Este capítulo pertenece al libro

Bioinformática con Ñ

Edición y coordinación:
Álvaro Sebastián y Alberto Pascual-García

Autores:
F. Abascal, J. Aguirre, E. Andrés-León, D. Bajic, D. Baú, J. A. Bueren-Calabuig,
Á. Cortés-Cabrera, I. Dotu, J. M. Fernández, H. G. D. Santos, B. García-Jiménez,
R. Guantes, I. Irisarri, N. Jiménez-Lozano, J. Klett, R. Méndez, A. Morreale,
A. Pascual-García, A. Perona, A. Sebastian, M. Stich, S. Tarazona, I. Yruela y R. Zardoya

Portada:
Enrique Sahagún (`http://www.scixel.es`)

Maquetación:
Álvaro Sebastián
Usando LaTeX(`http://www.latex-project.org`) y Texmaker (`http://www.xm1math.net/texmaker`)

Editorial:
Libro autoeditado e impreso por CreateSpace (`http://www.createspace.com`)

Depósito legal: SE-NNNNNNN

ISBN: NNNNNNN

Índice general

Capítulo 1

Evolución de estructura de proteínas

Alberto Pascual-García

1.1. Introducción

En la presente sección vamos a estudiar el impacto de los distintos eventos evolutivos que ocurren desde el punto de vista genético, en la evolución de las estructuras de proteínas. *Los distintos eventos evolutivos tendrán un impacto en la estructura de la proteína* que, al verse sometida al continuo proceso de selección, podrá será incorporado con mayor o menor probabilidad según cómo afecte a las posibilidades de reproducción del individuo. Cuantificar este impacto es una tarea difícil entre otras cosas porque, como vimos en el **??** y volveremos a ver en el presente capítulo, las estructuras de proteínas reales presentan cierta *robustez mutacional* que les permite acumular mutaciones (e incluso eventos más dramáticos como inserciones o deleciones) sin que su función se vea afectada. Así que para conseguir nuestro objetivo intentaremos inferir el impacto de los distintos eventos a través del análisis del espacio de todas las estructuras de proteínas conocidas, identificando en particular cuáles se nos presentan como dominantes en la evolución de las mismas.

1.2. Origen de nuevas proteínas

Para comenzar, vamos a recordar las *cuatro maneras principales de replicación de genes* que codifican las proteínas, si bien recomendamos al lector que tenga en cuenta en primer lugar los contenidos del capítulo de filogenia y evolución molecular (**??**).

- *Transferencia vertical*. Son aquellos genes que se transfieren desde progenitores a su descendencia. Los genes relacionados por este proceso se denominan *ortólogos*.
- *Transferencia horizontal*. Genes que se transfieren desde otro organismo de la misma o distinta especie. Como ejemplos podríamos tener la conjugación bacteriana, el intercambio entre mitocondria y núcleo o entre bacterias y virus a hospedadores eucarióticos y viceversa.
- *Duplicación génica*. Copia de un gen dentro de un mismo genoma acompañada de una diferenciación funcional, lo que se denomina neofuncionalización. Genes relacionados de este modo se denominan *parálogos*. Si la diferenciación funcional no tiene lugar, no existe presión selectiva para conservar las dos copias y uno de los dos genes se convierte en un *pseudo-gen* y se pierde.

- *Fusión génica*. A través de la fusión de dos o más fragmentos de estructura super-secundaria.

Dos genes relacionados a través de alguno de estos mecanismos evolutivos, se verán sometidos a una nueva serie de eventos que harán que la similaridad entre dichos genes disminuya. Como hemos adelantado, si los genes son ortólogos, normalmente reflejan el resultado de una *deriva aleatoria* (del inglés *random drift*) en la que ambos genes están sometidos a una presión selectiva para conservar la función de la línea ancestral. Pero si los genes son parálogos reflejaría más bien la adaptación de las distintas copias para una *diversificación de funciones*. Para una revisión integrada de los mecanismos que influyen en la evolución de proteínas recomendamos al lector la referencia [25].

Nos queremos preguntar qué eventos son dominantes en la evolución de las estructuras de proteínas. Es decir, qué eventos determinan esencialmente la similaridad (o disimilaridad) entre las proteínas a nivel, ya no de su secuencia, sino de su estructura. Como hemos comentado en la sección de plegamiento de proteínas, la estructura juega un papel determinante en el correcto funcionamiento de la proteína, y los efectos de *los diferentes eventos evolutivos posibles sobre las estructuras pueden ser muy distintos*. Por ejemplo, una *mutación puntual* podría prácticamente no afectar a la estructura si la mutación es sinónima o codifica para un aminoácido con un rol estructural similar. Como casi siempre en biología hay excepciones. Por ejemplo se han descrito polimorfismos silenciosos que dan lugar a distintas estructuras y funciones [18], lo cual va en contra del (mal llamado dogma) central de la biología (**??**). Lo que está claro es que esperamos cambios más dramáticos si existen *inserciones y deleciones*. Por ejemplo, podría cambiar por completo la estructura si se inserta un único nucleótido, ya que al traducirse el código genético en tripletes nos cambiaría el marco de lectura y por tanto su traducción (ver **??**). Además se han presentado ejemplos como en [1] en los que un número mínimo de mutaciones pueden cambiar tanto la estructura como la función como se ilustra en la Figura 1.1.

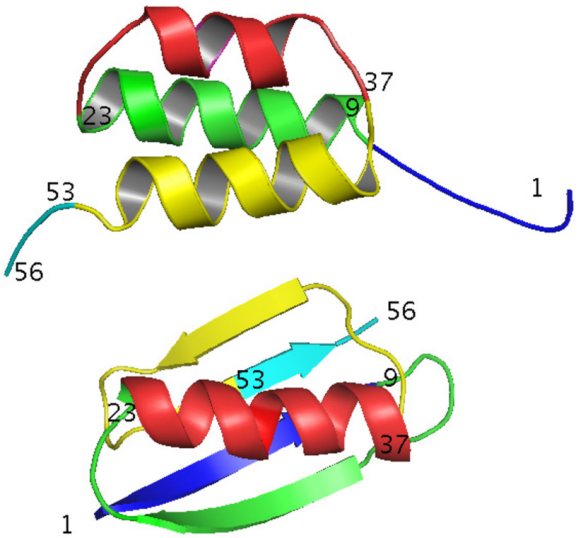

Figura 1.1: La proteína de Streptcoccus G presenta dos tipos de proteínas (G_A y G_B) que se unen a proteínas del suero sanguíneo, con estructuras distintas. En el estudio generan mutantes de cada uno de los grupos y consiguen que mantengan la estructura y función, a pesar de que las separan tan solo tres mutaciones [1]. Los números indican las posiciones de algunos aminoácidos, los mutados están en las posiciones 20, 30 y 45. Figura reproducida con permiso de los autores.

Continuando con el ejemplo, sabemos que la relación entre el número de sustituciones con respecto al de inserciones y deleciones varía entre alrededor de 5 veces más sustituciones hasta incluso 60 veces más [5], en función del organismo y de si es o no codificante la región estudiada. Pero, *aunque la frecuencia*

de las inserciones y deleciones es menor, su efecto sobre la estructura puede ser mucho más dramático. En la Figura 1.2 podemos ver el efecto de uno de los eventos evolutivos más dramáticos que se puede encontrar en la literatura, entre la luciferasa bacteriana y su homóloga, la flavoproteína no fluorescente que codifica el gen luxF de Photobacterium, con una deleción de aproximadamente 90 residuos descrita en [13].

