dominio B5 mapeado en la estructura en *T. thermophilus* (G) y en humano (H) son mostradas con bastones y conectados con líneas grises punteadas. Se indica la distancia en angstroms entre los átomos pesados más cercanos entre los aminoácidos de las posiciones predichas. Figura adaptada de [117].

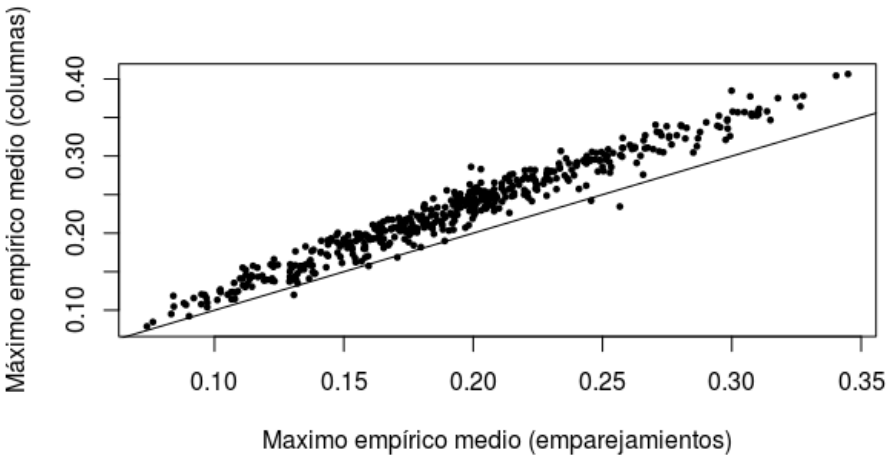


Figura A9. Máximo empírico medio para todo el conjunto de casos (490 IDDIs) utilizando las aleatorizaciones de emparejamientos (eje de abscisas) y las aleatorizaciones por columnas (eje de ordenadas).

