Nuevas aproximaciones computacionales para el estudio y la predicción funcional de dominios de proteínas

TESIS DOCTORAL

Daniel López López

Departamento de Biología Molecular
Facultad de Ciencias
Universidad Autónoma de Madrid

Julio 2013

Nuevas aproximaciones computacionales para el estudio y la predicción funcional de dominios de proteínas

Memoria que presenta para optar al
título de Doctor en Biología Molecular

Dirigida por el Doctor
Florencio Pazos Cabaleiro

Departamento de Biología Molecular
Facultad de Ciencias
Universidad Autónoma de Madrid

Julio 2013

A mis padres

"Estudia las frases
que parecen ciertas
y ponlas en duda."
Riesman, David

Agradecimientos

Cuando pienso en las personas que me han ayudado de alguna manera a hacer posible esta tesis doctoral me vienen a la cabeza multitud de nombres pero sin duda el mayor *culpable* de todo esto es mi director de tesis: Florencio Pazos. Gracias Sito por darme la oportunidad de entrar en tu grupo. Gracias por guiarme, enseñarme, corregirme y por tu gran paciencia para convencerme :). Sin ninguna duda estos años han sido una magnífica experiencia y de lo único que podría quejarme es de no haber podido aprender más de ti. Muchas gracias por todo.

Por supuesto también me gustaría agradecer a todos los compañeros con los que he coincidido durante estos años. Gracias por escuchar, criticar y aportar ideas para mejorar este trabajo. Gracias D. Ochoa también por tu ayuda técnica con R, LaTeXy en general estar dispuesto a echar una mano siempre con cualquier problema. Gracias Mónica por tu ayuda con las bases de datos y sobre todo con tu forma metódica y sencilla de sintetizar y exponer las ideas más complicadas. Gracias JC por mantener a punto las máquinas porque sin ellas difícilmente hubiéramos podido hacer algo. Gracias D. San León por el inigualable trabajo de β tester. Gracias Oli por tus cursos y por aportar siempre un punto de vista alternativo. Gracias a Trivi por su ayuda con la estadística, a Natalia por su ayuda con los predictores de desorden, a Toño por su ayuda con Java, a Jose Manuel, Dorota, Aldo, luis, Ponciano y Pablo por todo lo que pude aprender de vosotros. Y en general gracias a todos por los buenos momentos que hemos compartido, las conversaciones frikis y las discusiones/divagaciones de los cafés.

Si hay alguien que ha vivido esta tesis de cerca es sin duda Lorena. Muchas gracias por tus sugerencias con la tesis, por apoyarme y animarme en los momentos más duros, por aguantarme en los momentos más inaguantable pero sobre todo gracias por haber estado ahí siempre para cualquier cosa que he necesitado.

También me gustaría hacer un agradecimiento especial a la gente que he conocido a lo largo de estos años en el CNB, algunos de los cuales son y serán buenos amigos y que sin ninguna duda han hecho que mi estancia aquí sea

un verdadero placer. Muchas gracias.

No puedo acabar los agradecimientos sin mencionar a mis padres y mis hermanos. Esta tesis sencillamente no hubiera sido posible si no me hubieran *criado*, educado y apoyado en todo momento como lo han hecho. Muchas gracias por todo.

Lista de acrónimos

AA-TRNA	Aminoacil-tRNA
AJAX	JavaScript asíncrono y XML
AMP	Adenosín monofosfato
ATP	Adenosín trifosfato
AUC	Área bajo la curva ROC
DNA	Ácido desoxiribonucleico
EC	Comisión de enzimas
EF	Factor de elongación
FM-GO	Función molecular de GO
GDP	Guanosín bifosfato
GMP	Guanosín monofosfato
GO	Ontología de genes
GTP	Guanosín trifosfato
HTML	Lenguaje de marcado hipertextual
JSP	Página de servidor Java
MRNA	RNA mensajero
MSA	Alineamiento Múltiple de Secuencias
NGS	Secuenciación masiva (del inglés *Next generation sequencing*)

PARS ..... Análisis Paralelo de Estructura del RNA

PSSM ..... Matriz de puntuaciones específica de posición

RNA ..... Ácido ribonucleico

SDP ..... Posición determinante de especificidad

SNP ..... Polimorfismo de un nucleótido

tRNA ..... RNA transferente

μRNA ..... Micro RNA

VE ..... Velocidad de elongación de las proteínas (en la traducción)

Abstract

Obtaining experimental information on the structure, function and important residues for the proteins of a given organism is very time-consuming and expensive. For that reason, developing computational techniques for assigning functional features to protein sequences is an active area of research.

Almost all resources for predicting protein function assign functional terms to whole chains, and do not distinguish which particular domain is responsible for the allocated function. This is due to the fact that in the databases of functional annotations these methods use, these annotations are done on a whole-chain basis. Nevertheless, domains are the basic evolutionary and often functional units of proteins. Moreover, in many cases the domains of a protein chain have distinct molecular functions, independent from each other. For this reason, resources with functional annotations at the domain level, as well as methodologies for predicting function for individual domains adapted to these resources are required.

The main proposal of this thesis was to generate such two resources. We generated the first large-scale functional annotation at the domain level by annotating the SCOP structural domains with gene ontology terms. Additionally, we performed a large-scale comparison of these annotations with the ones implicit in the functional annotations of *InterPro* signatures, showing that the performance of this method is globally better.

Based on this database of functional annotations at the domain level, we developed a methodology for predicting the molecular function of individual domains and showed that this approach outperforms a standard method based on sequence searches in assigning functions. Additionally, we implemented this methodology on a web server for the concomitant prediction of fold, molecular function and functional sites at the domain level.

Although it is clear that the amino acid types are by far the main determinants of the functional features of proteins, several studies suggested that translational speed may also be playing a role in some cases.

However, a large scale comparative study on its relationship with a comprehensive diverse set of annotated functional features was missing. For that reason, we performed the first large scale analysis of the relationship between three experimental proxies of mRNA translation speed and the local features of the corresponding encoded proteins. We found that a number of protein functional and structural features are related to these mRNA properties. This results support the idea that the genome not only codes the protein functional features as sequences of amino acids, but also as subtle patterns of mRNA properties which, probably through local effects on the translation speed, have some consequence on the final polypeptide. Although the patterns found so far are in general very subtle, for particular cases with very clear patterns these could be used for predicting protein functional sites using single gene sequences. These results might have also implications for the heterologous expression of proteins.

Índice

Índice de figuras

Índice de Tablas

Capítulo 1

Introducción

Un microorganismo como *Escherichia coli* puede sintetizar a lo largo de su ciclo celular más de 4000 tipos de proteínas. El conjunto de proteínas expresadas en un determinado momento (proteoma) está implicado prácticamente en todos los procesos celulares, como la formación de estructuras, la comunicación, los mecanismos de señalización o la catálisis de transformaciones químicas de metabolitos (enzimas). Estudiar el funcionamiento y el papel que desempeña cada proteína en la célula no solo es interesante desde el punto del conocimiento básico, sino que puede ser útil para idear formas de modificación para nuestro beneficio.

Obtener experimentalmente este tipo de información es muy lento y costoso. La inmensa mayoría de las proteínas conocidas aún no han sido caracterizadas experimentalmente y se conoce poco sobre su función (Dimmer et al., 2012). Por el contrario, obtener la secuencia cruda de proteomas completos o parte de ellos es relativamente rápido y económico, y más aún con la llegada de tecnologías de secuenciación masiva (*Next-generation sequencing*) (Schuster, 2007). Consecuentemente, en la actualidad hay descritas más de 20 millones de secuencias de proteínas, donde en menos del 1 % de los casos se ha verificado experimentalmente su función.

La biología computacional ofrece herramientas que pueden dar indicios sobre la función de las proteínas basándose en su secuencia, estructura,

historia evolutiva o la asociación con otras proteínas. Dada la importancia y la enorme complejidad que presenta la predicción funcional de proteínas, ésta sigue siendo hoy en día un área activa de investigación.

1.1. Definición de función. Anotaciones funcionales

El primer problema que presenta la descripción funcional de una proteína se ilustra bien en el caso de la CbiF, que es una enzima implicada en la biosíntesis de la vitamina B12 (cobalamina). Concretamente, la CbiF transfiere un grupo metilo desde una S-adenosil-L-metionina al precursor de la vitamina B12 (cobalto-precorrina-4). La vitamina B12 es importante para el metabolismo, ayuda a la formación de glóbulos rojos, al mantenimiento del sistema nervioso central y también está implicada en la síntesis de DNA (MedlinePlus, 2005). Como puede verse, la función de la CbiF puede definirse desde distintos puntos de vista: molecular/encimático (metiltransferasa), metabólico (biosíntesis de cobalamina y DNA) y fisiológico (mantenimiento del sistema nervioso/tejido sanguíneo, a través de la B12). Este ejemplo ilustra cómo el concepto de "función de proteínas" es complejo y difícil de definir de manera unívoca.

Dado que la función de una proteína puede definirse de múltiples formas, los métodos predictivos deben también adaptarse a cada sistema de clasificación. Cuando se predice la función utilizando métodos automáticos a gran escala el problema se agrava por la necesidad de estandarizar y cuantificar la similitud de las funciones (Chagoyen y Pazos, 2010). Mientras que definir la similitud de secuencia o estructura puede ser relativamente sencillo, no hay *a priori* una medida sencilla para cuantificar la similitud entre las funciones de dos proteínas.

Para abordar este problema se han construido diferentes sistemas de clasificación u ontologías de funciones biológicas. Por ejemplo, la EC establece una clasificación jerárquica de las reacciones catalizadas por enzimas mediante cuatro dígitos (número EC). Cada dígito representa una

descripción específica de la enzima o de su actividad. En principio, las enzimas con más números EC en común empezando desde el primer dígito, que define clases enzimáticas generales, hasta el cuarto, que define sustratos específicos, deberían tener funciones más parecidas. Sin embargo, enzimas con los tres primeros dígitos idénticos pueden tener diferencias significativas en el proceso catalítico (Almonacid et al., 2010).

El proyecto GO (Harris et al., 2004) es una alternativa más general, ya que define un vocabulario controlado compuesto de términos que describen las propiedades de genes y productos de genes que se puede aplicar también a proteínas no enzimáticas. En la actualidad, GO se ha convertido en el estándar para representar la función de las proteínas mediante la asociación de uno o varios términos GO a éstas. Los términos GO se engloban en tres categorías que describen la localización celular (*cellular component*), el proceso biológico (*biological process*) o la función molecular (*molecular function*) específica. De hecho, GO define tres ontologías independientes donde sus términos establecen relaciones hijo-padre formando un grafo acíclico dirigido que se puede navegar desde funciones generales (p.e. "Enzima") hasta más específicas (p.e. "fosfolipasa A1"). De esta forma, una secuencia asociada a un término GO estará inherentemente asociada a todos sus términos padre. Asimismo, GO define códigos de evidencia para las anotaciones (asociaciones entre un término GO y una secuencia) distinguiendo la fuente de la anotación, según sea experimental, computacional, revisada por expertos o inferida automáticamente. Hay varios métodos para cuantificar la similitud funcional entre dos proteínas anotadas con términos GO teniendo en cuenta los términos que comparten así como la posición de éstos en el grafo de GO.

Quizás la principal limitación de GO cuando se aplica a proteínas es el carácter no posicional de los términos GO. Por ejemplo, no hay relación entre los términos GO que definen la actividad catalítica de una enzima y el centro activo donde se produce la reacción. Otras recientes ontologías como *Protein Feature Ontology* (Reeves et al., 2008) tratan de abordar este aspecto.

1.2. Métodos computacionales de predicción de función de proteínas

Tradicionalmente, "función protéica" hace referencia a la función molecular de una secuencia, como la actividad catalítica de una enzima, actividades de transporte y señalización de proteínas de membrana o la función de soporte de proteínas estructurales (p.e. citoesqueleto), entre otras. Estas funciones están englobadas dentro de la categoría FM-GO (Función molecular de GO). Dado que este aspecto concreto de la función lo determina únicamente la secuencia y la estructura, se han desarrollado multitud de metodologías y herramientas bioinformáticas para predecir la función molecular de las proteínas en base a estas propiedades.

1.2.1. Métodos basados en similitud de secuencia

La forma más sencilla de predecir la función molecular de una secuencia es mediante transferencia por homología (Frishman, 2007; Lee et al., 2007; Raes et al., 2007). La técnica consiste básicamente en alinear la secuencia de interés contra una base de datos de proteínas anotadas, como SWISS-PROT (UniProt Consortium, 2010), para encontrar un homólogo de función conocida del cual se transfieren las anotaciones funcionales a la primera. Entre los métodos de alineamiento más populares se encuentran PSI-BLAST (Altschul et al., 1997) y HMMER (Bateman et al., 1999). Una mejor estrategia consiste en tener en cuenta las anotaciones de todos los homólogos y transferir sólo las anotaciones estadísticamente significativas (enriquecimiento funcional) (Martin et al., 2004; Götz et al., 2008; Hawkins et al., 2009).

Sin embargo la transferencia por homología debe ser implementada con precaución. A menudo homología se confunde con similitud de función. En realidad, homología entre dos proteínas simplemente significa que tienen un origen evolutivo común, pero ello no implica necesariamente que hayan mantenido las propiedades desde entonces. En este contexto es conveniente

además distinguir entre genes ortólogos y parálogos. Los ortólogos son genes homólogos diferenciados por el proceso de especiación, mientras que los parálogos son el resultado de una duplicación génica (Fitch, 2000). Está generalmente asumido que los ortólogos tienen la misma función biológica en diferentes especies (Tatusov et al., 1997). En cambio los eventos de duplicación serían un factor clave en la evolución, ya que permiten a los parálogos adquirir nuevas funciones (Magadum et al., 2013). Por ello la función tiende a estar más conservada entre ortólogos que entre parálogos (Theissen, 2002). Varios estudios han cuantificado la relación entre parecido en secuencia y parecido en funcional (definido según diversos criterios) (Wilson et al., 2000; Devos y Valencia, 2000). Aunque una alta similitud de secuencia aumenta la confianza en la transferencia funcional de anotaciones, no hay un límite que pueda considerarse seguro. De hecho, pequeños cambios en la secuencia pueden a veces provocar cambios radicales en las propiedades de la función, como cambios en la actividad enzimática, o incluso la pérdida o adquisición de la actividad enzimática propiamente dicha (fig. 1.1). Para mitigar el problema de los parálogos, algunas herramientas infieren las relaciones evolutivas entre familias de proteínas a costa de aumentar considerablemente el tiempo de computación (Engelhardt et al., 2005).

La naturaleza multidominio de muchas proteínas también puede ser una fuente importante de errores. Generalmente, los dominios protéicos se definen como unidades evolutivas y estructuralmente independientes de las proteínas (Doolittle y Bork, 1993; Wetlaufer, 1973). Además, en la mayoría de los casos llevan a cabo funciones moleculares independientes hasta el punto de que incluso aparecen como proteínas aisladas en otros organismos (Marcotte, 1999). Las anotaciones funcionales de las bases de datos que almacenan secuencias completas, como SWISS-PROT (UniProt Consortium, 2010), no distinguen cuál es el dominio responsable para una determinada función, sino que asignan ésta a la cadena completa. Si la proteína cuya función queremos predecir no alinea con el dominio responsable de la función transferida sino con otro, la transferencia de esa anotación estaría totalmente injustificada.