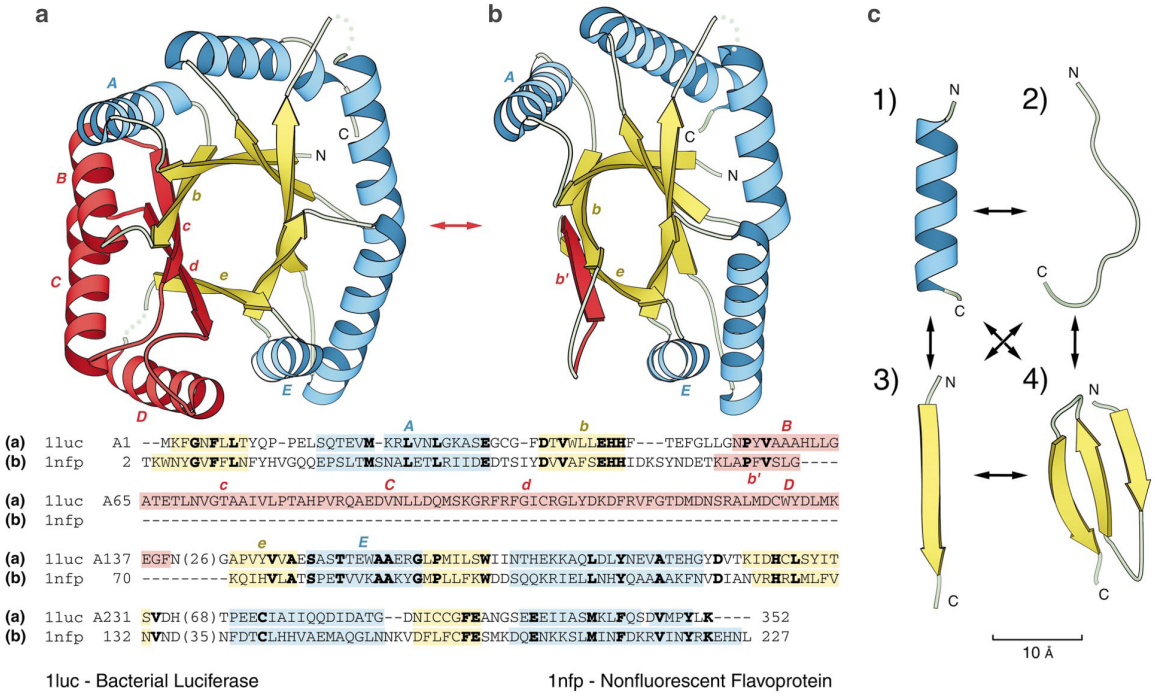

Figura 1.2: Comparación estructural entre la luciferasa bacteriana (izquierda) y la flavoproteína no fluorescente (derecha). Son dos proteínas homólogas a pesar de que su divergencia en secuencia es importante, como se ve en el alineamiento, con una (presumiblemente) deleción de 90 residuos. A la derecha se muestran cambios típicos en estructura secundaria observados [13]. Figura reproducida con permiso del editor.

Cambios tan dramáticos ponen en jaque nuestra habilidad para modelar estructura de proteínas. Por este motivo cabe preguntarse qué escenario esperaríamos encontrar en cada caso en función del tipo y frecuencia de los distintos eventos evolutivos. Es decir, *si consideráramos todas las proteínas conocidas, ¿cómo de parecidas son globalmente?, ¿tienen similaridades locales?* Si encontramos similaridades, ¿cómo lo debemos interpretar?, ¿es este parecido debido principalmente a que están relacionadas evolutivamente?, o es en cambio una consecuencia de que realizan una función similar aunque no tengan un origen evolutivo común (como veremos, casos de convergencia funcional). El entender la *relación secuencia-estructura-función* bajo la luz de la evolución, nos permitirá proponer métodos para predecir la función de una nueva proteína, utilizando su información en secuencia y estructura.

Por último, es importante destacar que vamos a trabajar con dominios estructurales, como se muestra en la Figura 1.3. Un *dominio estructural* lo podríamos definir como la unidad mínima en que se puede descomponer una cadena de aminoácidos de modo que sea autónoma desde el punto de vista evolutivo. Es decir, esperamos que sea una región conservada con una estabilidad termodinámica y accesibilidad cinética suficiente para realizar una determinada función. De modo que dada una proteína larga el primer paso es el intentar descomponerla en dominios, pues cada uno de ellos sería en principio susceptible de evolucionar autónomamente. Un reto importante consiste en entender cómo estos dominios han coevolucionado, es decir, de qué modo el hecho de que se encuentran en la misma proteína hn

influido en la evolución de cada uno como estructuras autónomas. La descomposición en dominios es un problema bioinformático complejo en tanto en cuanto no es sencillo determinar criterios objetivos para realizar una descomposición precisa. Remitimos al lector interesado a consultar [15].

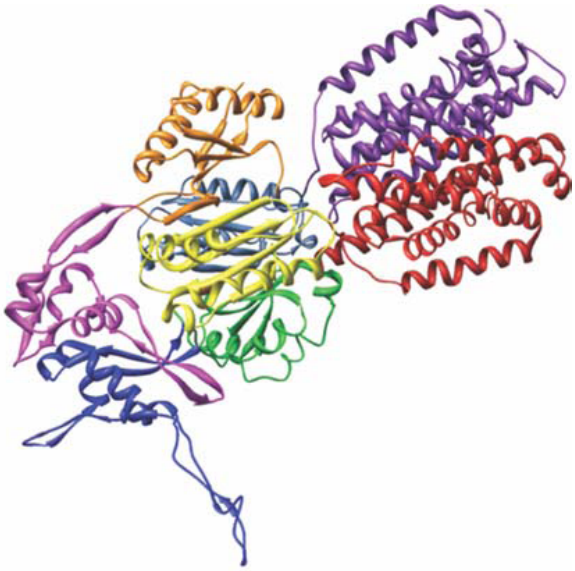

Figura 1.3: Ejemplo de descomposición en dominios de la cadena A de la proteína 1oy8 según la clasificación estructural de CATH [33]. Figura reproducida con permiso del editor.

1.3. Divergencia estructural gradual: especiación y duplicación génica

Comencemos considerando dos estructuras expresadas por genes homólogos, con una similaridad relevante. Su similaridad irá disminuyendo, y diremos por tanto que ambas estructuras divergen a lo largo del tiempo, a medida que los distintos eventos evolutivos se van acumulando en cada una de ellas. Si las *proteínas* son *ortólogas* ha habido un evento de especiación y dicha divergencia es debida esencialmente a la deriva génica. En ese sentido, *la acumulación de mutaciones es proporcional al tiempo que ha transcurrido desde la especiación*, y esa proporcionalidad nos permite decir que el modo en que las proteínas disminuyen su similaridad es gradual, es decir, sin brusquedades.

En cambio si las *proteínas* son *parálogas*, ha existido una duplicación génica y, tras este evento, los genes que no se transforman en pseudo-genes experimentan una aceleración en el ritmo de sustitución debido a una reducción en la *presión selectiva sobre una de las copias* y, quizá más importante, cambios en la regulación que modifican fuertemente la expresión génica. Estos genes duplicados divergerán por tanto en secuencia, estructura y eventualmente en función biológica. Sin embargo, en muchos casos las actividades bioquímicas se conservan (no necesariamente la función biológica) y, a pesar de que el ritmo de sustitución ha aumentado, podemos considerar en muchos casos que la divergencia es también aproximadamente gradual.

En general por tanto, el que la divergencia sea consistente con el reloj molecular, implica que pasado un tiempo t encontraríamos grupos de genes relacionados entre sí más fuertemente que con otros grupos de genes. Llamaremos a cada uno de estos grupos fuertemente relacionados *familias de genes*. Estas familias se obtienen "cortando" el árbol filogenético a un determinado umbral que, como hemos

apuntado, estará relacionado con el tiempo de divergencia. A este "corte" le llamaremos en adelante una *partición del espacio de genes*, ya que lo partimos *en distintas familias*. Y para cada una de las familias de proteínas, no existe discusión sobre su *origen monofilético*, es decir, sobre la *existencia de un ancestro común a todas las proteínas que pertenecen a la familia.*

Una propiedad interesante de la partición que contribuyó a determinar el origen monofilético de las familias de genes es la siguiente. Para cada una de las familias de genes, podemos contar el número de genes que contiene. Y de este ejercicio, podemos conocer la probabilidad de que una familia contenga un determinado número de genes, es decir, su *distribución de probabilidad*. Como vemos en la Figura 1.4 [17], se encontró que era una distribución muy apuntada hacia la derecha (tengamos en cuenta que la figura está en escala logarítmica), donde las familias no tienen un tamaño típico. Esta distribución sigue una ley de potencias y también es conocida como *distribución libre de escala* [35]. Estas distribuciones son importantes porque se han propuesto varios modelos que las generan que pueden ser compatibles con los escenarios evolutivos que consideramos, además de ser distribuciones encontradas frecuentemente en sistemas muy distintos, no solamente biológicos. En el capítulo de biología de sistemas recomendamos al lector dirigirse a la sección de redes complejas donde se explica la distribución y alguno de los modelos evolutivos asociados (**??**). Pero de momento nos basta entender que estos modelos son compatibles con el escenario en el que la duplicación génica junto con la divergencia gradual es dominante. Además, el hecho de que sea libre de escala, nos dice que *el escenario evolutivo dominante no ha cambiado a lo largo de la escala temporal.*

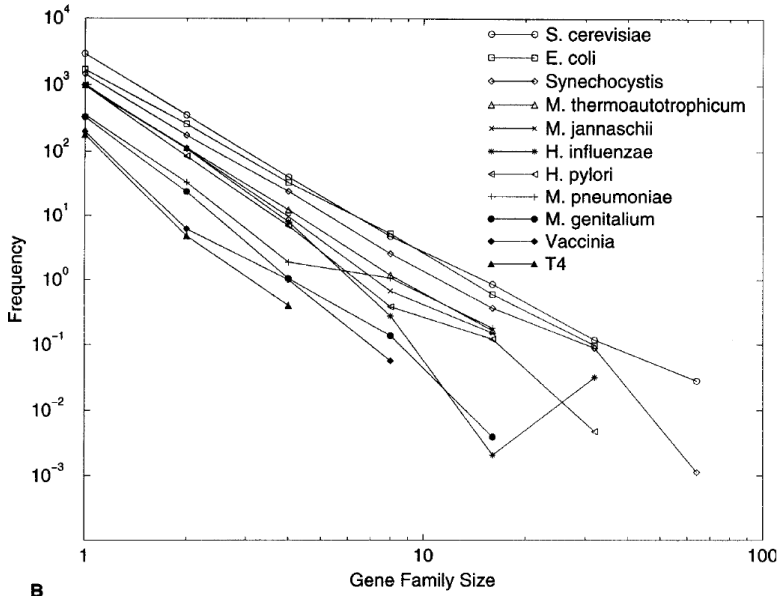

Figura 1.4: Histograma del número de familias con un tamaño determinado. El tamaño viene dado por el número de genes que contiene dicha familia. La gráfica muestra el histograma para distintos organismos. Aunque la pendiente de las curvas cambia no lo hace el tipo de distribución, que se ajusta muy bien siempre a una ley de potencias. Reproducido con permiso del editor.