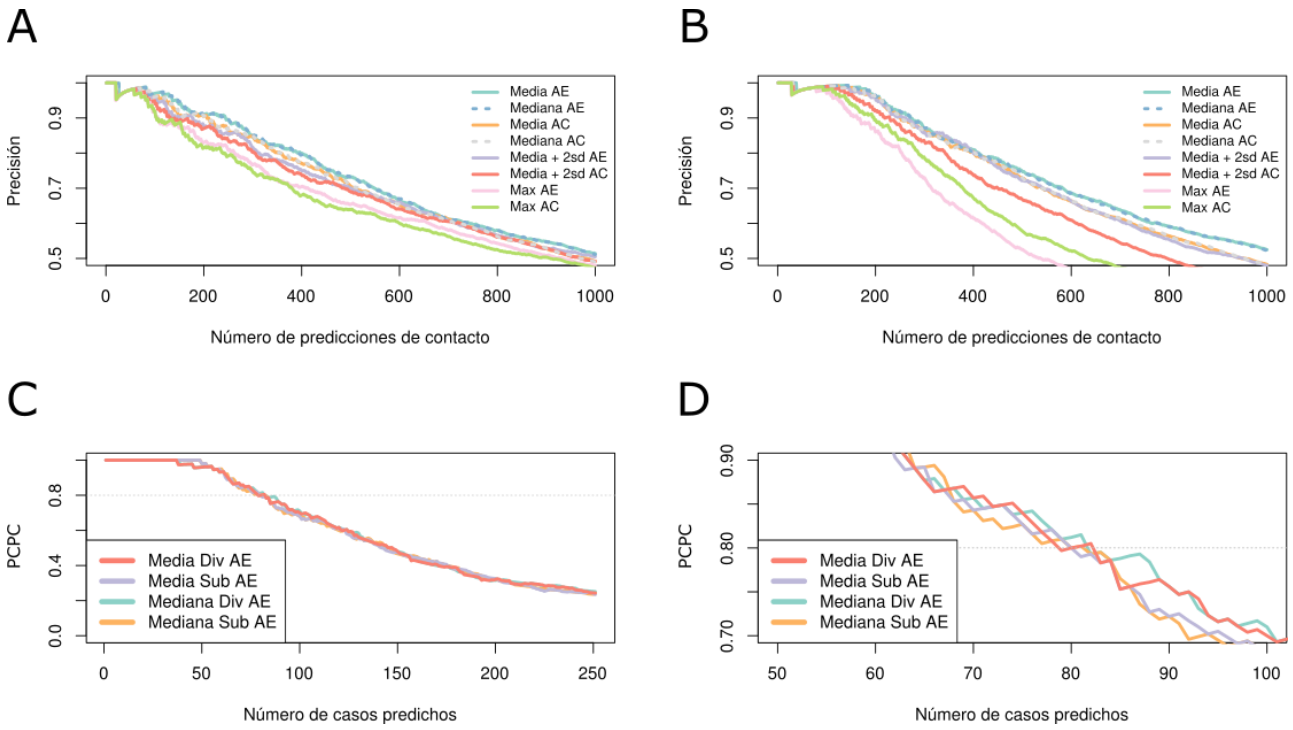


Figura A10. Precisión para las 16 estrategias de corrección de *score* APC mediante el uso del máximo *score* APC medio empírico. Precisión de las 8 estrategias con la operación de división (A) o sustracción (B) para las 1000 primeras predicciones de contacto. Para cada estrategia las predicciones están ordenadas por el *score* en orden decreciente. Se consideran las predicciones de todos los casos conjuntamente. Proporción de casos predichos correctamente (PCPC, con precisión ≥ 0.8) para 250 (C) casos o en el rango entre 50 y 100 casos (D).

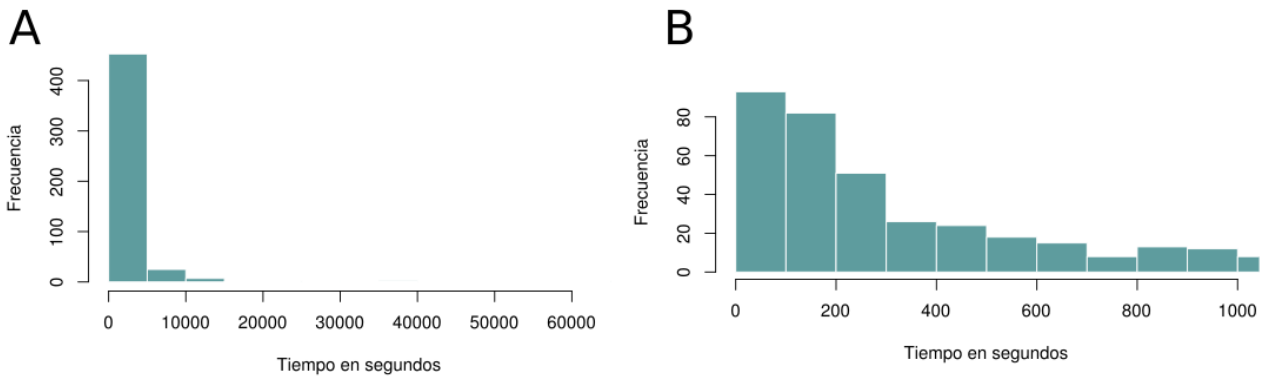


Figura A11. Tiempo de computación. Histograma del tiempo de computación en segundos para todo el rango de tiempos encontrado en el conjunto de datos (el máximo tiempo de computo es de 60 801 segundos) o para tiempos entre 0 y 1000 segundos.