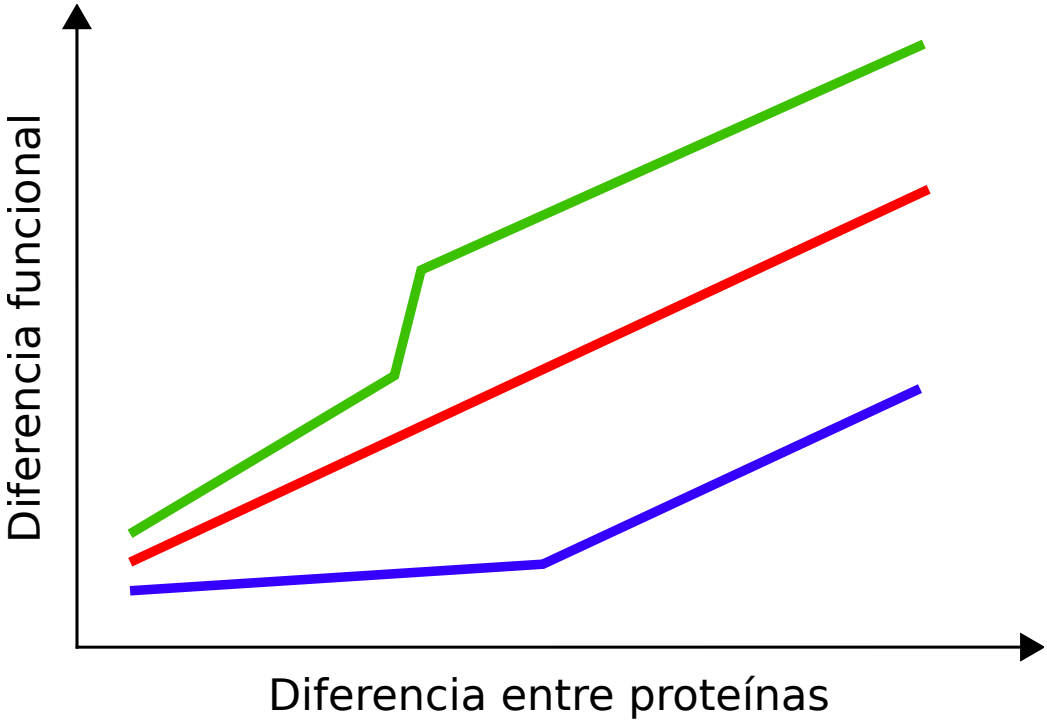


Figura 1.1: ***Relación entre similitud y función protéica****. El eje x representa la distancia entre dos proteínas, en términos de secuencia, estructura o alguna otra propiedad observable. El eje y es la distancia entre las mismas proteínas en términos de función biológica. Normalmente, los métodos de anotación asumen que cuanto más similares son dos proteínas mayor es la probabilidad de que compartan la función (línea roja). Pero esos cambios no tienen por qué ser graduales: la linea verde ilustra pequeñas variaciones en proteínas que provocan cambios sustanciales en su función molecular. La línea azul ilustra un ejemplo opuesto donde proteínas distantes llevan a cabo funciones bioquímicas estrechamente relacionadas. Figura adaptada de Erdin et al.*

La transferencia de función desde proteínas multifuncionales también puede provocar predicciones incorrectas. Esto es especialmente evidente en el caso de proteínas multifuncionales dependientes del contexto celular, como las metaloproteínas que unen distintos iones metálicos según la localización celular (Tottey et al., 2008). Un caso extremo lo forman las llamadas "*moonlightting proteins*", que son proteínas que llevan a cabo múltiples funciones significativamente diferentes (Jeffery, 1999, 2004). Por ejemplo, la cristalina-η es una proteína que juega un papel estructural en el cristalino de algunas especies mientras que actúa como una enzima en otros tejidos. Los homólogos de estas proteínas podrían retener sólo algunas de las funciones originales (Bateman et al., 2003). Como consecuencia, la transferencia funcional de anotaciones podría ser errónea o incompleta.

Finalmente, hay que tener en cuenta que las bases de datos de anotaciones de proteínas contienen errores, la mayoría de ellos causados por la incorrecta transferencia automática de anotaciones por homología (Linial, 2003). Actualmente, casi todas las anotaciones derivan de un pequeño grupo de proteínas (probablemente $\leq 5\,\%$) cuya información se ha obtenido directamente de forma experimental (Valencia, 2005). La aplicación sistemática de protocolos de anotación funcional automática tiene como consecuencia la propagación de errores que se ha llegado incluso a modelizar matemáticamente (Gilks et al., 2002).

1.2.2. Métodos basados en patrones

En algunos casos, una pequeña región de la secuencia puede ser suficiente para conservar la función de una proteína incluso si el resto de la proteína ha cambiado considerablemente a lo largo de la evolución. Por otra parte, proteínas no homólogas podrían haber adquirido el mismo motivo funcional de forma independiente (evolución convergente). Así, dos proteínas que no se podrían encontrar en una búsqueda de secuencia completa podrían aún así tener en común ciertos motivos que revelasen su relación funcional. Obviamente, si además de compartir los motivos tienen un determinado nivel de similitud global, aumenta la confianza de la

transferencia de anotaciones funcionales.

Existen muchas herramientas computacionales dedicadas a la identificación de motivos funcionales (Punta y Ofran, 2008; Attwood et al., 1999; Henikoff y Henikoff, 1996; Falquet et al., 2002; Puntervoll et al., 2003). Estos recursos normalmente ofrecen una colección de motivos generados manualmente por expertos o de forma automática empleando algoritmos de búsqueda de patrones. Quizás el más usado es *InterPro* (Hunter et al., 2012), que recopila motivos funcionales de proteínas de 11 bases de datos entre las que se incluyen algunas clasificaciones de familias de dominios como Pfam (Punta et al., 2012) o SUPERFAMILY (de Lima Morais et al., 2011). El meta-servidor *InterProScan* (Quevillon et al., 2005) permite hacer búsquedas por similitud de una secuencia problema contra estas bases de datos. Encontrar un motivo bien caracterizado en la proteína de interés podría ayudar a predecir su función.

Localización de residuos funcionales basada en conservación

Los residuos que son cruciales para la función de la proteína tienden a estar conservados entre proteínas homólogas. A menudo éstos se pueden identificar por patrones de conservación en MSA (Alineamiento Múltiple de Secuencias) de proteínas de la misma familia (Wallace et al., 2005; Notredame, 2007; Pietrosemoli et al., 2012). Obviamente, esta aproximación sólo es posible cuando están disponibles múltiples homólogos de la proteína de interés. Identificar los residuos conservados puede ser interesante incluso aunque no se conozca la función específica de esos residuos en la familia de proteínas, por ejemplo en experimentos de mutagénesis dirigida.

Pese a que la identificación de posiciones conservadas en un MSA podría parecer un problema trivial a primera vista, no hay una única forma de hacerlo. De hecho hay diversas aproximaciones que producen distintos resultados (Valdar, 2002). En la mayoría de los casos, esto es debido a que la conservación en los MSA no suele ser perfecta, es decir, rara vez se refleja en una columna del MSA con un sólo tipo de aminoácido. Esta conservación imperfecta puede deberse a múltiples factores, entre los que se

incluyen errores en el alineamiento y sustituciones conservativas (cambios entre aminoácidos con propiedades fisico-químicas similares). Por otra parte, la similitud global de las proteínas también debe tenerse en cuenta cuando se evalúa la conservación de una posición concreta. Si las proteínas son en general muy parecidas, la observación de una posición conservada es menos indicativa de su funcionalidad. Una de las mejores metodologías para detectar residuos funcionales en MSA está implementada en el servidor Conseq/Consurf (Armon et al., 2001), que tiene en cuenta la filogenia de las secuencias del MSA.

Otro patrón de conservación más sutil lo representan las posiciones diferencialmente conservadas en las distintas subfamilias del alineamiento, en caso de haberlas. Estas posiciones conservadas pero con un aminoácido distinto en cada subfamilia se han relacionado con determinantes de especificidad funcional, y por ello se conocen como SDP (Rausell et al., 2010).

En la mayoría de los casos, los residuos funcionales que forman parte del centro activo de una proteína están localizados relativamente cerca en el espacio y en su superficie para poder interaccionar con las moléculas sobre las que actúa la proteína. Consecuentemente, si está disponible la estructura tridimensional de alguno de los miembros de la familia de proteínas homólogas, se puede estudiar la agregación espacial y la exposición a la superficie de las posiciones conservadas o SDP. Esta información adicional se puede utilizar para filtrar los resultados y mejorar las predicciones.

1.2.3. Métodos basados en estructura

Un segundo tipo de medida de similitud entre proteínas se basa su estructura tridimensional. Al igual que ocurre con las secuencias, los métodos de predicción funcional basados en estructura asumen que ciertos aspectos e intervalos de similitud estructural pueden indicar similitudes funcionales.

La información estructural puede ayudar a predecir la función de

distintas maneras. Una alta similitud estructural entre dos proteínas puede revelar un origen evolutivo común. Dado que la estructura está mucho más conservada que la secuencia (Levitt y Gerstein, 1998), la información estructural permite detectar la homología incluso en ausencia significativa de similitud de secuencia. Por otro lado, esto mismo también provoca que las proteínas con la misma estructura puedan tener orígenes evolutivos y funcionales totalmente diferentes. Dos proteínas con la misma conformación espacial general e incluso residuos funcionales conservados (Whisstock y Lesk, 2003) pueden tener funciones no relacionadas. Del mismo modo, dos proteínas pueden tener la misma función a pesar de tener diferentes estructuras (Bartlett et al., 2003) (fig. 1.1).

La similitud estructural también puede ser útil para identificar fenómenos de convergencia evolutiva entre dos proteínas a causa de alguna restricción funcional común. Los efectores de virulencia procarióticos ofrecen un buen ejemplo de convergencia funcional. Algunas de estas proteínas se han adaptado para imitar las proteínas del hospedador y poder así interferir en su proceso biológico. Esto se consigue adoptando la misma estructura global o, más a menudo, características estructurales locales (Desveaux et al., 2006; Stebbins y Galán, 2001).

Se han desarrollado numerosos métodos para realizar comparaciones estructurales entre proteínas (Holm y Rosenström, 2010). Algunos utilizan bases de datos de clasificación estructural de dominios como SCOP (Andreeva et al., 2004) o CATH (Pearl et al., 2005).

Otros métodos predictivos se centran en la comparación de características estructurales locales. Algunas herramientas se basan en similitudes de geometría local de cavidades y *pockets* o de cargas electrostáticas en la superficie para realizar transferencias funcionales (Tseng et al., 2009; Brylinski y Skolnick, 2008; Shulman-Peleg et al., 2008; Kinoshita et al., 2002; Glaser et al., 2005; Laskowski, 1995). Otras utilizan colecciones de plantillas tridimensionales compuestas por unos pocos residuos funcionales con una geometría espacial definida (Polacco y Babbitt, 2006; Laskowski et al., 2005). Por ejemplo, los residuos Ser-His-Asp de la tríada catalítica de las serín proteasas no están seguidos

en la secuencia de aminoácidos y podrían ser difíciles de localizar en un análisis de secuencia. Aún así, su plantilla 3D podría encajar geométricamente en otras estructuras proteicas haciendo posible la identificación de nuevas proteasas (Watson et al., 2007).

1.2.4. Otros métodos de predicción funcional

Dado que existen diferentes métodos de predicción funcional basados en aproximaciones muy diferentes, lo esperable es que la predicción funcional mejore cuando se combinan distintos métodos. Este es el principio en el que se basan meta-servidores como ProFunc (Laskowski et al., 2005) o ProKnow (Pal y Eisenberg, 2005).

La mayoría de los métodos de predicción funcional de proteínas se basan en la transferencia de anotaciones desde proteínas similares en secuencia de función conocida. Sin embargo, Â¿qué alternativas existen cuando la proteína cuya función queremos predecir no tiene una similitud significativa con otras proteínas ya anotadas? Se han sugerido distintas aproximaciones para predecir la función proteica *de novo*. En lugar de basarse en la similitud con una proteína específica, estos métodos utilizan propiedades "genéricas" de la estructura y la secuencia comunes a proteínas con la misma función. De hecho, las proteínas con la misma función están sujetas a restricciones comunes (p.e. propiedades de un ligando, flexibilidad estructural) que pueden verse reflejadas en su estructura y su secuencia. Los métodos *de novo* normalmente se basan en algoritmos de aprendizaje automático (*machine learning algorithms*) capaces de capturar correlaciones significativas no triviales entre características y funciones (Jensen et al., 2002). Estos métodos normalmente son menos precisos que los de transferencia de anotaciones pero tienen mayor alcance, ya que permiten predicciones sobre regiones inexploradas del espacio de secuencias y posibilitan la anotación de genomas enteros.

1.3. Efectos de la velocidad de elongación en las proteínas

La redundancia del código genético permite que la mayoría de los aminoácidos sean codificados por más de un triplete de nucleótidos (codones sinónimos). El uso de codones sinónimos juega un papel clave en la regulación de la transcripción del mRNA, y ofrece a la evolución un mecanismo de control independiente de la secuencia de aminoácidos codificada. Por ejemplo, algunos cambios sinónimos en codones del mRNA pueden alterar los sitios de unión de μRNA reguladores, afectando así a la concentración total de mRNA y por tanto a los niveles de expresión global de la proteína (Brest et al., 2011). Asimismo, el intercambio entre codones sinónimos podría producir transcritos con diferente estructura secundaria y terciaria al modificar el patrón de apareamientos entre los nucleótidos. De hecho, pequeños cambios en la secuencia del mRNA pueden provocar grandes cambios estructurales (Chursov et al., 2012). Esto podría repercutir en la unión de los ribosomas y afectar a la VE (Velocidad de elongación de las proteínas (en la traducción)), ya que el mRNA debe ser desplegado para poder ser traducido por los ribosomas (Kudla et al., 2009). Finalmente, otra consecuencia de usar diferentes codones sinónimos para un determinado aminoácido está relacionada con la disponibilidad de los distintos aa-tRNA (Aminoacil-tRNA). Dado que cada aa-tRNA difiere en su concentración citoplasmática y por tanto en la disponibilidad para el ribosoma, la selección entre codones sinónimos afecta directamente la VE (Varenne et al., 1984; Sorensen, 2001). De hecho se ha visto que proteínas altamente expresadas están enriquecidas en codones "óptimos" (reconocidos por los aa-tRNA más abundantes) (Akashi y Eyre-Walker, 1998; Ernst, 1988). La relación entre el uso de codones y la concentración de tRNA juega un papel central en numerosos sistemas biológicos (Frenkel-Morgenstern et al., 2012).

Como se ha visto, el uso diferencial de codones sinónimos afecta al proceso global de la traducción. Esto sirve para regular globalmente la VE, afectando

así a la concentración final de la proteína. Sin embargo el uso diferencial de codones sinónimos también podría estar jugando un papel local en la traducción. En principio sería posible modular la VE de regiones concretas del mRNA seleccionando apropiadamente entre los distintos codones sinónimos. De hecho Dana y Tuller demostraron que la VE sufre grandes variaciones a lo largo del mRNA, influenciada por factores locales como la disponibilidad de aa-tRNA, la estructura secundaria del mRNA o la carga de los aminoácidos codificados (Dana y Tuller, 2012).