Esta observación abrió la puerta a preguntarse si este tipo de partición se encontraría también si hiciéramos el mismo ejercicio en el espacio de estructura de proteínas. Gracias a un artículo seminal debido a *Chothia y Lesk* [6] hace ya más de 25 años, y que abrió la puerta a comprender cuantitativamente las relaciones entre secuencia y estructura, sugería que así debería ser. En este artículo *observaron que la raíz de la desviación cuadrática media entre globinas* ($RMSD$, ver eq:ProtEvolucion:RMSD) *divergía regularmente con el número de sustituciones* (ver Figura 1.5). Esta observación nos invitaría a pensar

que *es posible construir también "familias" de estructuras a partir de medidas de similaridad* entre ellas, es decir, a establecer una clasificación. Esta fue la propuesta natural a la luz de las observaciones y por tanto la primera que vamos a desarrollar. Posteriormente vamos a ver otros escenarios en donde se sugiere la posible dominancia de otros mecanismos evolutivos como consecuencia de otras observaciones, también interesantes.

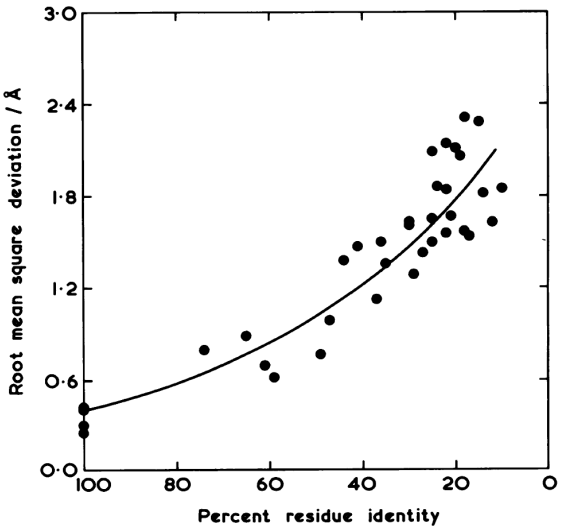

Figura 1.5: Relación entre la divergencia en secuencia, cuantificada mediante su porcentaje de identidad, y la divergencia en estructura cuantificada con la raíz de la desviación cuadrática media de sus posiciones espaciales. Ver las secciones de evolución en secuencia y alineamiento estructural para más detalles sobre estas medidas. Se observa que la divergencia en estructura para cada par de proteínas es proporcional a su divergencia en secuencia y por extensión al tiempo de divergencia con respecto a su ancestro común. Figura reproducida con permiso del editor.

1.3.1. Clasificación de estructura de proteínas

Supongamos que las estructuras divergen gradual y proporcionalmente con el tiempo y que sabemos construir una medida que, teniendo en cuenta dos estructuras cualesquiera, a y b, nos de una distancia indicativa de su disimilaridad estructural $d(a, b)$. Esto lo haríamos con un alineamiento de estructuras, que se puede consultar en la **??**. Si estas condiciones se cumplen, podemos deducir un resultado interesante. Imaginemos que consideramos dos proteínas homólogas a, b y una tercera proteína c que consideramos como referencia fuera del grupo. Si la acumulación de mutaciones, y por tanto la divergencia entre a y b, es constante a lo largo del tiempo, la siguiente relación se debe verificar aproximadamente $d(a, c) \approx d(b, c)$. Que se cumpla esta relación implica que estamos en situación de construir un árbol sin ambigüedades. El que podamos construir un árbol significa que podríamos además determinar un umbral por debajo del cual (si utilizamos como medida una distancia o disimilaridad, por encima si es una similaridad) las proteínas están globalmente relacionadas desde el punto de vista estructural formando conjuntos bien separados, es decir, podríamos construir una clasificación. A cada uno de estos *conjuntos de estructuras con una similaridad global* sustancial se les denomina *fold*(plegamiento).

No debemos confundir el plegamiento de una única proteína con este concepto, ya que el término en inglés es el mismo. Aquí de algún modo nos estamos refiriendo a los motivos estructurales que comparten todas las proteínas que están dentro de uno de estos conjuntos que llamamos *folds*. Es decir, es una especie de idea "platónica" a través de la cual se representa a todas las proteínas de dicho conjunto

tras reducir su variabilidad a la estructura mínima común. Y notamos también que, aunque esperamos que estos *folds* estén relacionados con las familias obtenidas mediante el análisis de las secuencias de genes, no tienen por qué ser necesariamente las mismas. De hecho, veremos más adelante que existen argumentos para esperar diferencias entre los conjuntos que se obtienen por análisis estructural respecto de los obtenidos por análisis de secuencias.

Las observaciones que hemos indicado motivaron la posibilidad de que se pudieran definir *folds*, y por extensión la búsqueda de *métodos para la creación de clasificaciones estructurales de proteínas*. Las clasificaciones más relevantes son las siguientes:

- *SCOP (Structural Classification Of Protein structures)* [22]. Ésta ha sido típicamente la clasificación de referencia. Es una clasificación en principio manual pero que, dado el elevado número de proteínas que hay que clasificar anualmente, es de esperar que utilicen herramientas automáticas de preprocesamiento (si bien es cierto que su actualización es la más lenta). Su éxito reside precisamente en el análisis experto de las estructuras, lo cual es al mismo tiempo una fuente de subjetividad. Se ha observado [27] que utilizan criterios más allá de los puramente estructurales para su clasificación, con un fuerte énfasis en las relaciones evolutivas. Por este motivo sí es de esperar que exista un acuerdo razonable entre los *folds* definidos en esta clasificación y los obtenidos por análisis de las secuencias pero la consistencia estructural de los grupos es discutible.

- *CATH (Class, Architecture, Topology, Homologous superfamily)* [24]. Esta clasificación es semiautomática. Realizan un preprocesamiento que incluye alineamientos estructurales automáticos junto con un algoritmo de aglomeración automática (en inglés algoritmo de *clustering*) de las nuevas estructuras, que son posteriormente analizadas por expertos. El punto fuerte de esta clasificación es su descomposición en dominios, que son bastante consistentes con otros métodos y con las evaluaciones manuales por expertos, y recientemente la incorporación de abundante información sobre la función de los dominios. Las siglas del nombre de la clasificación se refieren a los distintos niveles en los que clasifican las estructuras. Lo que llamamos fold en SCOP sería el equivalente al nivel de Topology en CATH.

- *FSSP (Fold Classification based on Structure-Structure Alignment of Proteins)* [16]. Esta clasificación es completamente automática y por tanto es la que encontraremos más actualizada. Se basa en la comparación estructural entre proteínas utilizando el algoritmo de alineamiento estructural DALI (que se explica en el **??**), junto con un umbral de significatividad. En el caso de DALI, al tratarse de un Z-score (**??**, **??**), el umbral utilizado es una similaridad mayor de dos. Aunque en este caso la clasificación se realiza de manera más objetiva, aún existe cierta arbitrariedad en la definición del umbral.

Existen dos críticas fundamentales a las dos primeras clasificaciones que está motivando que replanteen sus esquemas de clasificación [7]. El primero es sobre la definición de lo que se entiende por un *fold*. En la clasificación de SCOP, se dice que *dos proteínas comparten el mismo fold si tienen el mismo conjunto de estructuras secundarias principales con la misma conectividad*. Dilucidar cuáles son estructuras secundarias principales o simples embellecimientos es una tarea que debe ser definida de manera objetiva, lo cual no es sencillo si la clasificación no es automática.