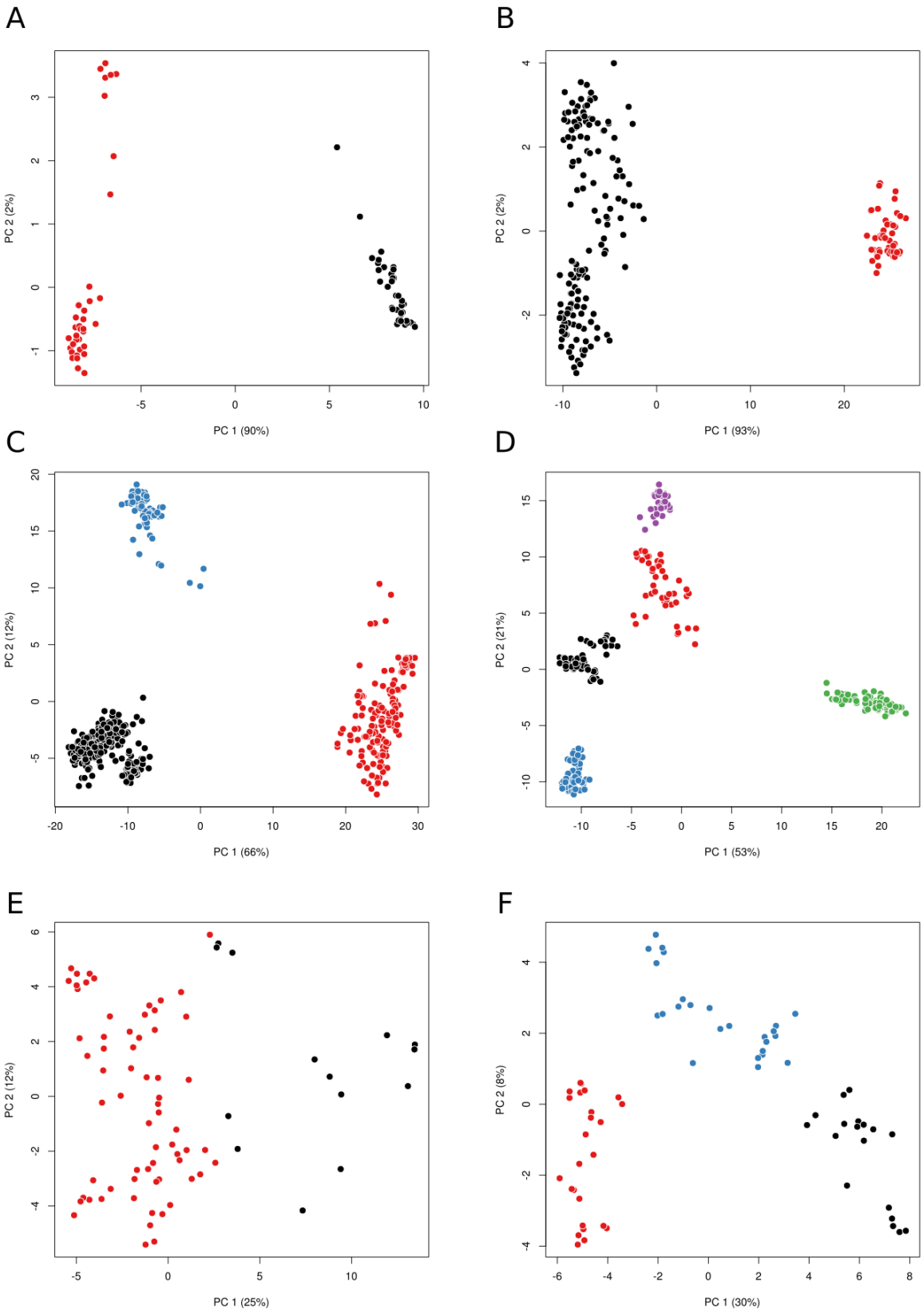


Figura A12. Se muestran las dos componentes principales de un análisis de componentes principales sobre las distancias BLOSUM62 entre las secuencias del alineamiento emparejado. Los panes A y B muestran casos con señales filogenética muy fuertes. Los paneles C y D casos con una señal filogenéticas bastante fuertes, mientras que los casos E y F muestran casos con una señal filogenética débil.