Estas variaciones locales de la VE podrían por tanto tener efectos en las proteína traducida de forma local. Se han realizado numerosos estudios que relacionan el uso diferencial de codones sinónimos y determinadas características estructurales locales de las proteínas. Algunas de estas características incluyen las regiones transmembrana (Power et al., 2004), dominios funcionales (Chartier et al., 2012), residuos con tendencia a la agregación (Lee et al., 2010) o determinados elementos de estructura secundaria (Li et al., 2012; Saunders y Deane, 2010; Thanaraj y Argos, 1996). En algunos casos, esta asociación podría reflejar un requerimiento en términos de VE. Por ejemplo, la modulación local de la VE podría permitir el correcto plegamiento de segmentos ya traducidos "frenando" la traducción de otros que podrían interferir con ellos y dándoles, por tanto, tiempo a plegarse (asumiendo el modelo aceptado de plegamiento co-traduccional) (Komar et al., 1999; Tsai et al., 2008). Asimismo, podría jugar un papel importante en el inicio de la transcripción. El hecho de que los primeros 20-30 codones se traduzcan más lentamente que el resto del mRNA, lo que se conoce como "rampa traduccional", podría estar relacionado con la correcta unión del ribosoma y el inicio del plegamiento del polipéptido emergente (Tuller et al., 2010; Dana y Tuller, 2012). Más aún, algunos estudios sugieren que la VE podría incluso estar implicada en algunos procesos de modificación post-traduccional (Zhang et al., 2009). Modificaciones como la miristilación (Wilcox et al., 1987) o la glicoxilación que ocurren durante la traducción, podrían requerir una determinada VE en las proximidades.

1.3.1. Relación entre VE y regiones funcionales de proteínas

Algunos estudios sugieren que la alteración artificial de la VE podría tener efectos significativos en la función de las proteínas codificadas. Por ejemplo Kimchi-Sarfaty et al. encontraron que una modificación de un solo nucleótido (SNP) en el gen MDR1 tenía un efecto drástico en la función de la proteína que codifica a pesar de que la mutación no alteraba la secuencia final de ésta (SNP sinónimo) (Kimchi-Sarfaty et al., 2007). Una posible explicación sugiere que el SNP estaría provocando una modificación local de la VE que afectaría a la estructura del sitio de unión de la proteína y por tanto a su función. Asimismo, Agashe et al. estudiaron los efectos que provocan los cambios en codones sinónimos del mRNA que codifica la enzima bacteriana FAE. Estos autores encontraron que, además de modificarse la concentración total de proteína sintetizada, los distintos mRNA sinónimos producían proteínas con actividad enzimática muy diferente (Agashe et al., 2012). De nuevo la alteración funcional podría deberse a cambios locales en la enzima posiblemente provocados por la modulación local de la VE.

Trabajos anteriores han estudiado la relación entre diferentes propiedades locales del mRNA relacionadas con la VE y aspectos estructurales de las proteínas. Sin embargo, hasta donde sabemos, nunca se ha hecho un estudio a gran escala que incluya anotaciones funcionales. La posible correlación entre VE y determinadas características funcionales de las proteínas podría ser la base para el desarrollo de métodos de predicción funcional *de novo* y tendría implicaciones para el diseño de mRNA para expresión heteróloga.

Capítulo 2

Objetivos

Conocer la función que desempeña cada proteína en un organismo es de vital importancia para comprender su biología. Además, este tipo de información se hace necesaria para diseñar maneras de modificar este proteoma en nuestro propio beneficio. Debido a la dificultad y el coste de la caracterización funcional y estructural de la proteínas utilizando métodos experimentales, el desarrollo de métodos computacionales de predicción sigue siendo hoy en día un área activa de investigación.

A pesar de que los dominios son las unidades funcionales, evolutivas y estructurales de las proteínas, la mayoría de estos métodos de predicción están diseñados para anotar cadenas completas mediante transferencia por homología. Para la mayoría de las aplicaciones esto no supone un problema. Sin embargo, en algunos casos resulta necesario distinguir los dominios concretos que realizan una determinada función molecular. Además, estos métodos están limitados por la disponibilidad de secuencias homólogas de características funcionales conocidas en las bases de datos. Por ello, resulta necesario seguir explorando métodos alternativos no basados en transferencia por homología.

En base a estos antecedentes y con el objetivo de entender, caracterizar y ayudar a predecir la función de las proteínas, nos propusimos los siguientes objetivos:

1. **Desarrollar la primera base de datos de anotaciones funcionales de dominios.** La base de datos debe generarse automáticamente a partir de recursos previamente desarrollados orientados a cadenas completas. Se debe idear un método de evaluación de dicho recurso. Adicionalmente, es necesario crear una interfaz web para permitir el acceso a este recurso.

2. **Implementar un método de predicción funcional de dominios.** El método debe basarse en una base de datos de anotaciones de dominios como la del objetivo anterior. Además, se debe diseñar un método para evaluar la capacidad de predicción, la sensibilidad y la especificidad del método.

3. **Construir una aplicación web para la predicción funcional de dominios.** La aplicación debe permitir la predicción funcional de dominios a partir de una secuencia de aminoácidos. La interfaz debe ser intuitiva, rápida y compatible con la mayoría de los navegadores. La aplicación web debe al mismo tiempo estar adaptada para utilizar distintas anotaciones funcionales a nivel de dominio.

4. **Estudiar la relación entre la variación local de la VE y determinados aspectos funcionales de las proteínas.** Para ello, se deben utilizar datos experimentales masivos de propiedades del mRNA relacionadas con la VE que proporcionen la suficiente resolución como para poder estudiar los efectos locales en las correspondientes proteínas.

Capítulo 3

Materiales y métodos

3.1. Anotación funcional de proteínas a nivel de dominio

El proyecto *GOA-PDB* (Dimmer et al., 2012) tiene como objetivo anotar todas las proteínas con estructura 3D depositada en *PDB* (Berman et al., 2000) mediante la asignación de términos GO (Harris et al., 2004) transferidos desde *Uniprot* (UniProt Consortium, 2010) (véase Apéndice A.2, Bases de datos). Estas anotaciones están pensadas para describir las funciones de las cadenas de forma global, sin tener en cuenta qué dominios estructurales son realmente necesarios para llevar cabo una determinada FM-GO.

SCOP2GO es la primera base de datos de anotaciones funcionales de proteínas a nivel de dominio. Dichas anotaciones se generan automáticamente mediante la transferencia de términos de la categoría FM-GO asignados inicialmente en *GOA-PDB* a una cadena completa, al dominio/s específico/s que lleva a cabo dicha FM-GO. Los siguientes apartados describen el procedimiento empleado para la generación automática de *SCOP2GO*, las diferentes aproximaciones que se han utilizado para evaluar la precisión de las anotaciones y la estructura del servidor web desarrollado para facilitar al usuario el acceso a este recurso.

3.1.1. Creación de la base de datos *SCOP2GO*

La base de datos *SCOP2GO* está basada en las anotaciones funcionales de *GOA-PDB*. Estas anotaciones consisten en asociaciones entre términos GO y cadenas completas de *PDB*. El proceso de creación de la base de datos *SCOP2GO* empieza con la extracción del subconjunto de cadenas *PDB* anotadas en *GOA-PDB* con un determinado término GO de la categoría FM-GO (fig. 3.1). El objetivo consiste básicamente en encontrar los dominios responsables de la FM-GO para todas estas cadenas.

Para ello, primero se obtienen los plegamientos de *SCOP* (véase Apéndice A.2, Bases de datos) de cada cadena anotada con el término GO. Algunas de estas cadenas pueden compartir dominios con el mismo plegamiento (fig. 3.1). El método asume que el plegamiento más frecuente en este conjunto de cadenas es, con mayor probabilidad, el responsable de llevar a cabo la FM-GO originalmente asociada a las cadenas completas. Los dominios asociados a ese plegamiento se anotan con ese término FM-GO y el proceso se itera hasta que todas las cadenas tengan al menos un dominio asociado a la FM-GO. En cada ciclo se asigna el término GO al plegamiento más frecuente y se descartan las cadenas con ese plegamiento. Los dominios de las cadenas descartadas no se tienen en cuenta a la hora de calcular las frecuencias de los plegamientos en los siguientes ciclos. Si dos o más plegamientos tienen una distribución similar (97 % de co-ocurrencia en misma la cadena), ambos dominios son asociados al término GO (p.e. pentágono-hexágono en la fig. 3.1). En este caso el método asumiría que es necesaria la presencia de más de un dominio para llevar a cabo la función.

Este proceso se puede entender como la búsqueda del mínimo número de plegamientos necesarios para explicar el hecho (observado) de que todas las cadenas del subconjunto tengan esa FM-GO. El proceso se repite para el resto de FM-GO cuya distancia a la raíz del árbol de GO es superior a dos, acumulando las anotaciones para cada dominio o conjunto de dominios.

La significación estadística de cada asociación plegamiento/función se

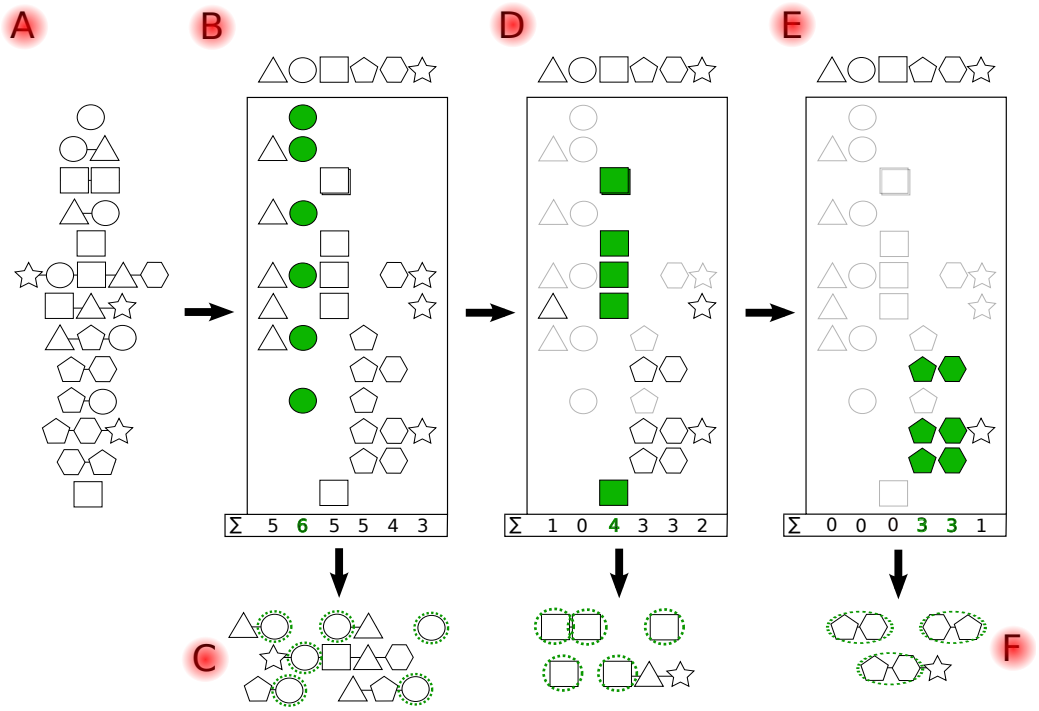


Figura 3.1: ***Esquema del método SCOP2GO.*** *(**A**) El punto de partida es el conjunto de cadenas PDB asociadas a una determinada FM-GO. Las formas representan distintos plegamientos según SCOP. (**B**) Se construye una matriz de cadenas y plegamientos, y se localiza el plegamiento predominante (círculo). (**C**) El término FM-GO se asigna a los dominios con ese plegamiento. (**D**) Las cadenas con algún dominio anotado se descartan y se excluyen para los siguientes recuentos. Se localiza el siguiente plegamiento con mayor frecuencia (cuadrado) y se repite el proceso anterior. (**E**) Cuando dos plegamientos tienen la misma frecuencia (pentágono-hexágono) ambos se asocian con la FM-GO (**F**).*

calculó en base a la distribución hipergeométrica

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \tag{3.1}$$

donde N denotaría el número total de cadenas en *SCOP* y n, K y k el subconjunto de cadenas con algún dominio clasificado con ese plegamiento, esa función o con el plegamiento y la función al mismo tiempo. De esta forma cada anotación lleva asociado un valor-P que refleja el grado de confianza de la asignación (véase Apéndice A.3, Pruebas estadísticas y normalizaciones).

3.1.2. Evaluación de las anotaciones funcionales de *SCOP2GO*

SCOP2GO es la primera base de datos de anotaciones funcionales de proteínas a nivel de dominio. Al no existir un recurso similar revisado por expertos o una referencia estándar con la que compararla, es imposible llevar a cabo una evaluación exhaustiva de las anotaciones. En los siguientes apartados se describen las distintas aproximaciones que se han empleado para evaluar parcialmente la base de datos.

3.1.2.1. Generación del conjunto de prueba *InterPro2GO*

Las anotaciones de dominios estructurales obtenidas mediante la creación de la base de datos *SCOP2GO* (sec. 3.1.1) se compararon con las anotaciones implícitas en *InterPro2GO* (véase Apéndice A.2, Bases de datos). La base de datos *InterPro2GO* asocia términos GO con entradas de *InterPro*. A pesar de que el objetivo de *InterPro2GO* no es localizar físicamente los dominios responsables de llevar a cabo cada FM-GO, se puede obtener una base de datos similar a *SCOP2GO* aprovechando la relación que existe entre *InterPro* y *SCOP*.

Para ello se transfieren las anotaciones desde las entradas de *InterPro* a los correspondientes dominios de *SCOP*. Cuando una entrada de *InterPro* está asociada a más de un dominio, estos se asocian en bloque con el

término GO. El proceso se repite para otros términos GO acumulando las anotaciones de forma similar a como se hizo con *SCOP2GO*. Finalmente se elimina la redundancia en las anotaciones, ya que un plegamiento de *SCOP* puede estar referenciado desde varias entradas de *InterPro* anotadas con el mismo término GO. El resultado final es una lista de plegamientos de *SCOP* anotados con términos GO similar a *SCOP2GO*.

3.1.2.2. Evaluación semiautomática de tres FM-GO relacionadas con la unión de grupos prostéticos

A pesar de que no es posible hacer una evaluación exhaustiva de todas las anotaciones de *SCOP2GO*, se puede realizar una evaluación parcial a gran escala de algunas FM-GO cuya asignación a los dominios puede verificarse utilizando otros recursos independientes. Este es el caso, por ejemplo, de FM-GO relacionadas con la unión a ciertos ligandos.

A continuación se detalla en procedimiento semiautomático utilizado para evaluar las anotaciones funcionales de dominios estructurales para términos GO relacionados con la unión a grupos prostéticos. En este trabajo se evaluaron tres términos GO: *ATP-binding*, *GTP-binding* y *heme-binding*. En cada caso se aplicó el mismo procedimiento tanto para las anotaciones de *SCOP2GO* como para las inferidas a partir de *InterPro2GO* (sec. 3.1.2.1).