Por otra parte, ambas clasificaciones establecen una jerarquía entre los distintos niveles que definen. Esto está motivado por el hecho de que, si podemos construir un árbol, podríamos en principio "cortarlo" en distintos umbrales de modo que los conjuntos estén unos dentro de otros. El problema surge principalmente porque en uno de los niveles más bajos que definen, el *nivel de Superfamilia*, se asume que las proteínas, además de estar relacionadas estructuralmente compartiendo el mismo *fold*, son homólogas.

Pero vamos a ver que, al no ser la clasificación basada exclusivamente en la estructura, existirán casos en los que no será posible clasificar sin ambigüedad (ver Tabla 1.1 para un resumen). Estos casos podrían surgir por ejemplo cuando encontramos convergencia funcional o si existe una aceleración en el ritmo de acumulación de mutaciones, veamos ambos casos. Entendemos por *convergencia funcional* el proceso evolutivo por el cual dos proteínas presentan la misma función (y en particular para el caso que nos ocupa presentan algún tipo de similaridad estructural, bien local o global) a pesar de no ser posible el reconocer un ancestro común para ambas. Por tanto, *proteínas con orígenes evolutivos distintos convergen estructuralmente debido a que los requerimientos funcionales han hecho que el proceso selectivo encuentre soluciones similares.* El problema con la clasificación surgirá al encontrar estructuras similares como para estar en el mismo *fold* (por ejemplo debido a que existe convergencia funcional) pero una de ellas es homóloga a proteínas con un *fold* muy distinto, pues la jerarquía nos exigirá automáticamente una separación en distintos *folds* para respetar el requerimiento de homología, que se considera que prevalece.

Otro ejemplo problemático es el de proteínas homólogas con *folds* distintos. Si tenemos dos proteínas ortólogas, en principio esperamos que conserven la misma función en ambos organismos y que exista una presión selectiva sobre la estructura que nos permita observar que ambas tienen una similaridad estructural global relevante. Pero si ambas proteínas son parálogas, sabemos que ha existido una duplicación génica y por tanto existe una relajación en la presión selectiva sobre una de ellas, que podría acumular mutaciones a un ritmo más elevado hasta el punto de que no podemos ya considerarla evolución gradual. Por tanto, como se ha observado que pocos cambios en la secuencia permiten obtener plegamientos significativamente distintos, asumir a priori que proteínas homólogas deben tener plegamientos similares puede suponer un problema a la hora de clasificar estructuralmente [1, 29] (ver Figura 1.6). A continuación vamos a ver algunas aproximaciones computacionales para investigar qué mecanismos evolutivos son relevantes y, por extensión, si la clasificación estructural está o no justificada.

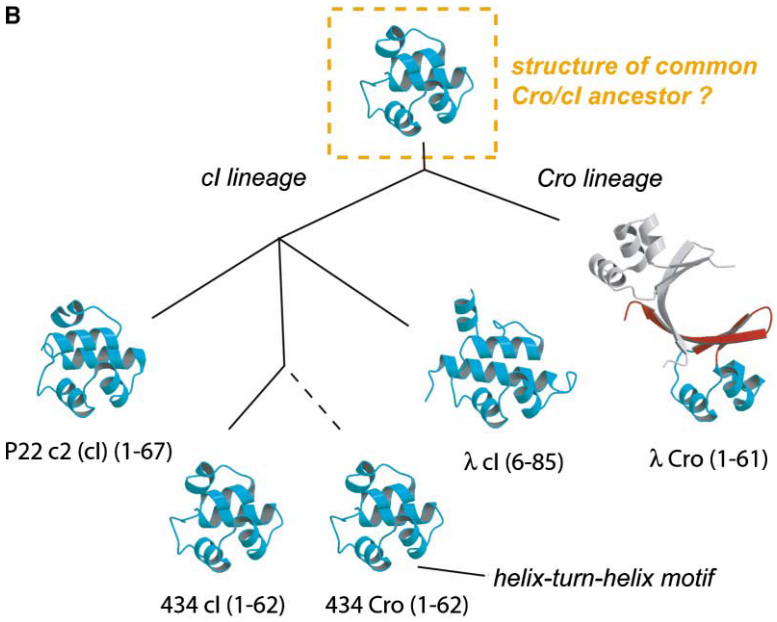

Figura 1.6: Divergencia estructural en la superfamilia Cro/cI, un grupo de factores de transcripción del fago lambda. A pesar de su elevada divergencia su homología ha sido demostrada [29], si bien existe incertidumbre en la reconstrucción. Observamos que la divergencia estructural es muy notable para λ Cro, que forma además un dímero. Este tipo de ejemplos suponen un reto para las clasificaciones estructurales jarárquicas como CATH y SCOP.

Categoría	Ejemplo	Definición/Propiedades	Problemas
Clase estructural	α/β	Composición global de los elementos de estructura secundaria	
Fold	barril TIM	Mismos elementos de estructura secundaria *principales* con la misma conectividad. Asunción a priori de relación evolutiva monofilética	Casos de convergencia funcional
Superfamilia	Aldolasas	Similaridad en secuencia reconocible	Aceleraciones evolutivas con cambios estructurales que dan lugar a estructuras que se considerarían folds distintos
Familia	Aldolasas clase I	Similaridad en secuencia significativa	Aceleraciones evolutivas con cambios estructurales que dan lugar a estructuras más parecidas a las de otras familias (o folds)
Grupos de ortólogos	2-keto-3deoxy-6-phosphogluconato aldolasa	Relaciones ortólogas dentro del conjunto de especies, conservación de la actividad bioquímica y, a menudo, la función biológica	Típicamente bien resuelto

Tabla 1.1: Ejemplo de clasificación jerárquica de proteínas. Para cada categoría se muestra un ejemplo y cómo se determina la pertenencia de una proteína a dicha categoría. La jerarquía implica que las propiedades de los niveles superiores se observan en los niveles inferiores, lo que puede inducir algunos problemas en los casos señalados en la última columna. Tabla adaptada de [19].

1.3.2. Cuantificando la divergencia estructural

La pregunta que surge a nivel bioinformático es cómo de fundamentada es la hipótesis de que la partición que obtendríamos en el espacio de estructura de proteínas es consistente con el que encontramos a nivel de secuencias de genes. En primer lugar queremos saber si la divergencia estructural es proporcional a la divergencia evolutiva para proteínas cuyas estructuras no son tan parecidas como en el caso del trabajo de Chothia y Lesk. En su trabajo, utilizaron el $RMSD$ pues sus proteínas estaban muy relacionadas evolutivamente y se podían superponer sin problemas, pero esta medida puede generar sesgos importantes si las proteínas son muy divergentes, para lo cual debemos acudir a otras medidas más complejas. Por ejemplo, podemos utilizar la llamada *divergencia de contactos*, que es tratada en detalle en la **??**. Esta medida utiliza el llamado *solapamiento de contactos*, del inglés *contact overlap*, que recordamos mide la *fracción de residuos alineados entre dos proteínas que se encuentran a una distancia menor de un determinado umbral*. Y lo que realiza es una transformación de el solapamiento de contactos análogamente a como se transforma la identidad en secuencia para obtener una medida de divergencia evolutiva.

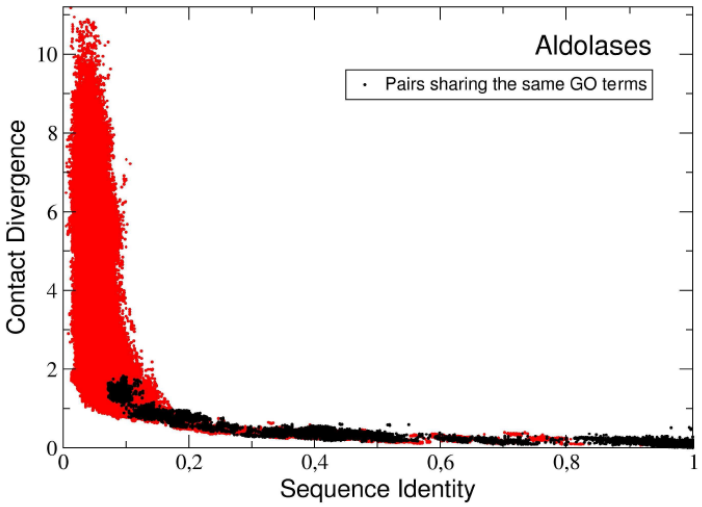

Figura 1.7: Divergencia estructural frente a divergencia en secuencia para la numerosa superfamilia de las aldolasas. Frente a la medida utilizada en la Figura 1.5, esta medida permite cuantificar divergencias más importantes y podemos observar el efecto de la función en dicha divergencia. Aquellos pares que comparten la misma función están marcados en negro, y vemos que están muy constreñidos estructuralmente. Sin embargo existen pares con distinta función y elevada similaridad estructural.