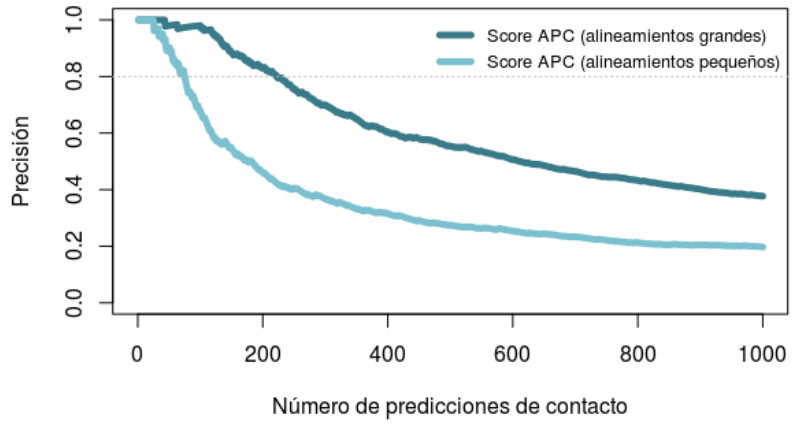


Figura A13. Precisión en función del número de predicciones para el *score* APC separando los casos con 500 o más secuencias en los alineamientos emparejados (alineamientos grandes) de los que tienen menos de 500 secuencias (lineamientos pequeños).

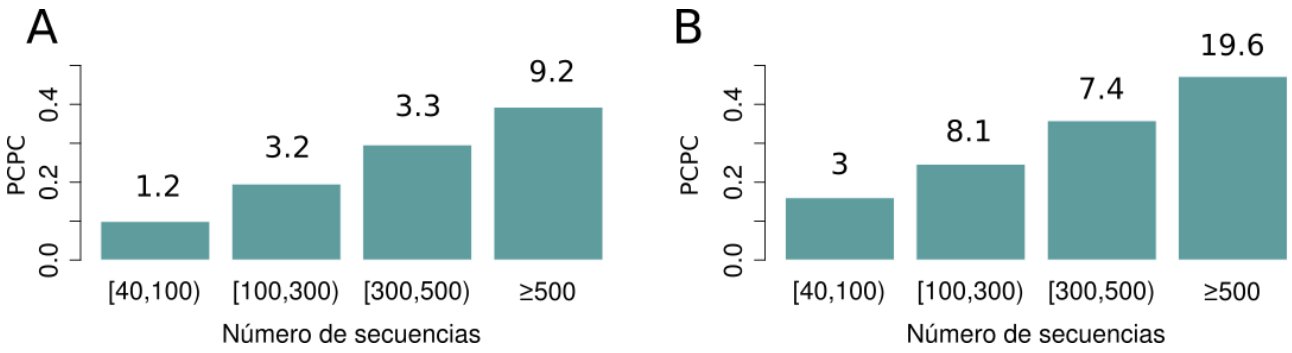


Figura A14. Proporción de casos predichos correctamente (PCPC) usando una precisión de 0.8 (A) o una precisión de 0.5 (B) en función del número de secuencias en el alineamiento emparejado. El número encima de cada barra el indica el número de predicciones en promedio.