La evaluación consiste básicamente en cuantificar el grado de interacción de cada dominio estructural (o conjunto de dominios) con el grupo prostético al que supuestamente se une según el término GO asignado (fig. 3.2). Para la cuantificación también se tuvieron en cuenta ligandos "funcionalmente equivalentes" con el objetivo de obtener más flexibilidad en la evaluación. Por ejemplo, en el caso de *GTP-binding*, además del GTP, se incluyeron ligandos como el GDP o el GMP. Un ligando se considera equivalente si tiene un coeficiente de similitud de Tanimoto (Holliday et al., 2002) superior a 0,95 respecto al grupo prostético original. En el caso de *ATP-binding* se incluyó manualmente el AMP a pesar de tener un coeficiente de 0,87. Los datos de similitud se obtuvieron del servidor

superligands (Michalsky et al., 2005).

La evaluación de cada dominio anotado con alguna de las tres FM-GO se realizó de forma independiente. Para ello primero se extrajeron los dominios de *SCOP* pertenecientes a la misma categoría "dominio proteico". Con esto se obtienen distintos cristales de la misma proteína así como proteínas cristalizadas en distintas condiciones o con diferentes ligandos. Para aquellos dominios cristalizados con el ligando (o alguno de los ligandos funcionalmente equivalentes) se recuperó de *MSD* (Boutselakis, 2003) la lista de residuos de la proteína que contactan con dicho ligando. Finalmente, se calculó el porcentaje de estos residuos pertenecientes al dominio anotado en *SCOP2GO* con el término GO (solapamiento). El solapamiento global del dominio anotado con el término GO viene determinado por el mayor de los solapamientos de los dominios equivalentes. Si ni el dominio ni ninguno de sus variantes une el ligando (o un ligando equivalente) el solapamiento global será próximo a 0. Por tanto la anotación con una determinada FM-GO de unión será tanto mejor cuanto mayor sea el solapamiento.

3.1.2.3. Conjunto de proteínas multidominio

Las proteínas multidominio discutidas por Bashton y Chothia se pueden utilizar como conjunto de prueba adicional para la evaluación de *SCOP2GO*. Estos autores hicieron un estudio comparativo entre la función de 45 proteínas multidominio y las correspondientes proteínas homólogas monodominio. En la mayoría de los casos se observa que los dominios de las proteínas multidominio conservan la función de los correspondientes homólogos.

Para evaluar *SCOP2GO* se revisaron manualmente las predicciones para cada uno de los dominios de este conjunto de proteínas multidominio, y se evaluaron en base a las funciones de las correspondientes proteínas monodominio. Para ello primero se recuperaron los identificadores de *SCOP* tanto para el dominio evaluado como para su homólogo de la proteína monodominio. En todos los casos se verificó la relación de homología

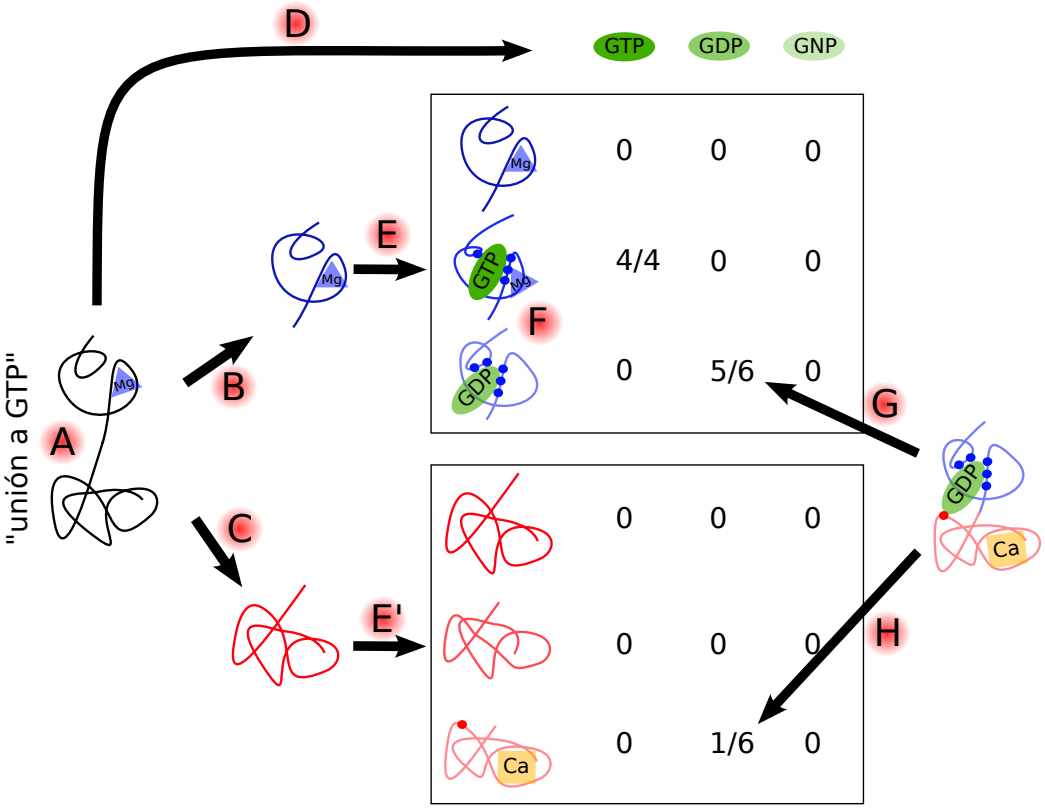


Figura 3.2: ***Procedimiento de evaluación automático de FM-GO de unión a ligandos.*** *Las anotaciones para los términos GO relacionados con la unión a ligandos se pueden evaluar buscando estructuras cristalizadas con dichos ligandos. Una cadena proteica con dos dominios está anotada con un término GO de unión a un ligando (GTP-binding en el ejemplo) (**A**). SCOP2GO transfiere la anotación al dominio responsable. La predicción puede apuntar al dominio correcto (**B**) o al incorrecto (**C**). El primer paso es recuperar la lista de grupos prostéticos que pueden estar representando esa función de unión (**D**). Se recuperan también las entradas de SCOP con la misma categoría "domain" que el dominio predicho para tener todos los cristales disponibles en PDB del dominio evaluado (**E**, **E'**). Si la predicción fue correcta probablemente alguno de los dominios une uno de los grupos prostéticos que representa la función (**F**). Se calcula el porcentaje de residuos que contactan con el grupo prostético según MSD (**G** –puntos de colores–) que están dentro del dominio predicho (puntos azules). Para una predicción incorrecta (**E'**), casi ningún residuo del dominio contacta con esos grupos prostéticos (**H**).*

comprobando que ambos dominios pertenecen a la misma superfamilia de *SCOP* (véase Apéndice A.2, Bases de datos). Por último se comparó la función predicha por *SCOP2GO* para el dominio de la proteína multidominio con la función de la proteína monodominio. En algunos casos no se pudo realizar la evaluación de la predicción al no estar identificado el dominio homólogo o no estar disponible la correspondiente entrada en *SCOP*.

3.1.2.4. Comparación con GOtcha

Para ilustrar el problema de usar sistemas basados en anotaciones de cadenas completas se simuló la predicción funcional de dominios utilizando el servidor *GOtcha* (Martin et al., 2004). Este servidor permite predecir la función de una secuencia problema a partir de las anotaciones funcionales de secuencias similares. En este caso, se hicieron predicciones funcionales independientes para cada dominio del factor de elongación 1α de *Sulfolobus solfataricus* (*PDB*:1jny, cadena A). Para ello se recuperó la secuencia de cada dominio utilizando la base de datos *astral* (Chandonia et al., 2004) y se envió de forma separada al servidor *GOtcha*.

3.1.3. Interfaz de usuario

El servidor web *SCOP2GO*, que permite consultar estas anotaciones FM-GO a nivel de dominio, fue implementado siguiendo un modelo de tres capas (fig. 3.3). La capa de datos se encarga de almacenar y recuperar la información biológica a través del gestor de bases de datos *MySQL*. Todas las propiedades que requieren un tiempo de computación considerable (como el cálculo del valor-P) son precalculadas para disminuir el tiempo de acceso.

La lógica de la aplicación se lleva a cabo mediante un *servlet* principal, que se encarga de controlar el flujo general de la aplicación (fig. 3.3). Este componente es el único que se comunica con la capa de datos para solicitar información y/o modificar la base de datos. También es el encargado de responder a los eventos procedentes de la capa de usuario o interfaz de

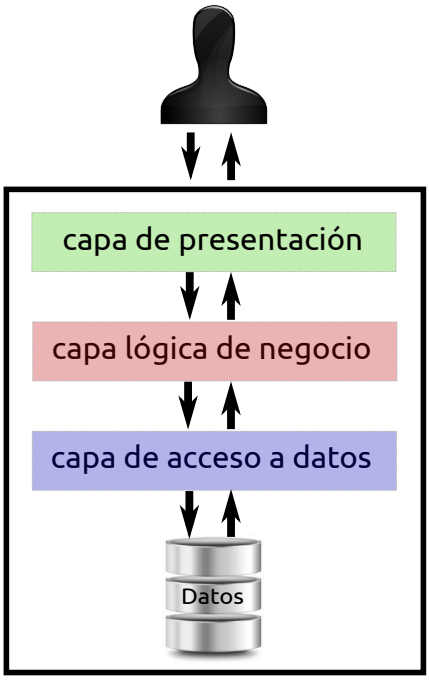


Figura 3.3: ***Modelo de programación de tres capas.*** *El usuario (arriba) interacciona únicamente con la capa de presentación (interfaz gráfica). La capa lógica recibe las peticiones del usuario y envía las respuestas tras el procesado. Esta capa es la encargada de la comunicación con la capa de acceso a datos para solicitar el almacenamiento o la recuperación de datos. La capa de datos accede a los mismos mediante uno o más gestores de base de datos (abajo).*

usuario.

La interfaz de usuario está programada principalmente en JSP. Los JSP reciben información del *Servlet* y en función de ésta generan una vista en HTML. Las peticiones del usuario son redirigidas al *Servlet*. En los casos en los que es necesario cambiar dinámicamente el contenido de la página web (por ejemplo, al expandir la ontología de GO) se utiliza *Javascript*. Para mostrar la estructura tridimensional de un dominio se utiliza la miniaplicación Java Jmol (http://www.jmol.org).

El servidor web incorpora además un formulario para que cualquier usuario sugiera correcciones en la base de datos para una determinada anotación funcional. Estas sugerencias son posteriormente revisadas manualmente.

3.2. Predicción funcional y estructural de dominios proteicos

En esta sección se describe la metodología empleada para la creación de la colección de perfiles funcionales basada en dominios estructurales anotados con términos FM-GO en *SCOP2GO* (PERFILES_GO). Además, se describe el método utilizado para evaluar la capacidad predictiva de estos perfiles para determinar funcional y estructuralmente los dominios de una proteína. Finalmente, se incluye un apartado con las características técnicas del servidor COPRED específicamente desarrollado para facilitar las predicciones en secuencias de aminoácidos utilizando estos perfiles.

3.2.1. Colección de perfiles GO y plegamientos

El punto de partida para generar la colección de perfiles es el recurso *SCOP2GO* descrito anteriormente (sec 3.1). Primero se eliminaron los multidominios (conjunto de dominios asociados juntos a una función) así como los dominios cuyas cadenas *PDB* están marcadas como "mutantes" o "permutaciones circulares". De esta forma, los dominios anotados

funcionalmente en *SCOP2GO* se pueden agrupar como una matriz de plegamientos de *SCOP* (Andreeva et al., 2004) y términos FM-GO (fig. 3.4). Para cada par plegamiento/función se eliminó la redundancia en función de la identidad de secuencia ($\leq$ 40 %) mediante T-coffee (Notredame et al., 2000). Una vez eliminada la redundancia, las entradas en la matriz con menos de tres secuencias fueron descartadas.

El siguiente paso consiste en la generación de un alineamiento estructural múltiple de los dominios restantes con la misma estructura/función (fig. 3.4). La mayoría de los programas de alineamiento estructural múltiple están limitados a un número relativamente pequeño de estructuras, que resultaría insuficiente en este caso. Por esta razón, se usó Dali_lite (Holm y Park, 2000) para generar alineamientos binarios entre cada dominio y un dominio patrón, y se apilaron estos alineamientos binarios para obtener un alineamiento estructural pseudo-múltiple. Como dominio patrón se utilizó aquel con la longitud de secuencia más próxima a la media del conjunto. Este procedimiento se aplicó para todas las entradas en la matriz.

Finalmente, se generó un perfil PSSM de *psi-blast* para cada uno de estos alineamientos estructurales (PERFIL_GO). Por tanto, estos PERFILES_GO representan la distribución de aminoácidos en cada posición de un alineamiento múltiple de los dominios con la misma estructura/función.

3.2.2. Búsqueda de una secuencia contra la colección

El programa *rps-blast* permite comparar una secuencia contra una colección de matrices PSSM. Dado que cada PERFIL_GO está asociado a un plegamiento de *SCOP* y un término FM-GO, la búsqueda de una secuencia contra la colección creada produce una lista de resultados que representan predicciones funcionales y estructurales de regiones (dominios) de esta secuencia (fig. 3.4). Además, debido a la forma de generar los PERFILES_GO (sec. 3.2.1), se espera que sus posiciones conservadas se correspondan con sitios importantes para la FM-GO en ese plegamiento o