En la Figura 1.7 podemos observar la relación entre la divergencia de contactos y la similaridad en secuencia para un conjunto de proteínas homólogas de una superfamilia de CATH muy grande, en donde se pueden encontrar divergencias estructurales importantes. Vemos que *la divergencia de contactos está correlacionada linealmente con la divergencia evolutiva para valores bajos de esta variable*, lo que sugiere que crece también linealmente con el tiempo. Sin embargo, para similaridades en secuencia bajos, observamos una explosión en la divergencia de contactos que ya no se correlaciona con la similaridad en secuencia. Esta explosión se relaciona en la gráfica con cambios en la función, si bien lo contrario no es necesariamente cierto, dado que hay proteínas con distinta función y similaridad estructural significativa. Sin embargo, aquellas *proteínas con la misma función están claramente constreñidas desde el punto de vista estructural*. La figura también sugiere que proteínas con distintas funciones han sufrido aceleraciones evolutivas en el ritmo de divergencia estructural, que podrían estar relacionadas con la

presencia de inserciones y deleciones.

Así pues, vemos que *la clasificación estructural podría estar justificada hasta cierto umbral de distancia estructural*, pero deberíamos de tener problemas en clasificar si incorporamos además algún criterio tal como la información en secuencia (por extensión la información evolutiva) o la función. Podríamos tener problemas también si eventos evolutivos más dramáticos, como las inserciones y deleciones, tuvieran un efecto sobre las estructuras que impidieran la clasificación. Antes de aproximarnos a los métodos computacionales orientados a examinar si es posible una partición compatible con la divergencia estructural gradual, vamos a discutir ejemplos de estos eventos más dramáticos.

1.4. Evolución estructural mediante ensamblaje de módulos

Como hemos mencionado, las proteínas están formadas por dominios que se pueden considerar autónomos desde el punto de vista estructural, funcional y evolutivo. En algunas ocasiones en los genes eucarióticos las proteínas corresponden a varios exones, donde cada dominio estructural correspondería a un exón, lo que sugiere que *las proteínas multidominio están formadas a través de un proceso de ensamblaje de exones*. De hecho, proteínas con funciones nuevas se forman frecuentemente por el ensamblaje de dominios preexistentes [1, 29].

Del mismo modo se ha propuesto [21, 23] que los primeros dominios se han formado a través del ensamblaje de pequeños fragmentos polipeptídicos (del orden de 20 aminoácidos) que, sin ser globulares, tienen una estructura secundaria bien definida. La aparente periodicidad de los elementos de estructura secundaria apoyaría esta teoría [13, 20]. A estos elementos por debajo del nivel de dominio les llamamos *estructuras supersecundarias*. Se han propuesto varios mecanismos para definir estos fragmentos:

1.4.1. Longitud típica y posible origen evolutivo común: péptidos ancestrales

Si consideramos una cadena de monómeros, podemos calcular la *probabilidad $P_\delta(l)$ de que dos monómeros entren en contacto* (estén a menos de una distancia espacial δ que fijaremos) dado que hay l monómeros que los separan. Cuando estén muy cerca el uno del otro, la probabilidad será muy baja pues existe cierta rigidez intrínseca que impedirá a los monómeros entrar en contacto. Y si están muy lejos, los efectos entrópicos impedirán que entren en contacto con una probabilidad razonable. Así que existirá una *longitud óptima para la cual entrarán en contacto con máxima probabilidad* [4], que está relacionada con la denominada *longitud de persistencia*. Esta longitud óptima se puede estimar teóricamente mediante la teoría de polímeros [10], y dependerá de la naturaleza de los monómeros. Por ejemplo, para las prolinas esperamos tener que la probabilidad se maximiza para valores de l alrededor de 200 monómeros y para las glicinas de tan sólo 4. Pero en general, como en las proteínas tenemos cantidades bajas de ambas esperamos tener unos valores de l de entre 20 y 30 monómeros.

Si recorremos cadenas de proteínas contando el número de monómeros que separan cualquier par de contactos entre monómeros, encontraremos que la distribución de tamaños tiene un *máximo* precisamente alrededor de los 25 *monómeros*. Este tamaño está por encima de la longitud típica de un elemento de estructura secundaria pero por debajo de la longitud de un dominio estructural. Además, estos fragmentos cerrados no se encuentran ocasionalmente aquí y allá en las estructuras sino que, al recorrerlas, los encontramos uno tras otro [3], lo que nos lleva a pensar sobre un posible origen evolutivo.

Por ejemplo, si pensamos en un escenario en el que aún no existieran proteínas con funciones complejas, es razonable pensar en estos módulos como antecesores de las proteínas actuales, pues presentan una

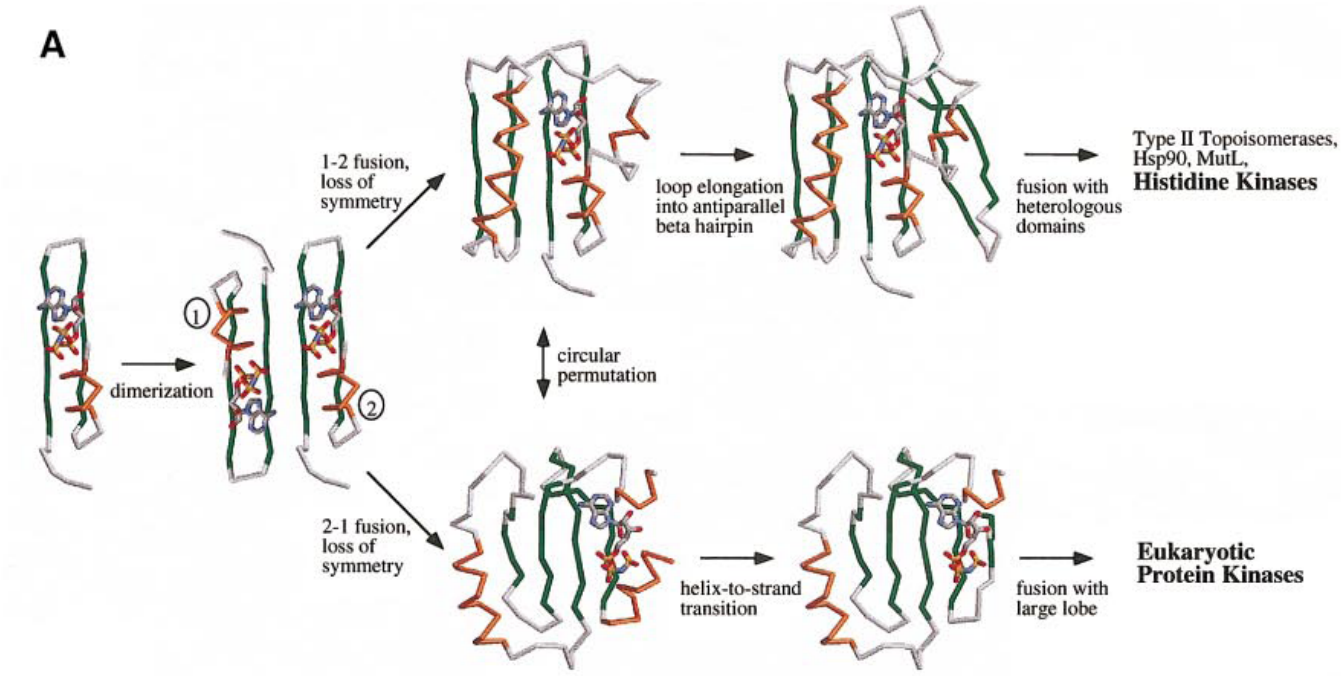

Figura 1.8: Trayectoria evolutiva especulativa descrita en [21], para explicar las similaridades locales encontradas entre los dominos de unión a ATP en histidina kinasas bacterianas y kinasas eucarióticas. El fragmento ancestral inicial, de tipo $\alpha\beta\beta$, se habría dimerizado y fusionado en orden inverso en cada trayectoria. A pesar de no tener una elevada similaridad en secuencia entre ambas, sus correspondientes elementos $\alpha\beta\beta$ se alinean estructuralmente con una resolución menor a 2Å. Figura reproducida con permiso del editor.

longitud y estabilidad termodinámica mínima para empezar a realizar procesos catalíticos rudimentarios de manera semiautónoma. Por este motivo *se ha propuesto que algunas de las repeticiones de fragmentos cerrados de esta longitud encontrados sistemáticamente en las proteínas que podemos hoy observar, podrían ser las reliquias de péptidos ancestrales* a partir de los cuales se formaron estructuras más complejas por duplicación y fusión de los (también pequeños) genes asociados. En la Figura 1.8 mostramos un ejemplo de la trayectoria evolutiva de dos proteínas bajo este escenario especulativo [21]. Pero, ¿cómo compaginar este escenario con los eventos evolutivos dominantes que conocemos? ¿cómo se relaciona la función de estos fragmentos ancestrales con la función de las proteínas actuales? Pues bien, se postula también que *la formación de dominios estructurales más largos, de entre* 100 *y* 150 *aminoácidos, podría ser compatible con un momento posterior en el que se encuentra un nuevo escenario de estabilidad debido a la circularización del DNA ancestral*, ya que el tamaño óptimo de cierre del DNA (determinado de nuevo por su flexibilidad) es de alrededor de 400 pares, consistente con el tamaño típico de los dominios actuales.