	Interfaz completa		Interfaz representativa	
Conjunto	**Antes de filtrar**	**Después de filtrar**	**Antes de filtrar**	**Después de filtrar**
Precisión global				
inter	0.81 ± 0.05	0.91 ± 0.04	0.74 ± 0.06	0.83 ± 0.05
intra	0.95 ± 0.01	0.96 ± 0.01	0.90 ± 0.01	0.91 ± 0.01
Precisión promediada por interfaz				
inter	0.77 ± 0.08	0.88 ± 0.06	0.69 ± 0.09	0.78 ± 0.08
intra	0.90 ± 0.02	0.93 ± 0.02	0.87 ± 0.02	0.89 ± 0.02

Tabla A1. Precisión de las predicciones de contactos en eucariotas. A) Precisión de contactos global para todos los pares de posiciones con un *z-score* coevolutivo mayor que 8. B) Precisión promediada por interfaz, esto es, se computa la precisión para cada interfaz por separado considerando todos los pares de posiciones y obtuvo el promedio de estas precisiones. Se muestra la precisión obtenida antes y después de aplicar el filtro de casos con alineamiento de baja calidad en eucariotas. Los errores estándar se obtuvieron mediante *bootstrapping* (con 10000 iteraciones con reemplazamiento). Tabla adaptada de [117].

	Correlaciones parciales								
	Todos los casos			< 1000 secuencias			≥ 1000 secuencias		
	Num. sec.	Num. pos.	Señal Filogenia	Num. sec.	Num. pos.	Señal Filogenia	Num. sec.	Num. pos.	Señal Filogenia
Umbral objetivo (precisión 0.8)	0.38	-0.66	0.34	0.43	-0.68	0.34	-0.69	-0.57	0.03
Umbral objetivo (precisión 0.5)	0.27	-0.65	0.31	0.33	-0.66	0.32	-0.61	-0.49	-0.14
Máximo empírco medio (AE)	0.28	-0.72	0.38	0.37	-0.74	0.38	-0.76	-0.61	-0.19
Máximo empírco medio (AC)	0.39	-0.76	0.35	0.49	-0.78	0.35	-0.76	-0.62	-0.15

Tabla A2. Correlaciones parciales de Spearman para los umbrales óptimos (con precisiones 0.8 y 0.5) y los máximos empíricos medios de las aleatorizaciones de emparejamientos (AE) y por columnas (AC) con respecto a las variables número de secuencias (Num. sec.), número de posiciones (Num. pos.) y señal filogenética (Señal filogenia). Se muestran para 3 conjuntos de casos, todos los casos (n=490), casos con menos de 1000 secuencias no redundantes en el alineamiento emparejado (<1000 secuencias; n=457) y con 1000 o más secuencias no redundantes (≥ 1000 secuencias; n=33).

ANEXO II. PUBLICACIONES

Durante el transcurso de esta tesis he participado en los siguientes trabajos.

- Trabajos relacionados con la tesis

    - Artículo publicado en *Proceedings of the National Academy of Sciences* (se adjunta a continuación)

Rodriguez-Rivas, J., Marsili, S., Juan, D. and Valencia, A., 2016. Conservation of coevolving protein interfaces bridges prokaryote–eukaryote homologies in the twilight zone. *Proceedings of the National Academy of Sciences*, 113(52), pp.15018-15023.

    - Manuscrito enviado a revista para publicación

Rodriguez-Rivas, J., Marsili, S., Juan, D. and Valencia, A., 2019. Increasing the reliability and applicability of coevolution-based inter-protein contact predictions through background correction. Manuscrito enviado para publicación

- Trabajos publicados no relacionados con la tesis

    - Artículo en *PLoS computational biology*

Abascal, F., Ezkurdia, I., Rodriguez-Rivas, J., Rodriguez, J.M., del Pozo, A., Vázquez, J., Valencia, A. and Tress, M.L., 2015. Alternatively spliced homologous exons have ancient origins and are highly expressed at the protein level. *PLoS computational biology*, 11(6), p.e1004325.

    - Artículo en *Nucleic acids research*

Rodriguez, J.M., Rodriguez-Rivas, J., Di Domenico, T., Vázquez, J., Valencia, A. and Tress, M.L., 2017. APPRIS 2017: principal isoforms for multiple gene sets. *Nucleic acids research*, 46(D1), pp.D213-D217.