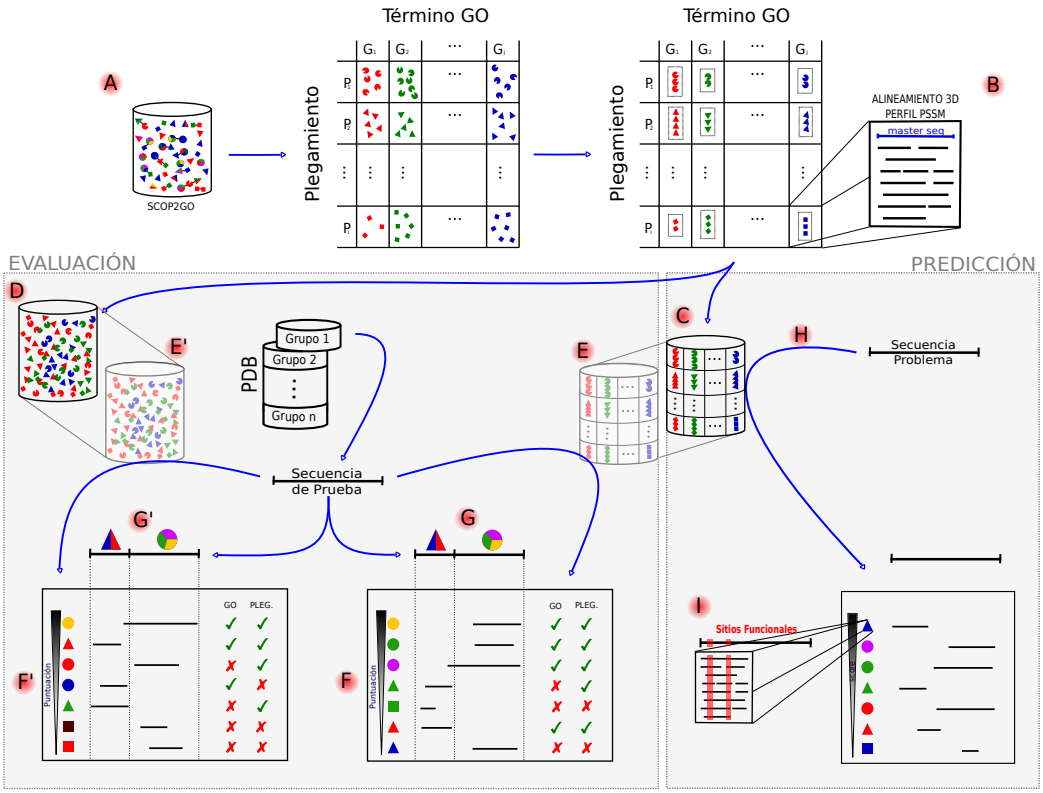


Figura 3.4: ***Esquema del método de predicción y evaluación.*** *El punto de partida es la base de datos de anotaciones de dominios SCOP2GO (**A**). Las formas representan el plegamiento de SCOP de los distintos dominios. Los colores indican el término FM-GO originalmente asociado al dominio en SCOP2GO. A partir de los alineamientos estructurales de dominios con el mismo plegamiento se generan perfiles PSSM (**B**) y se compilan en una base de datos (**C**). Para la evaluación, se genera una base de datos equivalente con los mismos dominios incluidos en los perfiles PSSM (**D**). Para evaluar ambos recursos para una secuencia problema, se reconstruyen las bases de datos excluyendo esta secuencia y todas sus homólogas (**E**, **E'**). Al buscar la secuencia problema en ambos recursos se obtiene una lista de resultados que pueden ser interpretados como predicciones estructurales y funcionales asociadas a sus dominios (**F**, **F'**). Las predicciones generadas con ambos recursos pueden contrastarse con las anotaciones originales en SCOP2GO (**G**, **G'** –triángulo y círculo multi-coloreado–). Para la predicción, se busca una secuencia desconocida (en términos de función y estructura de sus dominios) contra la base de datos de perfiles PSSM (**H**). Los resultados pueden interpretarse como predicciones de función y estructura de sus dominios. Además, el patrón de conservación de los alineamientos estructurales asociados a los perfiles PSSM encontrados puede ayudar a identificar residuos funcionales de la secuencia problema (**I**).*

para el mantenimiento general de la estructura. Por esta razón, inspeccionar el alineamiento entre la secuencia problema y el perfil puede ayudar también a detectar residuos funcionales (fig. 3.4).

3.2.3. Evaluación de la capacidad predictiva

Una de las ventajas de método presentado es que los PERFILES_GO están construidos con dominios con la misma función, en lugar de basarse en grupos de dominios relacionados por su identidad de secuencia (p.e. familias y superfamilias). Para poner de manifiesto la ventaja de usar estos perfiles frente a una predicción estándar basada en transferencia por homología, comparamos las predicciones hechas mediante la colección de PERFILES_GO con aquellas hechas utilizando una base de datos equivalente con exactamente las mismas secuencias pero sin agruparlas según su función (PSI_BLAST) (fig. 3.4).

Para evaluar cada uno de estos recursos se construyó un conjunto de prueba a partir de la base de datos *PDB* agrupada al 30 % de identidad descargada directamente del sitio *RCSB* (Berman et al., 2000) (fig. 3.4). El conjunto de prueba consiste en una secuencia representativa de cada uno de estos grupos. Para ello, se tomó la primera secuencia de cada grupo con algún dominio anotado en *SCOP2GO*. A pesar de que los dos recursos evaluados están basados en dominios, se utilizó la cadena completa para simular un escenario de aplicación real de predicciones con el método presentado aquí.

Para cada cadena del conjunto de prueba se siguió el siguiente procedimiento. Primero, se reconstruyeron los dos recursos evaluados eliminando todos los dominios pertenecientes a alguna de las cadenas *PDB* del mismo grupo que la secuencia de prueba (fig. 3.4). En el caso de PERFILES_GO esto implica además volver a compilar los perfiles PSSM que contienen alguna de estas secuencias similares a la cadena de prueba. Con esto se simula un escenario de predicciones sin homólogos claros en la base de datos. La búsqueda de la secuencia problema contra los dos recursos evaluados da como resultado dos listas ordenadas de predicciones de estructura/función para regiones (dominios) de la secuencia problema

(fig. 3.4). Dado que se conoce la estructura/función de los dominios de la secuencia problema, cada predicción puede ser etiquetada como "acierto" o "fallo" en términos de función y estructura según la anotación original de *SCOP2GO* (fig. 3.4). En todo caso, la región predicha se tiene en cuenta para la evaluación; la predicción sólo es correcta si alinea con un dominio de la secuencia problema con esa estructura/función (fig. 3.4). Para comprobar esto se realizan alineamientos binarios con *BLAST* entre la región predicha de la secuencia problema y cada uno de sus dominios tomados de *ASTRAL* (Chandonia et al., 2004).

3.2.4. Servidor web *COPRED*

El servidor web *COPRED* implementa el método de predicción descrito anteriormente (sec. 3.2.1). Este servidor fue diseñado para facilitar la caracterización estructural y funcional de los dominios de una proteína a partir de su secuencia de aminoácidos. Al igual que *SCOP2GO*, está programado siguiendo un modelo de programación en tres capas (fig. 3.3).

La interfaz de usuario está programada principalmente en JSP y *javascript*. Esta capa se encarga de presentar los datos al usuario en función de la información recibida por el controlador (capa lógica). Para facilitar la implementación de ciertas funcionalidades, como las peticiones AJAX en segundo plano, se hizo uso de la librería JQuery. Además, en la vista exploratoria de los resultados se utiliza la librería Threejs. Esta librería permite visualizar y manipular representaciones 3D de moléculas usando la tecnología *WebGL* (Tavares, 2012). De esta forma, no es necesario instalar accesorios (*plugins*) o mini-aplicaciones Java (*applets*) para obtener una primera vista exploratoria. Por el contrario, el navegador necesita tener activada la compatibilidad con esta tecnología. Para explorar en detalle el alineamiento de la secuencia problema con un perfil determinado se utilizó *Jalview* (Waterhouse et al., 2009). De forma adicional, se utilizó Jmol para mostrar el modelo tridimensional implícito de la región (dominio) de la secuencia problema encajado por el método en un determinado perfil. Este modelo se genera a partir de los carbonos α del miembro representativo del

perfil, renombrando los residuos según corresponda en la secuencia problema.

La capa lógica o de controlador se encarga de gestionar las peticiones del usuario y devolver los resultados (fig.3.3). Al introducir una secuencia se crea un trabajo que es gestionado mediante un sistema de colas multi-hilo configurable según la potencia del servidor. El resultado de cada trabajo se almacena en el servidor durante 30 días para poder recuperarlo mediante un identificador único.

Los datos necesarios para el funcionamiento del servidor, como la base de datos de perfiles, las estructuras *PDB* o los alineamientos usados para compilar cada perfil, se almacenan en un sistema de ficheros.

3.3. Propiedades del mRNA y características locales en las proteínas

Para comprender la relación entre propiedades locales del mRNA no relacionadas con los aminoácidos codificados y regiones funcionales en las correspondientes proteínas se realizó un análisis masivo utilizando datos experimentales de propiedades del mRNA y anotaciones funcionales de *UniProt* (UniProt Consortium, 2010).

3.3.1. Datos experimentales masivos de propiedades del mRNA

Se utilizaron tres propiedades del mRNA relacionadas con la VE obtenidas experimentalmente: ocupación ribosomal y concentración de tRNA para *Escherichia coli*, y estructura secundaria del mRNA para *Saccharomyces cerevisiae*. En cada caso, se obtuvieron estos datos representados como una lista de valores continuos para cada nucleótido para todos los mRNA del organismo.

Perfiles de ocupación ribosomal. Los perfiles de ocupación ribosomal, utilizados como aproximación de la VE, se obtuvieron a partir de los datos generados por Li et al. (Li et al., 2012) para *E. coli* (GEO:GSM872393). La técnica utilizada está basada en la secuenciación masiva de los fragmentos de mRNA protegidos por los ribosomas en una muestra congelada. La alineación de esos fragmentos con el genoma de referencia permite cuantificar de forma indirecta la densidad ribosómica en el mRNA (en unidades arbitrarias) con una resolución de un solo nucleótido. Se asume que las regiones con alta ocupación ribosómica se traducen más lentamente, ya que hay más ribosomas "parados" ahí en el momento del experimento. Debido a que algunas posiciones presentaban valores extremos posiblemente a causa de un error o artefacto en la técnica, se optó por excluir del conjunto de datos aquellos valores alejados de la media más de 1,5 veces la desviación estándar.

Estructura secundaria del mRNA. En el caso de la estructura secundaria del mRNA, se utilizaron los datos de levadura obtenidos por Kertesz et al. (Kertesz et al., 2010) mediante la técnica denominada PARS. Esta técnica permite obtener de forma masiva el contenido de estructura secundaria (grado de apareamiento) de los mRNA. Al igual que los perfiles de ocupación ribosomal, esta técnica permite obtener valores cuantitativos para cada nucleótido del mRNA. Para ello, una muestra de RNA se digiere de forma independiente con dos enzimas: la RNAsa V1 y la nucleasa S1. La RNAsa V1 corta preferentemente el RNA de doble hélice, mientras que la nucleasa S1 actúa principalmente en el de cadena sencilla. El nucleótido exacto sobre el que actúa cada enzima se determina mediante secuenciación masiva y alineamiento de los fragmentos con el genoma de referencia. El grado de apareamiento de cada nucleótido se cuantifica como el logaritmo del cociente del número de veces que actúa cada enzima en ese nucleótido. Debido a que la distribución de valores para cada uno de los cuatro nucleótidos difiere considerablemente, los datos se normalizaron transformándolos a valores-Z. Para un nucleótido, el valor-Z viene dado

por:

$$Z_n = \frac{x - \mu_n}{\sigma_n} \tag{3.2}$$

donde n representa el valor PARS del nucleótido, y μ_n y σ_n la media y la desviación típica de los valores PARS para ese tipo de nucleótido.

Concentración de tRNA. Se utilizaron los datos de concentración citoplasmática de los distintos aa-tRNA en *E. coli* obtenidos experimentalmente por Dong et al. (Dong et al., 1996). Estos autores identificaron, separaron y cuantificaron los niveles de concentración de cada aa-tRNA en un cultivo celular de *E. coli* en fase de crecimiento. Debido a que la metodología empleada para el estudio de la relación entre una propiedad del mRNA y las características locales de las proteínas requiere valores cuantitativos asociados a nucleótidos, se asignó el valor de concentración de tRNA a cada uno de los tres nucleótidos del codón.

3.3.2. Descripción general del procedimiento de evaluación

La figura 3.5 sintetiza el proceso seguido para evaluar la relación entre una propiedad del mRNA cuantificable a nivel de nucleótido y las características estructurales y funcionales locales de las correspondientes proteínas. Para llevar a cabo este análisis se combinaron las anotaciones proporcionadas por *Uniprot* con los datos experimentales de los correspondientes mRNA descritos anteriormente (sec. 3.3.1). Dado que las anotaciones de *UniProt* no incluyen regiones desordenadas, se añadieron anotaciones predichas con *IUPred* (Dosztányi et al., 2005) y *ANCHOR* (Mészáros et al., 2009). En ambos casos se utilizaron los parámetros por defecto de estos programas.

Para una determinada anotación proteica, el proceso empieza alineando por el primer residuo todas las instancias de esa anotación (fig. 3.5). Esto es necesario porque las instancias del mismo tipo de anotación pueden diferir

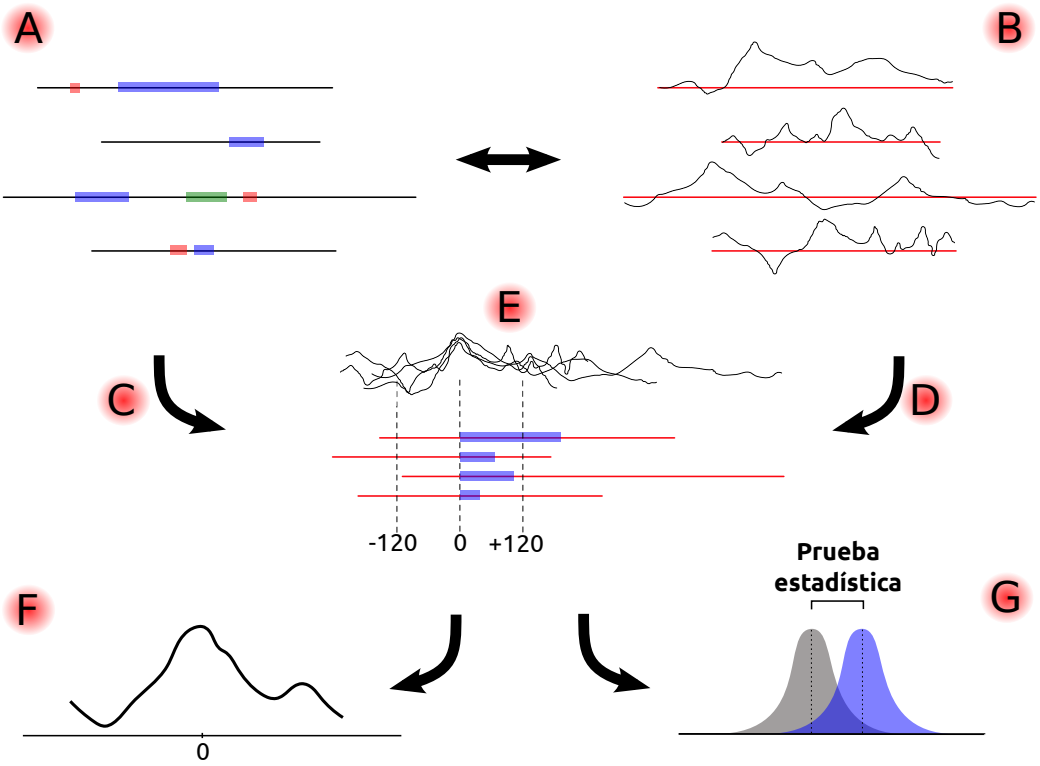


Figura 3.5: ***Esquema de la metodología usada para estudiar la relación entre las propiedades locales del mRNA y las características funcionales/estructurales de las proteínas que codifican.*** *(**A**) Conjunto de proteínas de un organismo (líneas negras) con distintas anotaciones regionales marcadas (cajas de colores). (**B**) Conjunto de mRNA de ese organismo (líneas rojas) con los correspondientes valores cuantitativos para cada nucleótido según la propiedad del mRNA analizada (estructura secundaria, ocupación ribosomal o concentración de tRNA -líneas negras-). (**C**) Para una determinada anotación (caja violeta), todas las instancias en el proteoma son alineadas respecto al primer residuo. Esto determina un alineamiento equivalente para las regiones de los correspondientes mRNA (**D**) y su propiedad asociada (**E**). El patrón promedio representa el comportamiento de la propiedad del mRNA analizada en las proximidades de la anotación proteica (**F**). Para calcular la significación de la relación entre los patrones característicos (**E**) y la propiedad funcional de la proteína, se compara la distribución de correlaciones entre cada par de ventanas de valores (en azul) con una distribución equivalente generada a partir de ventanas de valores del mismo tamaño obtenidas aleatoriamente del proteoma (en gris).*

en su longitud (p.e. sitios de unión de distinto tamaño). A su vez, este alineamiento determina el alineamiento de los correspondientes mRNA que codifican estas proteínas anotadas en *UniProt*. De esta manera, se puede calcular un patrón característico de la anotación promediando los valores de la propiedad del mRNA para cada uno de sus nucleótidos (fig. 3.5).