1.4.2. Relación con la función y búsqueda por recurrencia

Pero, independientemente de su origen, existe un creciente interés por la búsqueda de estos fragmentos ante su posible relevancia funcional [9, 32, 34]. Si hablamos de relevancia funcional, tenemos que hablar también de la secuencia. Hemos visto que existen proteínas homólogas que, aun conservando la misma función, tienen una similaridad en secuencia indistinguible de lo esperado por azar [26]. Por tanto, dado que el número de residuos es aún más pequeño para un fragmento que para una proteína, la

probabilidad de encontrar similaridades significativas en una secuencia tan corta es muy baja.

Un primer paso para abordar el problema lo encontramos en la base de datos de *BLOCKS* [14]. En esta base de datos se extraen fragmentos conservados de grupos clasificados por INTERPRO [2]. Curiosamente, la mayoría de estos fragmentos $\sim 75\,\%$ tienen tamaños de entre 10 y 40 aminoácidos. Con estos bloques en secuencia, podríamos dividir una proteína en regiones en función de su correspondencia con los bloques. Se puede observar que existen correspondencias claras a lo largo de una proteína [11], con otras regiones sin asignación que podrían tener un rol únicamente de soporte a las regiones funcionales (ver Figura 1.9)

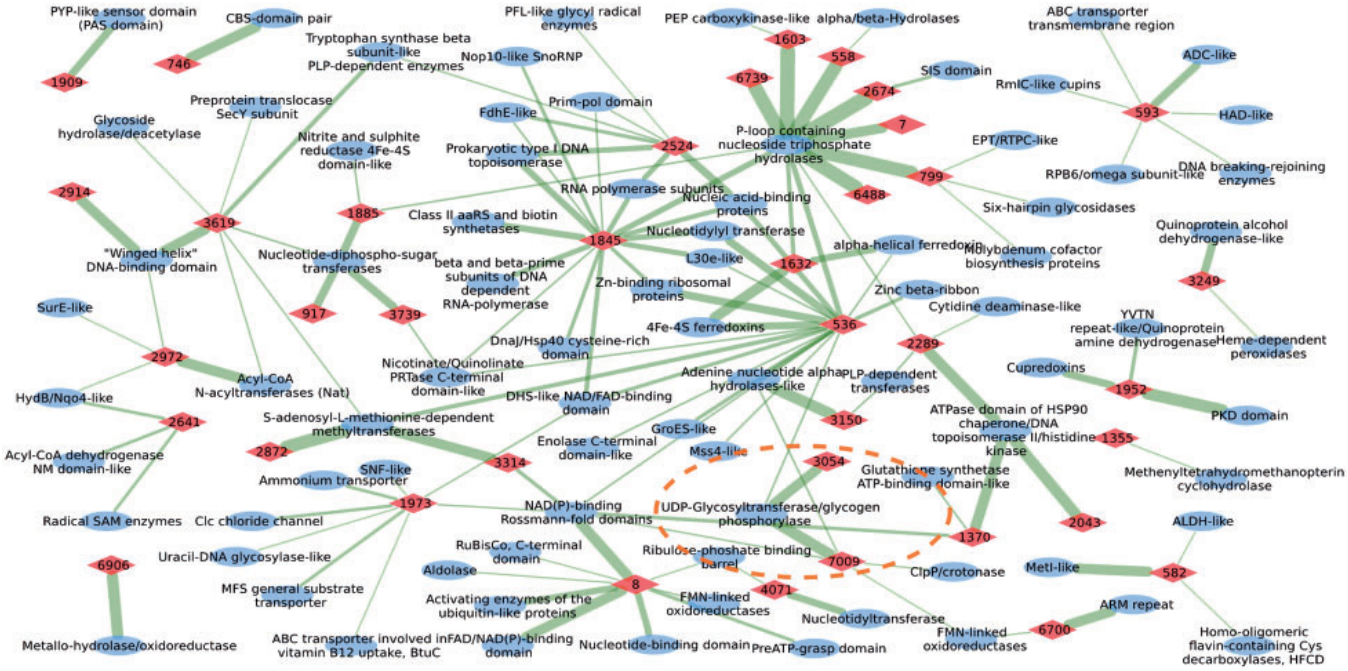

Figura 1.9: Cada superfamilia (elipses azules) está conectada a otra superfamilia si comparten algún fragmento (rombos naranjas). Los fragmentos se han obtenido a partir de perfiles construidos considerando proteomas completos de archaea junto con la hipótesis de la existencia de péptidos ancestrales, y posteriormente se han buscado similaridades en el espacio de estructuras [12]. Figura reproducida con permiso del editor.

Así que se han propuesto distintas aproximaciones para extraer fragmentos. La mayoría de las aproximaciones se basan en la búsqueda de motivos recurrentes a partir de una clasificación previa [9, 32, 33], acompañado de algún método para extraer estructuras consenso como técnicas de reducción de la dimensionalidad (**??**), o algún criterio más teórico como el mencionado tamaño típico a partir de la longitud de persistencia. Cada autor se refiere a los fragmentos con nombres diferentes, como *legos*, *Smotifs*, etc. Sorprendentemente, *con tan sólo* 70 *legos es posible cubrir el* 90 % *de los plegamientos conocidos* [32] lo que nos lleva a pensar en su utilidad en el modelado de proteínas. Un ejemplo de cómo se puede cubrir una proteína desconocida con fragmentos lo representamos en la Figura 1.10. Además, se observa que aquellos fragmentos más frecuentes tienen propiedades características como un mayor número de contactos internos, mientras que otros más raros contienen *loops* más largos. Se ha observado además que cuando se identifica un nuevo *fold*, no se identifican nuevos fragmentos sino una nueva combinación de los mismos [9].

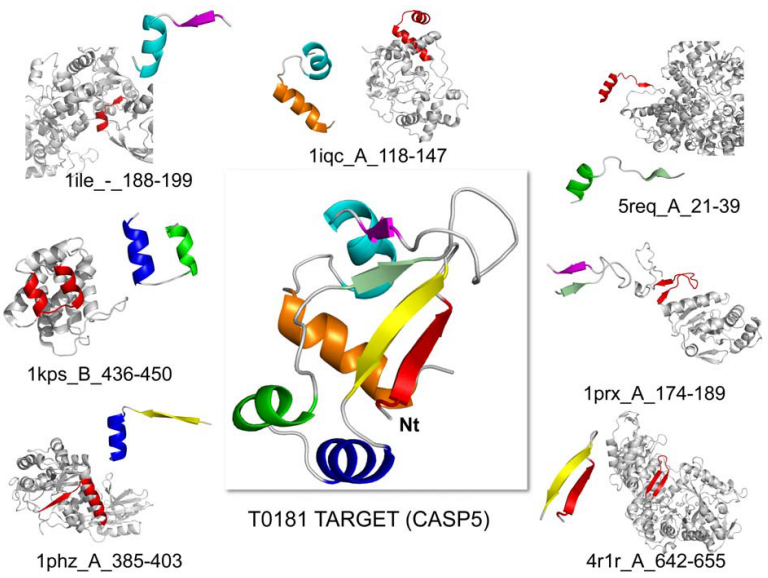

Figura 1.10: Cobertura de una proteína problema en la competición de CASP 5 mediante fragmentos encontrados en otras proteínas [9]. Hay que notar que en este caso algunos de los fragmentos no cumplen con el requerimiento de ser un loop cerrado, como se asume en algunas de las hipótesis evolutivas discutidas. Figura reproducida con permiso de los autores.

1.5. Divergencia estructural Vs. ensamblaje de módulos: consecuencias para la clasificación estructural y la modelización de estructura de proteinas

Si estos fragmentos pueden tener un rol claro funcional cabe preguntarse si la clasificación estructural tiene sentido ya que, *si toda las proteínas conocidas pueden ser construidas por un conjunto relativamente pequeño de fragmentos, deberíamos encontrar que muchos de los folds que pretendemos construir tienen similaridades locales entre sí*. Y esto es precisamente lo que se ha encontrado cuando se han comparado distintos *folds*: es posible relacionar muchos de ellos dentro de un umbral de similaridad significativo [27, 31], atribuibles a fragmentos de estructura supersecundaria. Si esto es así, a priori deberíamos de pensar que no es posible clasificar proteínas pues, como hemos explicado anteriormente, no vamos a encontrar ningún umbral dentro del cual los grupos sean realmente disjuntos. Esto será así a no ser que haya algún argumento desde el punto de vista evolutivo que justificara un "salto" en la similaridad y, a pesar de compartir localmente ciertos fragmentos, se pudiera decir que existe aún una separación clara entre los grupos debidas a similaridades globales. Este salto evolutivo estaría justificado por el hecho de que cualquier combinación de fragmentos ensamblados tiene una probabilidad baja de ser termodinámicamente estable. Si no fuese así, un escenario en el que el espacio es totalmente continuo y no es posible clasificar sería lo que esperaríamos. Así pues, *deberían existir ciertas combinaciones favorecidas, y estas combinaciones formarían los folds*, y las similaridades locales serían debidas o bien a que existen ciertos fragmentos ancestrales compartidos, o a mecanismos evolutivos más dramáticos como grandes inserciones correspondientes a fragmentos completos.