Para determinar la significación de estos patrones se utilizaron ventanas de valores de la propiedad del mRNA para los 120 nucleótidos en torno al inicio de la anotación. Primero se calculó la distribución de las correlaciones para todos los pares de estas ventanas de valores. A continuación, generó una distribución equivalente a partir de 1000 ventanas de valores del mismo tamaño tomadas al azar del conjunto de mRNA (fig. 3.5). Finalmente, se compararon ambas distribuciones mediante una prueba t de Student para muestras desapareadas (véase Apéndice A.3, Pruebas estadísticas y normalizaciones).

Capítulo 4

Resultados y discusión

4.1. Anotación funcional de dominios de proteínas

Los dominios son las unidades estructurales, evolutivas y funcionales de las proteínas. En la mayoría de los casos llevan a cabo funciones moleculares independientes hasta el punto de que incluso aparecen aislados como proteínas independientes en otros organismos (Marcotte, 1999). A pesar de ello, la mayoría de los recursos describen funcionalmente las proteínas de manera global, sin tener en cuenta el papel que desempeña cada dominio. Esto ocurre incluso en bases de datos orientadas a dominios cuyas anotaciones funcionales derivan de recursos basados en la anotación funcional de cadenas completas (Lopez y Pazos, 2009).

Por ejemplo, el EF 1α está compuesto por tres dominios estructurales (fig. 4.1). El dominio N-terminal (*SCOP*:d1jnyA3, plegamiento c.37) cambia su conformación espacial al unir e hidrolizar GTP. El dominio 2 (*SCOP*:d1jnyA1) adopta una estructura en barril-beta y está involucrado en la unión de aa-tRNA. El dominio 3 (C-terminal) (*SCOP*:d1jnyA2) también adopta una estructura en barril-beta, aunque está involucrado tanto en la unión de aa-tRNA como en la del EF-1β (fig. 4.1).

Esta proteína (la cadena completa) está anotada en *GOA-PDB* con tres

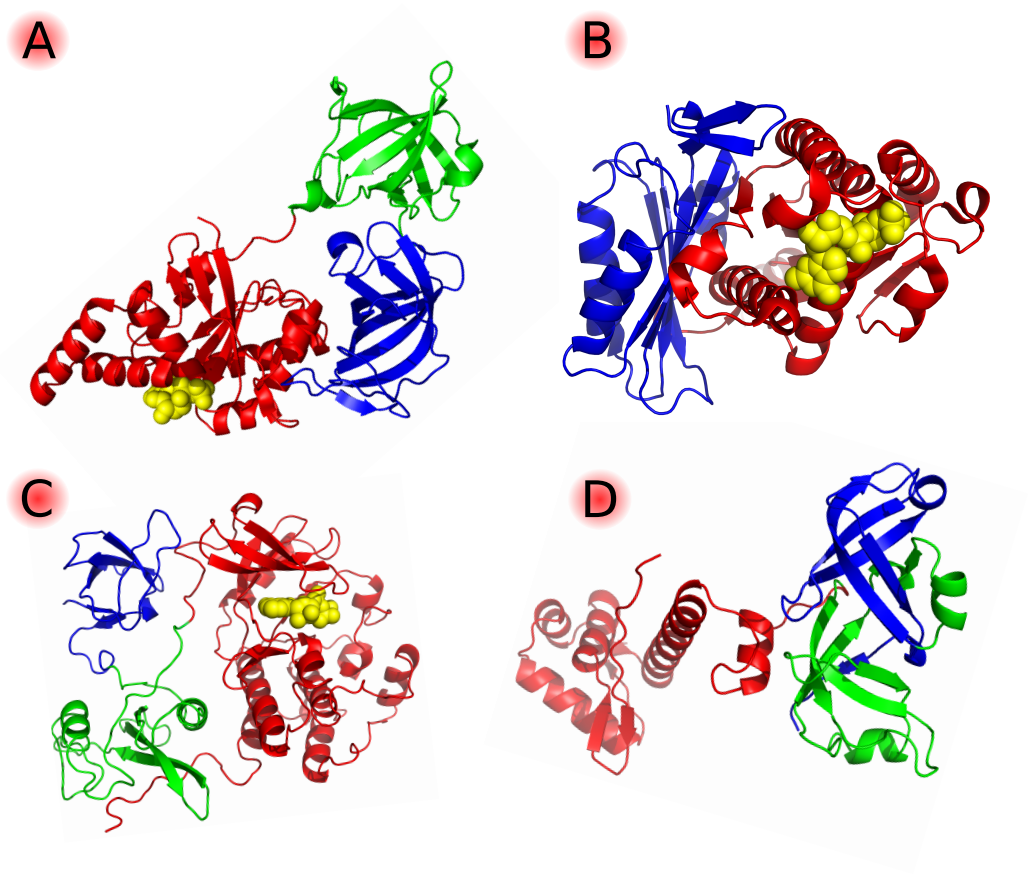


Figura 4.1: ***Ejemplos de proteínas con dominios con distinta función molecular***. *(**A**) Factor de elongación 1α (PDB:1jnyA). La figura muestra la estructura 3D de esta proteína con los dominios marcados según la descripción de SCOP. El GDP está coloreado en amarillo. (**B**) FtsZ (PDB:1w5bA). Se muestran los dos dominios estructurales de SCOP íntimamente ligados y el GTP coloreado en amarillo. (**C**) Tirosina fosfatasa (PDB:2srcA). La figura muestra la estructura 3D de esta proteína con el ANP (análogo de ATP) en amarillo. (**D**) Regulador transcripcional dependiente de molibdeno (PDB:1b9mA). Los dos dominios Mop de unión a molibdeno se muestran en verde y azul, mientras que el de unión a DNA está en rojo. Figuras generadas con Pymol (Schrödinger, 2010).*

Tabla 4.1: ***Comparación de las anotaciones de Pfam/InterPro y SCOP2GO para los dominios de EF-1α.*** *Cada fila indica la anotación funcional según la base de datos. SCOP2GO asocia la "Actividad EF" con dos dominios (d1jnyA3-d1jnyA1), indicando que ambos son necesarios para llevar a cabo dicha función.*

Dominio	Pfam/InterPro	SCOP2GO		
d1jnyA3	Unión a GTP	Unión a GTP	-	Activ. EF
d1jnyA1	Unión a GTP	-	Unión a *RNA*	
d1jnyA2	Unión a GTP	-	-	-

términos FM-GO: "unión a GTP", "actividad GTPasa" y "actividad EF". En *Pfam*, cada dominio está automáticamente anotado con el término "unión a GTP" a pesar de que la información textual disponible para cada entrada (PF00009, PF03143, y PF03144) diferencia claramente sus funciones (tabla 4.1). Esto es debido a que las anotaciones electrónicas de *Pfam* son transferidas directamente desde los correspondientes dominios de *InterPro*, que se anotan automáticamente mediante *InterPro2GO* (véase Apéndice A.2, Bases de datos).

El objetivo de *InterPro2GO* no es sin embargo anotar regiones de proteínas. Este recurso está diseñado para facilitar la anotación de nuevas secuencias (cadenas completas) a partir de coincidencias locales con dominios de *InterPro*. En este sentido, la asociación de la función "unión a GTP" con el segundo y tercer dominio sería apropiada, ya que una secuencia con estos dominios probablemente se trate de un EF y por tanto interaccione con el GTP aunque no sea el dominio alineado.

Este problema en las anotaciones tiene consecuencias también en los sistemas de predicción que las usan. La interpretación de los resultados de servidores basados en anotaciones de cadenas completas en términos de dominios físicos podría llevar a conclusiones erróneas. Si por ejemplo se utiliza el servidor *GOtcha* para predecir la función de cada dominio del EF-1α de forma independiente (véase Materiales y Métodos, sec. 3.1.2.4), éste asigna el término GO "unión a GTP" a cada dominio con

Tabla 4.2: ***Predicción funcional utilizando el servidor GOtcha.*** *La tabla muestra la puntuación asignada a la asociación de cada dominio del EF-1α con el término GO:0005525 ("unión a GTP") cuando se predice de forma independiente la función de cada dominio.*

	d1jnyA1	d1jnyA2	d1jnyA3
Puntuación	0,92	0,90	0,92
Desv. típica	0,27	0,27	0,26
Probabilidad est.	0,54	0,45	0,52

prácticamente la misma puntuación (tabla 4.2). Estos resultados muestran que sistemas entrenados a partir de anotaciones funcionales de cadenas completas no pueden ser usados para predecir funciones de dominios, y resalta la importancia de desarrollar y utilizar recursos como *SCOP2GO*.

4.1.1. Ejemplo de anotación funcional de dominios

SCOP2GO (Lopez y Pazos, 2009) fue la primera base de datos de anotaciones funcionales de dominios de proteínas. A pesar de que se empezaba a reconocer la importancia de los recursos orientados a dominios (Riley, 2007), cuando desarrollamos *SCOP2GO* no existía ninguna otra base de datos con anotaciones funcionales explícitas de dominios usando un vocabulario funcional.

La base de datos se genera automáticamente mediante la transferencia de términos FM-GO originalmente asociados a cadenas completas de *PDB* a sus dominios estructurales (véase Materiales y Métodos, sec. 3.1.1).

Por ejemplo, en el caso del EF-1α discutido anteriormente (fig. 4.1), *SCOP2GO* identifica correctamente el dominio N-terminal (*SCOP*:d1jnya3, c.37) como el único involucrado en la unión de GTP (tabla 4.1). Esto se debe a que *SCOP2GO* utiliza como "conjunto de entrenamiento" todas las proteínas anotadas con la función "unión a GTP" en lugar de restringirse a secuencias similares al EF-1α anotadas con dicha función. El plegamiento c.37 es el más abundante en proteínas que unen GTP. De hecho, en ocasiones es el único dominio de la cadena en proteínas con esta FM-GO.

Dado que el método asume que el plegamiento más frecuente es con mayor probabilidad el responsable de llevar a cabo la función, *SCOP2GO* asocia la función únicamente al primer dominio (plegamiento c.37), descartando cualquier implicación de los otros dos.

SCOP2GO también asigna la función "actividad EF" al conjunto del primer y segundo dominio (*SCOP2GO*:d1jnya3-d1jnya1) (tabla 4.1). Estos dos dominios son frecuentes en otros EF en los que a veces no está presente el dominio C-terminal. Por ello el método interpreta que la función "actividad EF" puede llevarse a cabo sin este último dominio. Además, dado que los dos dominios suelen aparecer juntos en cadenas anotadas con esa función (tienen una distribución similar), el método interpreta que ambos son necesarios para llevar a cabo la actividad EF.

Finalmente, a pesar de que tanto el dominio 2 como el 3 del EF-1α interaccionan con el RNA, *SCOP2GO* asocia la función "unión a RNA" únicamente con el dominio intermedio (tabla 4.1). Esta falta de anotaciones funcionales para algunos dominios es aún más evidente en otros casos, como en el EF-Tu (homólogo de EF-1α). Mientras que *SCOP2GO* identifica correctamente el dominio N-terminal como el responsable de la unión al GTP, ningún dominio aparece asociado con la actividad EF. Esto se debe a que el EF-Tu no está originalmente anotado en *GOA-PDB* con esta función. Estos dos últimos ejemplos ilustran la necesidad de interpretar las anotaciones de *SCOP2GO* sólo en términos de anotaciones positivas. El método permite asociar uno o varios dominios a un término GO con un cierto nivel de confianza (Valor-P). Sin embargo, el hecho de que un dominio no presente una determinada anotación no implica necesariamente que no la tenga. Esto pone de manifiesto además una importante característica del método: *SCOP2GO* transfiere anotaciones funcionales desde la cadena completa al dominio responsable pero no puede "predecir" nuevas funciones.

4.1.2. Evaluación de *SCOP2GO*

Evaluar la capacidad global de *SCOP2GO* para transferir correctamente las anotaciones a los dominios no es trivial, ya que no existe ningún método similar ni un conjunto de prueba estándar con el que comparar. De hecho uno de los objetivos de *SCOP2GO* es precisamente generar un recurso así. No obstante, la evaluación parcial utilizando diferentes aproximaciones y recursos alternativos podría dar indicios del comportamiento general del método.

4.1.2.1. Conjunto de proteínas multidominio

Las 45 proteínas multidominio estudiadas en detalle por Bashton y Chothia (véase Materiales y Métodos, sec. 3.1.2.3) constituyen un excelente conjunto de prueba para evaluar parcialmente *SCOP2GO*. Dado que los autores discuten tanto la función de estas proteínas multidominio como la de sus correspondientes proteínas homólogas monodominio, las predicciones de *SCOP2GO* para las primeras se pueden evaluar manualmente basándose en las segundas. Por ejemplo, el regulador transcripcional dependiente de molibdeno modE (PDB: 1b9m, fig. 4.1) actúa como represor del operón modABC. Al interaccionar con molibdato esta proteína cambia su conformación espacial, lo que le permite reconocer una región del DNA y modificar los niveles de expresión del operón. La proteína modE tiene tres dominios: el dominio I y dos dominios Mop. El dominio I forma un motivo hélice-giro-hélice capaz de unir DNA. Este dominio es homólogo al represor del operón marRAB (marR, PDB:1jgs), que está implicado en la activación de genes de resistencia a antibióticos y de estrés oxidativo (tabla 4.3). Los otros dos son dominios Mop de unión a molibdato/tunstato, homólogos a la proteína II de unión a molibdeno-pterina. *SCOP2GO* asigna el término GO GO:0003700 (”actividad factor de transcripción”) al primer dominio y GO:0030151 (”unión a ion molibdeno”) a cada uno de los restantes, en perfecta concordancia con las funciones de las proteínas homólogas a cada dominio (tabla 4.3). Al igual que en este ejemplo, en la mayoría de los casos de este conjunto *SCOP2GO* asigna una función a los dominios de proteínas

multidominio que coincide o puede explicarse por la función de la proteína homóloga monodominio (véase Apéndice B, Conjunto de proteínas multidominio). No obstante, éste es un conjunto relativamente pequeño, y no puede usarse como medida global de la eficiencia del método.

Tabla 4.3: ***Predicción de SCOP2GO para el regulador transcripcional dependiente de molibdeno ModE****. Cada dominio de la proteína ModE tiene una proteína homóloga monodominio con función similar a la asignada por SCOP2GO.*

Prot. Multidom.(SCOP)	Func. SCOP2GO	Prot. Monodom.
ModE (d1b9ma1)	factor de transcripción	marR
ModE (d1b9ma3)	unión a ion molibdeno	Proteína II
ModE (d1b9ma4)	unión a ion molibdeno	Proteína II

4.1.2.2. Evaluación a gran escala

Las predicciones de *SCOP2GO* que involucran algunos términos FM-GO pueden ser indirectamente evaluadas de forma automática. Por ejemplo, las FM-GO relacionadas con la unión a grupos prostéticos se pueden evaluar mediante la cuantificación del grado de unión del correspondiente ligando con el dominio predicho (solapamiento). De esta forma, las asociaciones "correctas" tenderán a presentar altos valores de solapamiento, ya que la mayoría de los residuos que interaccionan con el lingando pertenecen al dominio predicho. En cambio, las asociaciones incorrectas tendrán un solapamiento cercano a 0.

La tabla 4.4 muestra el resultado obtenido para la evaluación de las FM-GO de unión a los grupos prostéticos ATP, GTP y *hemo*. En la mayoría de los casos, los dominios anotados con esos términos GO presentan un solapamiento con el ligando correspondiente superior al 70 % (tabla 4.4). Los resultados muestran que la base de datos *SCOP2GO* predice correctamente el dominio estructural responsable de la unión a los grupos prostéticos evaluados en la mayoría de los casos. Además, los altos valores del AUC indican que el valor-P asociado a las predicciones de *SCOP2GO* puede ser usado como medida del grado de confianza de una predicción, ya

que los valores bajos tienen a estar asociados a predicciones correctas y viceversa (véase Apéndice A.3, Pruebas estadísticas y normalizaciones). Pese a ser una evaluación parcial, en principio no hay ninguna razón para pensar que la sensibilidad para predecir correctamente otros términos GO no asociados a unión vaya a ser considerablemente distinta.

Tabla 4.4: ***Evaluación de la predicción de tres FM-GO relacionadas con la unión a grupos prostéticos****. Para cada termino GO, las distintas columnas muestran de izquierda a derecha: el número de dominios estructurales predichos con esa función; el porcentaje de ellos que une el ligando (al menos el 70 % de solapamiento); el porcentaje de dominios correctamente predichos (revisado manualmente); y el AUC como medida del poder discriminativo del valor-P para separar predicciones correctas de las incorrectas.*

		Predicciones correctas		
	Dominios	**Solap.**$\geq 70\,\%$	**Asig. manual**	**AUC**
Unión a GTP	524	92,75 %	95,23 %	0,991
Unión a *HEMO*	1945	94,14 %	96,40 %	0,934
Unión a ATP	2962	68,47 %	82,44 %	0,998

A pesar de los buenos resultados obtenidos siguiendo el método de evaluación automática, especialmente para GTP y *hemo*, en ocasiones *SCOP2GO* no es capaz de predecir correctamente el dominio que une el grupo prostético. Algunas de estas predicciones incorrectas son debidas a errores en la anotación original de *GOA-PDB*. Por ejemplo, la cochaperona GroES (chaperonina 10KDa) está incorrectamente anotada en *GOA-PDB* con la función "unión a ATP". Esto posiblemente se debe a la estrecha vinculación con la chaperona GroEL (chaperonina 60KDa), la cual une e hidroliza ATP durante su ciclo funcional. Este es un claro ejemplo de anotación funcional basada en "complejo proteico", donde cada miembro es anotado con las FM-GO de todo el conjunto. De los 2962 dominios evaluados para ATP, 188 pertenecen a proteínas de la familia GroES donde probablemente está ocurriendo lo mismo. Estos casos reducen claramente la eficiencia para el ATP (68,47 %) y explican en parte la diferencia con el GTP (92,75 %) y el grupo *hemo* (94,14 %) (tabla 4.4).

Algunas predicciones son consideradas "incorrectas" por el método automático de evaluación aun cuando *SCOP2GO* identifica correctamente el dominio de la cadena responsable de la unión al ligando. Este método de evaluación se basa en el cálculo del solapamiento del ligando con el dominio asociado a la FM-GO. Sin embargo, para poder evaluar el solapamiento es necesario que el dominio, o alguno de sus homólogos, haya sido cristalizado con el ligando o los ligandos estructuralmente similares. En caso contrario, el solapamiento es 0 y por tanto la predicción se considera incorrecta (véase Materiales y Métodos, sec. 3.1.2.2). Por ejemplo, algunos de los dominios anotados por *SCOP2GO* con la FM-GO "unión a *hemo*" sólo han sido cristalizados con los ligandos *sirohemo*, *hemo-C* o *hemo-D*. A pesar de ser claramente equivalentes al grupo *hemo*, estos ligandos fueron descartados en la evaluación automática por presentar un coeficiente de Tanimoto inferior a 0,95 con el *hemo*, por lo que el solapamiento calculado en estos casos era 0. Estos y otros falsos negativos derivados del propio método de evaluación se reducirían si existiesen más proteínas cristalizadas con el ligando de interés. De hecho, la evaluación automática utilizando versiones más recientes de *SCOP* produce mejores resultados. Cuando éstos y otros casos similares son corregidos manualmente el porcentaje de aciertos aumenta moderadamente (tabla 4.4).

4.1.2.3. Comparación con las anotaciones de InterPro2GO

Una de las principales características de *SCOP2GO* es la transferencia de anotaciones funcionales basada en la similitud estructural entre dominios (plegamientos) en lugar de restringirse a secuencias homólogas. Los programas de predicción funcional basados en transferencia de anotaciones desde secuencias homólogas están limitados por la capacidad de detectar dichos homólogos. Por debajo de un cierto umbral de similitud de secuencia, la detección de homólogos resulta prácticamente imposible. Sin embargo, la estructura está mucho más conservada que la secuencia (Levitt y Gerstein, 1998). La utilización de información estructural permite detectar homología entre proteínas incluso en ausencia significativa de similitud de secuencia.

Debido a esta característica, el método empleado para generar *SCOP2GO* puede utilizar más secuencias y secuencias más variadas para identificar los dominios responsables de llevar a cabo una determinada función.

Para poner de manifiesto la ventaja de usar este tipo de información estructural, las anotaciones de *SCOP2GO* se compararon con las anotaciones de regiones de proteínas (dominios de *InterPro*) obtenidas a partir de *InterPro2GO* (véase Materiales y Métodos, sec. 3.1.2.1). Para ello se evaluaron las mismas tres FM-GO de unión a grupos prostéticos siguiendo un procedimiento similar al método de evaluación automático empleado en *SCOP2GO*. Es importante remarcar que esta base de datos de anotaciones de dominios derivada de *InterPro2GO* se generó únicamente para comparar los resultados de la evaluación con los obtenidos para *SCOP2GO*. Como se discutió anteriormente, *InterPro2GO* no ha sido diseñado para asociar términos GO a regiones concretas de proteínas, sino para localizar motivos característicos de proteínas con una determinada función. No obstante constituye un buen conjunto de referencia y puede servir para ilustrar los problemas y las limitaciones de interpretar los resultados de los métodos actuales basados en similitud de secuencia a nivel de dominio.

La figura 4.2 muestra el porcentaje de aciertos tanto para *SCOP2GO* como para la base de datos de anotaciones de dominios equivalente generada a partir de *InterPro2GO*. En ambos casos se consideraron predicciones correctas aquellas con un solapamiento $\geq 70\,\%$. Los resultados muestran que *SCOP2GO* supera el porcentaje de aciertos para GTP y ATP, mientras que los aciertos para el grupo *hemo* son prácticamente iguales en ambos casos (fig. 4.2).

La ausencia de diferencia para la predicción de unión al grupo *hemo* probablemente se debe a que muchas de las proteínas que unen este ligando tienen un único dominio (87 %) en comparación con el número de proteínas monodominio que unen GTP (31 %) o ATP (34 %). De hecho, las proteínas con un solo dominio no necesitan ningún método de predicción funcional de dominios, ya que cualquier predicción que involucre a la cadena completa forzosamente deberá incluir el sitio de unión al ligando. Esto da cierta

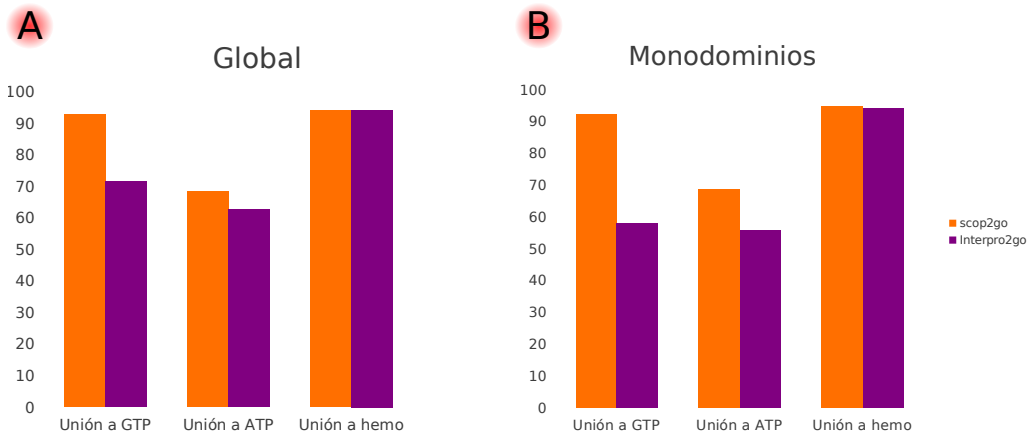


Figura 4.2: ***Comparación de aciertos entre SCOP2GO e InterPro2GO para tres FM-GO diferentes****. Porcentaje de dominios predichos con la función que realmente unen el grupo prostético (solapamiento* $\geq$ *70 %) para distintos términos GO. (**A**) Todas las predicciones. (**B**) Sólo las predicciones que incluyen un único dominio.*

ventaja a las predicciones inespecíficas que asocian grandes regiones de la proteína con una determinada función. Por ejemplo, la proteína de división celular FtsZ tiene dos dominios pero sólo el N-terminal (plegamiento c.32) está involucrado en la unión a GTP (fig. 4.1). *SCOP2GO* asocia la función "unión a GTP" a ambos dominios (SCOP: d1w5a1-dw5a2; valor-P 52,8E-59), ya que no es capaz de discernir el dominio que realmente une el GTP. El motivo es que estos dominios están estrechamente asociados y aparecen siempre juntos, o al menos en las cadenas anotadas con esta función en *GOA-PDB*. Aunque estrictamente hablando debería considerarse un falso positivo, la predicción es considerada correcta durante el proceso de evaluación automática, ya que el sitio de unión del GTP se encuentra dentro del intervalo de la cadena predicha con esa función. *InterPro2GO* también asocia ambos dominios de FtsZ con la función "unión a GTP". Sin embargo, mientras que en *SCOP2GO* las predicciones que involucran a varios dominios son relativamente raras, en *InterPro2GO* pueden suponer hasta el 30 % del total de predicciones (tabla 4.5). Cuando se evalúan únicamente predicciones funcionales para dominios aislados el porcentaje de

aciertos de *InterPro2GO* disminuye considerablemente, mientras que el de *SCOP2GO* se mantiene prácticamente igual (fig. 4.2).

Tabla 4.5: ***Predicciones que involucran más de un dominio.***

	SCOP2GO	**InterPro2GO**
Unión a GTP	7,61 %	34,74 %
Unión a *hemo*	1,28 %	16,29 %
Unión a ATP	1,69 %	1,18 %

4.1.3. Servidor web *SCOP2GO*

El servidor web *SCOP2GO* contiene 70658 anotaciones funcionales de dominios o grupos de dominios pertenecientes a 37923 cadenas *PDB* (fig. 4.3). Estos dominios han sido anotados usando 1297 términos GO específicos. El formulario de búsqueda permite obtener las anotaciones de dominios a partir de identificadores *PDB*, dominios de *SCOP* o términos GO. Además, el servidor incluye otras funcionalidades, como la posibilidad de visualizar la jerarquía de GO para cada anotación, explorar las anotaciones funcionales originales de *GOA-PDB* o visualizar la estructura tridimensional de la cadena en 3D mediante la mini-aplicación Jmol. La aplicación web está preparada además para que cualquier usuario sugiera correcciones en la base de datos en vistas a crear en el futuro una versión revisada manualmente.

La base de datos *SCOP2GO* se puede acceder desde la dirección: http://csbg.cnb.csic.es/scop2go/

4.2. Predicción funcional de dominios de proteínas

Existen múltiples métodos de predicción funcional de proteínas. La mayoría de ellos, especialmente aquellos basados en similitud de secuencia,

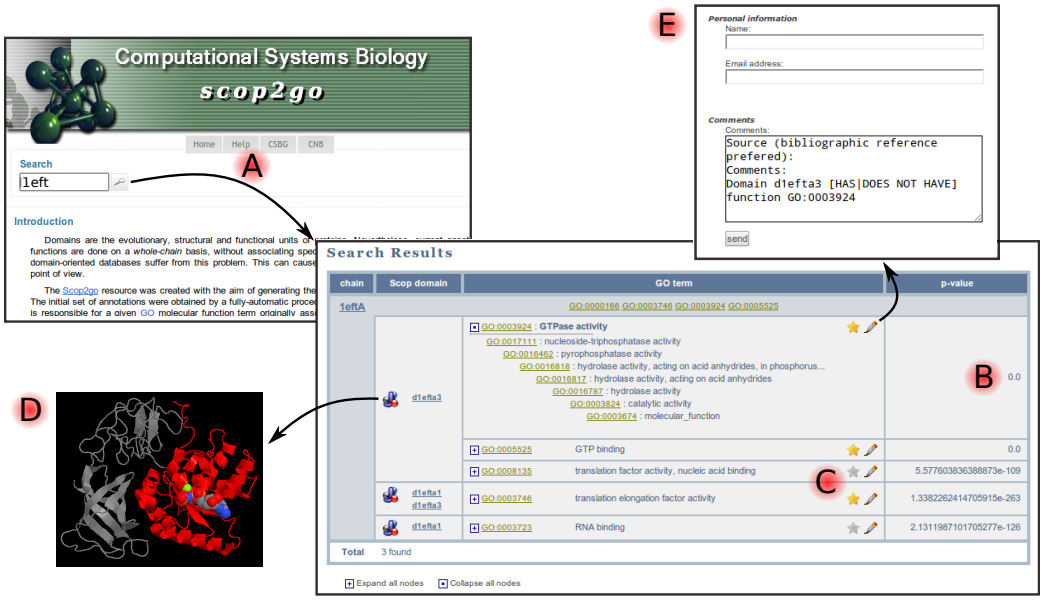


Figura 4.3: ***Capturas de pantalla del servidor web SCOP2GO (http://csbg.cnb.csic.es/scop2go/).*** *(**A**) La búsqueda del PDB 1eft devuelve las anotaciones funcionales de cada uno de sus dominios. (**B**) Cada predicción está asociada a un nivel de confianza (valor-P). (**C**) Las anotaciones funcionales revisadas manualmente aparecen marcadas con una estrella amarilla. (**D**) La estructura 3D del PDB se puede explorar visualmente mediante la mini-aplicación java Jmol (dominio predicho en rojo). (**E**) Los usuarios pueden aportar sugerencias para la revisión manual de la base de datos.*

están pensados para asignar funciones a la cadena completa, sin distinguir el dominio responsable de llevar a cabo cada FM-GO concreta. En la mayoría de los casos no es un problema de la metodología usada sino una consecuencia de los recursos en los que se basan. En estos recursos, las funciones están asociadas a cadenas completas y por tanto son éstas las que se transfieren a la secuencia problema. Como se discutió en la sec. 4.1, dada la relativa independencia funcional de los dominios, resulta imprescindible desarrollar recursos orientados a dominios y consecuentemente métodos predictivos a este nivel.