¿Es posible testar computacionalmente este escenario? Se han sugerido algunas aproximaciones para buscar de manera objetiva un salto en la similaridad de las estructuras de proteínas, con el objeto de determinar un umbral intrínseco en el cual es posible clasificar las estructuras. Como hemos señalado anteriormente, el disponer de clases de equivalencia estructurales con características comunes desde

el punto de vista evolutivo, funcional o ambos, sería de gran utilidad a la hora de modelar nuevas estructuras, ya que estas clases se pueden utilizar como plantillas a partir de las cuales construir los modelos.

Una de estas aproximaciones [8] se basa en la *teoría de redes complejas*, como ya hemos comentado anteriormente. Esencialmente el proceso consiste en *construir una red uniendo aquellas proteínas que sean más parecidas estructuralmente*, lo que generará en un primer momento grupos disjuntos. Al ir reduciendo la similaridad, algunos de los conjuntos irán creciendo y, eventualmente, algunos de estos grupos se irán uniendo entre sí. De nuevo surge la pregunta, ¿hasta cuándo juntar y, por tanto, en qué momento cortar el árbol que se va formando? En particular nos interesará el conjunto más grande ya que, si alcanza cierto tamaño crítico, nos puede indicar la existencia de una transición. En el trabajo que mencionamos se observó que existe una transición brusca del tamaño de este conjunto gigante, que se denomina percolación. La *percolación* es lo que ocurre por ejemplo cuando hacemos un café en una cafetera italiana. Mientras calentamos el agua, unas pequeñas burbujas de vapor atravesarán el café condensándose y formando pequeñas burbujas, que serían similares a nuestros conjuntos. A medida que sigamos calentando, el vapor encontrará ciertos caminos preferentes y algunos de estos conjuntos crecerán más que los demás hasta un punto en el cual el vapor percola, y atraviesa el café sin dificultad (momento en el cual el café empieza a fluir en el recipiente). Sabemos que, entre la formación de las primeras burbujas y el momento en el que el vapor fluye, existe un punto crítico en la transición que está bien descrito por la distribución de probabilidad del tamaño de las burbujas [28]. Y esta distribución es precisamente la misma que comentábamos en la Figura 1.4, una distribución libre de escala. Por tanto si, al construir nuestra red, encontramos un punto crítico que se pueda describir de manera similar, estaríamos ante una transición de fases similar a la del café pero en el universo de estructuras de proteínas. Este punto lo encontraríamos monitorizando también el tamaño del conjunto más grande con la esperanza de observar una transición clara. En este trabajo, se observó efectivamente una transición en el conjunto gigante. Y en el *punto de la transición*, además se obtuvo una *distribución libre de escala para el número de estructuras similares a cada una de las estructuras consideradas*, como se observa en la Figura 1.11.

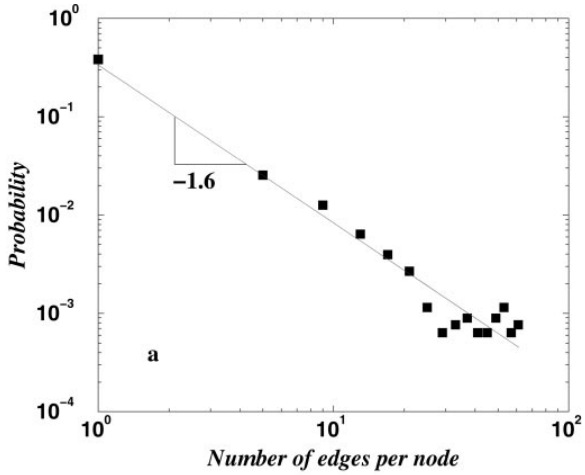

Figura 1.11: Distribución de probabilidad del número de relaciones de similaridad significativa que presenta cada estructura en la transición de percolación [8]. La distribución es consistente con la encontrada en la Figura 1.4. Figura reproducida con permiso del editor.

Observar esta transición es determinante en nuestra comprensión de la estructura del universo de proteínas, y ha sido posteriormente respaldada por otras aproximaciones. Por ejemplo, y como decíamos

anteriormente, si dos proteínas están relacionadas entre sí porque su distancia $d(a, b)$ es pequeña, esperaríamos que al compararlas con una tercera proteína se cumpliera el que $d(a, c) \approx d(b, c)$. Si esta relación no se cumple decimos que la *transitividad* está comprometida, es decir, el hecho de que estando a relacionada con b y estando b relacionada con c resulte que a y c no están relacionadas. El que la transitividad se respete permite formalmente definir clases de equivalencia, que es nuestro *proxy* a los conjuntos de estructuras de proteínas. Pues bien, basándose en esta definición de *clase de equivalencia*, se puede medir la violación de transitividad en un proceso de aglomeración que permite detectar también una transición consistente con el resultado anterior [27].

Esta transición es importante también porque nos indicaría que existe una *dualidad entre una concepción discreta y continua del espacio de estructura de proteínas* [30]. Para *similaridades altas*, tendríamos estructuras relacionadas preferentemente por duplicación génica y posterior divergencia. Para los lectores interesados, un modelo computacional que respalda esta hipótesis utilizando modelos sencillos de plegamiento de proteínas (como los explicados en el capítulo de plegamiento de proteínas (**??**) se describe en [36]. Encontraríamos grupos disjuntos que constituirían una clasificación objetiva de las estructuras. Para *baja similaridad* estructural (más allá del punto crítico), tendríamos en cambio que están relacionadas al nivel de fragmentos de estructura super-secundaria, y el mecanismo evolutivo que lo explicaría podría ser, o bien una consecuencia del modelo en el que hemos discutido la posibilidad de que existan fragmentos ancestrales, o bien por algún mecanismo de transferencia horizontal. Ambos escenarios son representados en la Figura 1.12.

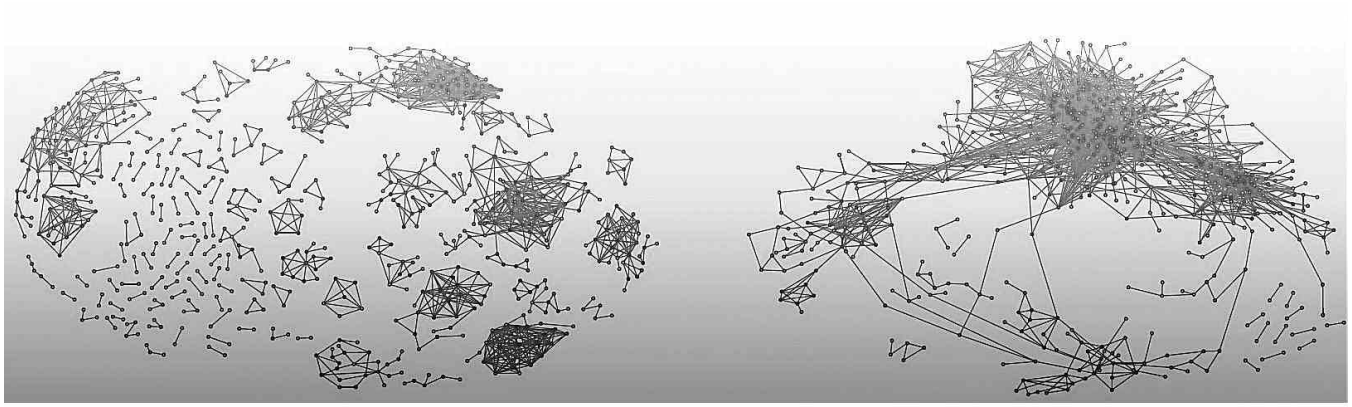

Figura 1.12: Representación artística de las relaciones de similaridad (enlaces) entre proteínas (nodos) para un umbral de significatividad alto (izquierda) compatible con un espacio discreto y por tanto con una clasificación estructural, y un umbral de significatividad más bajo (derecha) en el que las similaridades son locales y la representación es continua, solo compatible con una red y no con un árbol (detalles en [27]). Figura reproducida con permiso de los autores.

Sin embargo, desde el punto de vista de la modelización, el reciente interés por los fragmentos de estructura supersecundaria nos proporciona algunas novedades a tener en cuenta. Ya que si encontramos fragmentos de estructuras similares entre distintos *folds* que conservan la misma función, podemos estudiar la evolución en secuencia que han sufrido estos fragmentos conectándolos a través de una red neutral [11]. Una *red neutral* es una red que se construye relacionando secuencias que están separadas por mutaciones puntuales, y el lector interesado en profundizar en este campo encontrará un ejemplo sobre evolución neutral de RNA en la **??**. El relacionar a través de la secuencia estas redes condicionado a que compartan la misma estructura y función, nos abrirá muchas puertas para entender la evolución, ya que recorreremos distintos *folds* y organismos. Además, la estructura de la red nos muestra que hay algunas secuencias más conectadas que otras, lo que indica que son "sumideros" desde el punto de vista

evolutivo.

Cuando tengamos una secuencia cuya estructura desconocemos, ¿qué aproximación deberíamos seguir entonces? Pues haciendo uso de la dualidad que se observa en el espacio de estructura de proteínas, probablemente hay que apostar por *una aproximación que combine el uso de una clasificación estructural para intentar aproximarnos a una estructura con parecido global, y de una base de datos de fragmentos que nos permita refinarla.* Por tanto, el éxito de la modelización pasa sin duda por entender en profundidad los mecanismos evolutivos que subyacen, lo que abrirá puertas a nuevas aproximaciones que están aún por escribir.

Agradecimientos

El autor agradece la lectura crítica del manuscrito a Julián Echave, y la aportación de parte del material a Ugo Bastolla.

1.6. Bibliografía

[1] P. A. Alexander, Y. He, Y. Chen, J. Orban, and P. N. Bryan. A minimal sequence code for switching protein structure and function. *Proceedings of the National Academy of Sciences*, 106(50):21149–21154, Dec. 2009.

[2] R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, E. Birney, M. Biswas, P. Bucher, L. Cerutti, F. Corpet, M. D. R. Croning, R. Durbin, L. Falquet, W. Fleischmann, J. Gouzy, H. Hermjakob, N. Hulo, I. Jonassen, D. Kahn, A. Kanapin, Y. Karavidopoulou, R. Lopez, B. Marx, N. J. Mulder, T. M. Oinn, M. Pagni, F. Servant, C. J. A. Sigrist, and E. M. Zdobnov. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Research*, 29(1):37–40, Jan. 2001.

[3] I. N. Berezovsky, A. Y. Grosberg, and E. N. Trifonov. Closed loops of nearly standard size: common basic element of protein structure. *FEBS Letters*, 466(2-3):283–286, Jan. 2000.

[4] I. N. Berezovsky and E. N. Trifonov. Loop fold nature of globular proteins. *Protein Engineering*, 14(6):403–407, June 2001.

[5] J.-Q. Chen, Y. Wu, H. Yang, J. Bergelson, M. Kreitman, and D. Tian. Variation in the ratio of nucleotide substitution and indel rates across genomes in mammals and bacteria. *Molecular Biology and Evolution*, 26(7):1523–1531, July 2009.

[6] C. Chothia and A. M. Lesk. The relation between the divergence of sequence and structure in proteins. *The EMBO Journal*, 5(4):823–826, Apr. 1986. PMID: 3709526 PMCID: PMC1166865.

[7] A. L. Cuff, I. Sillitoe, T. Lewis, O. C. Redfern, R. Garratt, J. Thornton, and C. A. Orengo. The CATH classification revisited-architectures reviewed and new ways to characterize structural divergence in superfamilies. *Nucleic Acids Research*, 37(Database):D310–D314, Jan. 2009.

[8] N. V. Dokholyan, B. Shakhnovich, and E. I. Shakhnovich. Expanding protein universe and its origin from the biological big bang. *Proceedings of the National Academy of Sciences*, 99(22):14132–14136, Oct. 2002.

[9] N. Fernandez-Fuentes, J. M. Dybas, and A. Fiser. Structural characteristics of novel protein folds. *PLoS Comput Biol*, 6(4):e1000750, Apr. 2010.

[10] P. J. Flory and M. Volkenstein. Statistical mechanics of chain molecules. *Biopolymers*, 8(5):699–700, 1969.

[11] Z. M. Frenkel and E. N. Trifonov. From protein sequence space to elementary protein modules. *Gene*, 408(1-2):64–71, Jan. 2008.

[12] A. Goncearenco and I. N. Berezovsky. Prototypes of elementary functional loops unravel evolutionary connections between protein functions. *Bioinformatics*, 26(18):i497–i503, Sept. 2010.

[13] N. V. Grishin. Fold change in evolution of protein structures. *Journal of Structural Biology*, 134(2-3):167–185, May 2001.

[14] S. Henikoff, J. G. Henikoff, and S. Pietrokovski. Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations. *Bioinformatics*, 15(6):471–479, June 1999.

[15] T. A. Holland, S. Veretnik, I. N. Shindyalov, and P. E. Bourne. Partitioning protein structures into domains: Why is it so difficult? *Journal of Molecular Biology*, 361(3):562–590, Aug. 2006.

[16] L. Holm and C. Sander. Dali/FSSP classification of three-dimensional protein folds. *Nucleic Acids Research*, 25(1):231–234, Jan. 1997.

[17] M. A. Huynen and E. v. Nimwegen. The frequency distribution of gene family sizes in complete genomes. *Molecular Biology and Evolution*, 15(5):583–589, May 1998.

[18] C. Kimchi-Sarfaty, J. M. Oh, I.-W. Kim, Z. E. Sauna, A. M. Calcagno, S. V. Ambudkar, and M. M. Gottesman. A silent polymorphism in the MDR1 gene changes substrate specificity. *Science*, 315(5811):525–528, Jan. 2007.

[19] E. V. Koonin, Y. I. Wolf, and G. P. Karev. The structure of the protein universe and genome evolution. *Nature*, 420(6912):218–223, Nov. 2002.

[20] J. Lee and M. Blaber. Experimental support for the evolution of symmetric protein architecture from a simple peptide motif. *Proceedings of the National Academy of Sciences*, 108(1):126–130, Jan. 2011.

[21] A. N. Lupas, C. P. Ponting, and R. B. Russell. On the evolution of protein folds: Are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? *Journal of Structural Biology*, 134(2-3):191–203, May 2001.

[22] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of molecular biology*, 247(4):536–540, Apr. 1995. PMID: 7723011.

[23] S. OHNO. *Evolution by gene duplication.* Springer-Verlag, 1970.

[24] C. Orengo, A. Michie, S. Jones, D. Jones, M. Swindells, and J. Thornton. CATH - a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1109, Aug. 1997.

[25] C. Pál, B. Papp, and M. J. Lercher. An integrated view of protein evolution. *Nature Reviews Genetics*, 7(5):337–348, May 2006.

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

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

[28] F. Radicchi and S. Fortunato. Explosive percolation in scale-free networks. *Physical Review Letters*, 103(16):168701, Oct. 2009.

[29] C. G. Roessler, B. M. Hall, W. J. Anderson, W. M. Ingram, S. A. Roberts, W. R. Montfort, and M. H. J. Cordes. Transitive homology-guided structural studies lead to discovery of cro proteins with 40 % sequence identity but different folds. *Proceedings of the National Academy of Sciences*, 105(7):2343–2348, Feb. 2008.

[30] R. I. Sadreyev, B.-H. Kim, and N. V. Grishin. Discrete-continuous duality of protein structure space. *Current Opinion in Structural Biology*, 19(3):321–328, June 2009.

[31] J. Skolnick, A. K. Arakaki, S. Y. Lee, and M. Brylinski. The continuity of protein structure space is an intrinsic property of proteins. *Proceedings of the National Academy of Sciences*, 106(37):15690–15695, Sept. 2009.

[32] J. D. Szustakowski, S. Kasif, and Z. Weng. Less is more: towards an optimal universal description of protein folds. *Bioinformatics*, 21(Suppl 2):ii66–ii71, Oct. 2005.

[33] C.-H. Tai, V. Sam, J.-F. Gibrat, J. Garnier, P. J. Munson, and B. Lee. Protein domain assignment from the recurrence of locally similar structures. *Proteins: Structure, Function, and Bioinformatics*, 79(3):853–866, 2011.

[34] E. N. Trifonov and Z. M. Frenkel. Evolution of protein modularity. *Current Opinion in Structural Biology*, 19(3):335–340, June 2009.

[35] X. F. Wang and G. Chen. Complex networks: small-world, scale-free and beyond. *IEEE Circuits and Systems Magazine*, 3(1):6 – 20, 2003.

[36] K. B. Zeldovich, P. Chen, B. E. Shakhnovich, and E. I. Shakhnovich. A first-principles model of early evolution: Emergence of gene families, species, and preferred protein folds. *PLoS Comput Biol*, 3(7):e139, July 2007.