Para cubrir esta necesidad, desarrollamos un método de predicción funcional de dominios de proteínas basado en *SCOP2GO* (véase Materiales y Métodos, sec. 4.1). El método utiliza una colección de perfiles PSSM (PERFILES_GO) derivada de alineamientos estructurales de dominios anotados con el mismo término FM-GO en *SCOP2GO* (véase Materiales y Métodos, sec. 3.2). Dado que todos los dominios dentro de un perfil comparten el mismo plegamiento, el método implícitamente asigna además un plegamiento de *SCOP* a los dominios de la secuencia problema. Adicionalmente, los patrones de conservación de posiciones en estos perfiles podrían ayudar a identificar posibles sitios funcionales en la secuencia objeto de la predicción.

En las siguientes secciones se describen y discuten los resultados de la evaluación del método y así como las funcionalidades del servidor desarrollado específicamente para realizar las predicciones.

4.2.1. Evaluación del método de predicción

Para evaluar la ventaja de usar PERFILES_GO frente a una predicción estándar basada en transferencia por homología se realizaron predicciones funcionales y estructurales sobre un conjunto de prueba de 1009 cadenas completas de *PDB* de función y estructura conocida (véase Materiales y Métodos, sec. 3.2.2). En cada caso, se eliminaron de los perfiles las cadenas similares a la evaluada para simular un escenario sin homólogos claros en base de datos. De forma similar, se realizaron predicciones utilizando una base

de datos con las mismas secuencias (dominios anotados funcionalmente) sin estar agrupadas por términos GO en perfiles precompilados (PSI_BLAST). Para evaluar la capacidad predictiva de ambas aproximaciones se comparó en cada caso la mejor predicción (menor valor-E) con la anotación original *SCOP2GO*. El proceso completo se sintetiza en la figura 3.4.

Este método de evaluación automática permite comparar las predicciones hechas con PERFILES_GO frente a las realizadas utilizando un método estándar basado en homología. No obstante, los métodos basados en homología normalmente utilizan recursos basados en anotaciones de la cadena completa. De hecho, utilizando este método de evaluación sólo se pone de manifiesto la ventaja de usar perfiles estructurales frente a usar el mismo conjunto de secuencias sin agrupar *a priori* en perfiles. Valorar realmente la diferencia entre usar un recurso basado en anotaciones funcionales de cadenas en lugar de PERFILES_GO, que utiliza *SCOP2GO* como fuente de anotaciones, requeriría una evaluación mucho más exhaustiva que se aleja de los objetivos de esta tesis doctoral.

4.2.1.1. Ejemplos

En este apartado se describen algunos ejemplos de predicciones de PERFILES_GO y PSI_BLAST que ilustran las ventajas y limitaciones de la predicción funcional de dominios basada en PERFILES_GO así como la complementariedad con otros métodos existentes.

El primer ejemplo es la subunidad β de la ATP-sintasa mitocondrial (PDB:1w0kD). La mejor puntuación de PERFILES_GO es para el perfil GO:0005524/c.37 contra la región central de la proteína. Este perfil corresponde al término GO "unión a ATP" en el plegamiento *SCOP* "bucle-P que contiene hidrolasas de nucleótidos trifosfatos". En este ejemplo, el perfil encontrado alinea prácticamente con todo el dominio c.37 de unión a ATP de la proteína evaluada. Por lo tanto, el método automático de evaluación considera que tanto la predicción funcional como la estructural son correctas para esta proteína. En cambio, la búsqueda

equivalente con PSI_BLAST encuentra como mejor candidato una secuencia con función "actividad dihidropteroato sintasa" (GO:0004156), una actividad enzimática que no es dependiente de ATP, y el plegamiento "barril-α-β TIM" (c.1).

Otro ejemplo es el de la hemoproteína citocromo c-L (PDB:2mtaC). Mientras que PERFILES_GO predice correctamente la unión al grupo *hemo* (GO:0020037) en el plegamiento "Citocromo C" (a.3), el mejor resultado de PSI_BLAST es una proteína de unión a DNA (GO:0003677) con el plegamiento d.218. Además, en este caso las dos posiciones más conservadas del perfil GO:0020037/a.3 alinean con los dos residuos (Cis^{79} y Cis^{82}) implicados en la unión covalente del grupo *hemo* del citocromo c-L (fig. 4.4). Este ejemplo ilustra una ventaja adicional del método de predicción funcional de dominios basado en PERFILES_GO: este método puede dar indicios sobre posibles sitios funcionales al tiempo que proporciona predicciones de función y estructura.

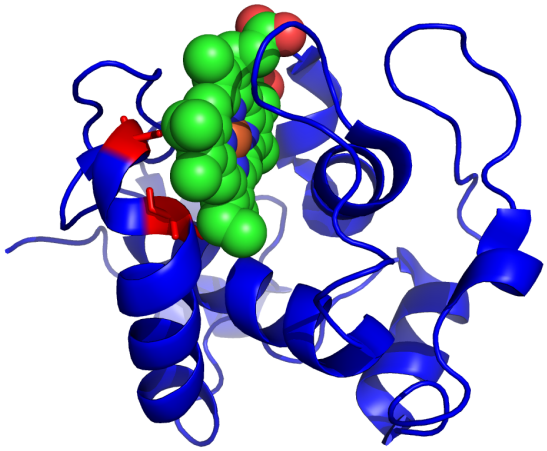


Figura 4.4: ***Citocromo c-L****. Los residuos en rojo (Cis^{79} y Cis^{82}) se unen covalentemente al grupo hemo.*

El siguiente ejemplo ilustra un problema de este método: las predicciones están directamente afectadas por la calidad de las anotaciones originales de *SCOP2GO*, ya que la colección de PERFILES_GO se construye a partir de este recurso. Para la ligasa aspartato-tRNA

(PDB:1b8aB), PSI_BLAST identifica correctamente el dominio de unión al ATP de esta proteína. Por el contrario, PERFILES_GO asigna incorrectamente la función "unión a ATP" (perfil GO:0005524/b.40) a la región N-terminal de la proteína. De hecho, este perfil no debería existir, ya que no existen proteínas con dominios de ese plegamiento concreto que unan ATP. Sin embargo, en *SCOP2GO* hay ejemplos de algunos de estos dominios incorrectamente anotados con esta función. El problema se debe a dos motivos: existen muchas instancias en *PDB* de dominios con plegamiento b.40 cristalizados en forma de fragmentos aislados, y las proteínas con el dominio b.40 suelen tener también un dominio de unión al ATP. Estos fragmentos, al estar anotados con la función de la cadena completa ("unión a ATP"), modifican artificialmente la distribución de plegamientos confundiendo la metodología de *SCOP2GO* (véase Materiales y Métodos, sec. 3.1.1). No obstante, como se discutió en la sección 4.1, este tipo de fallos en la anotación de *SCOP2GO* deberían ser anecdóticos y disminuirán a medida que aumente el número de proteínas cristalizadas (preferentemente completas) anotadas funcionalmente y se vaya revisando manualmente la base de datos *SCOP2GO*.

El siguiente ejemplo pone de manifiesto problemas asociados con la evaluación automática de estas predicciones, discutida en detalle en el siguiente punto. Para la quinasa de la caseína 1 (PDB:2csnA), PERFILES_GO identifica correctamente la función "actividad proteína quinasa" en el plegamiento "proteína similar a quinasa" (perfil GO:0004672/d.144). PSI_BLAST también predice correctamente el plegamiento d.144 pero lo asocia a la unión de ATP (GO:0005524). A pesar de que realmente el único dominio de la quinasa de la caseina 1 es dependiente de ATP, en este caso el método de evaluación automático considera que la predicción funcional de PSI_BLAST es incorrecta (falso negativo). Esto se debe a que la función "unión a ATP" no está originalmente asociada a la proteína en *GOA-PDB* y por tanto *SCOP2GO* no anota su dominio con este término GO.

Otro ejemplo que pone de manifiesto problemas asociados con la evaluación automática de estas predicciones es el caso de la

proteína-tirosina fosfatasa (fig. 4.1). En este caso, PERFILES_GO identifica correctamente la "actividad proteína-tirosina fosfatasa" en el plegamiento "proteína fosfotirosina" (perfil GO:0004725/c.45) del dominio N-terminal de la proteína. Sin embargo, la anotación original de *SCOP2GO* asocia este dominio con un término GO más general (GO:0004721 –"actividad fosfoproteína fosfatasa"–, que es padre directo de GO:0004725). A pesar de que en este caso la predicción funcional de PERFILES_GO es correcta (y, de hecho, más precisa), el método automático de evaluación la marca como un fallo, ya que no coincide con *SCOP2GO*. Por el contrario, la predicción de PSI_BLAST para esta proteína se contabiliza como un acierto, ya que en este caso recupera el término original (menos específico).

A pesar de que este tipo de falsos negativos y otros artefactos del método automático de evaluación afectan en cierto grado a los porcentajes de aciertos en la evaluación automática, en principio cabe esperar que afecte a ambas metodologías por igual, ya que no hay ninguna razón para pensar que PERFILES_GO o PSI_BLAST tienda a recuperar términos más o menos específicos.

4.2.2. Evaluación global

La figura 4.5 muestra las curvas ROC de las predicciones funcionales y estructurales generadas para las 1009 cadenas de *PDB* utilizando ambos métodos. Como se aprecia, tanto PERFILES_GO como PSI_BLAST presentan un elevado poder discriminatorio (fig. 4.5). Sin embargo, estos resultados muestran que las puntuaciones asociadas a las predicciones hechas usando los PERFILES_GO reflejan mejor la capacidad de recuperar correctamente la FM-GO originalmente asociada a un dominio de la secuencia problema (fig. 4.5).

Esto probablemente es debido a que cada perfil de la colección de PERFILES_GO se genera a partir de los alineamientos estructurales de todos los dominios asociados a un determinado término GO, y no solo de aquellos con un origen evolutivo común. De esta forma los perfiles capturan la información de todo el espacio de secuencias asociado a una FM-GO en

un plegamiento en lugar de restringirse a secuencias similares con la misma función.

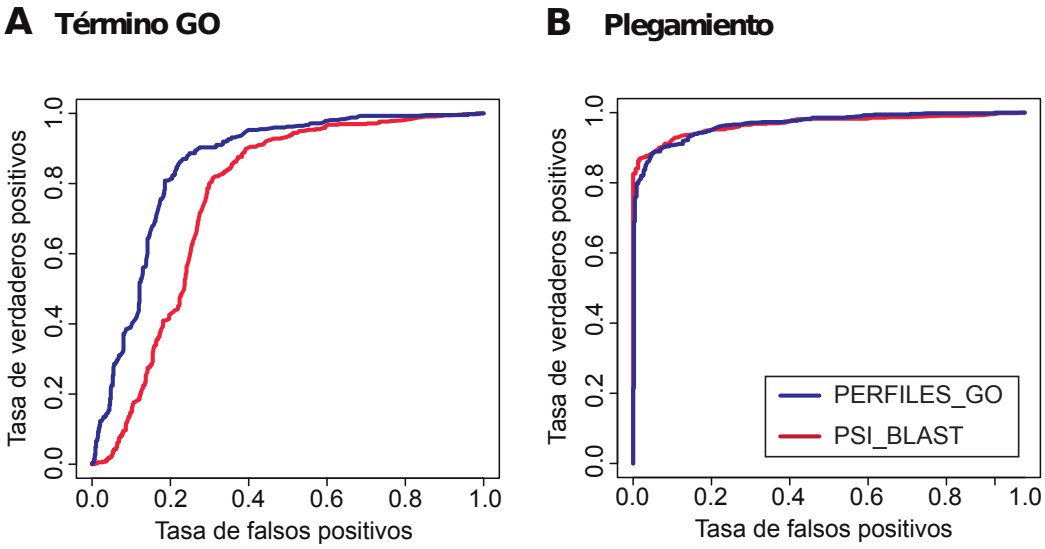


Figura 4.5: ***Evaluación a gran escala de la capacidad predictiva****. Las curvas ROC muestran la capacidad de ambos métodos para discriminar predicciones correctas de incorrectas para el conjunto de proteínas de prueba. Se tiene en cuenta únicamente la mejor predicción (menor valor-E). Se consideran correctas aquellas predicciones que recuperan el término GO (**A**) o el plegamiento de SCOP (**B**) del correspondiente dominio originalmente anotado en SCOP2GO.*

La diferencia entre ambos métodos varía cuando se tiene en cuenta la especificidad de los términos GO predichos (representando ésta por la "profundidad" en el grafo de GO) (fig. 4.6). Cuanto mayor es la distancia del término FM-GO predicho a la raíz del árbol de GO (mayor especificidad), menor es la diferencia entre PERFILES_GO y PSI_BLAST. Las curvas ROC para los dos métodos son prácticamente iguales cuando se evalúan sólo términos GO con una distancia a la raíz ≥ 6 (fig. 4.6). A pesar de que la distancia a la raíz no es un criterio perfecto para cuantificar la especificidad de los términos GO, es una manera sencilla de obtener una primera aproximación. Así, esta variación en la diferencia entre los métodos podría explicarse por el hecho de que las funciones más específicas están mejor reflejadas a nivel de secuencia y por tanto pueden ser capturadas por los métodos comunes basados en secuencia. Por el

contrario, las funciones generales (p.e. actividad hidrolasa) pueden haber divergido mucho o incluso tener un origen evolutivo distinto. Esto hace que sus características a nivel de secuencia sólo puedan ser capturadas por métodos basados en perfiles que no hayan sido construidos por homología para poder incluir secuencias más diversas.

4.2.3. Servidor web COPRED

La aplicación web *COPRED* permite que cualquier usuario pueda usar la metodología descrita anteriormente. Para ello implementa un sistema de búsqueda contra los 2938 PERFILES_GO procedentes de los alineamientos estructurales de dominios con la misma función. Estos perfiles engloban 580 términos FM-GO y 244 plegamientos de *SCOP* diferentes.

La figura 4.7 muestra capturas de pantalla representativas de la interfaz de usuario de COPRED. En la página de inicio, el usuario puede introducir una secuencia de aminoácidos como única entrada. Algunos parámetros de búsqueda se pueden modificar pinchando sobre "*advanced options*". La página de inicio también incluye un botón para rellenar el formulario con un ejemplo para probar el servidor.

Cuando se realiza una búsqueda, que normalmente lleva solo unos segundos, los resultados se muestran en una sola página con varios paneles desplegables. El primero representa la secuencia problema (azul en la fig. 4.7). Al expandir este panel se puede obtener cierta información de la secuencia de entrada así como el identificador del trabajo (*job ID*), que será necesario para recuperar posteriormente los resultados sin necesidad de realizar de nuevo la búsqueda. Los siguientes paneles muestran el resultado de la predicción, que básicamente consiste en una lista ordenada de perfiles alineados en distintas regiones de la secuencia problema. Para cada perfil se muestra el término FM-GO, el plegamiento de *SCOP* y el nivel de significación del alineamiento (valor-E), que viene reflejado también por el color de las barras (de verde a negro). De forma adicional, las posiciones conservadas de los perfiles, que podrían dar indicios sobre posibles sitios funcionales, se marcan con puntos rojos. La barra deslizante superior

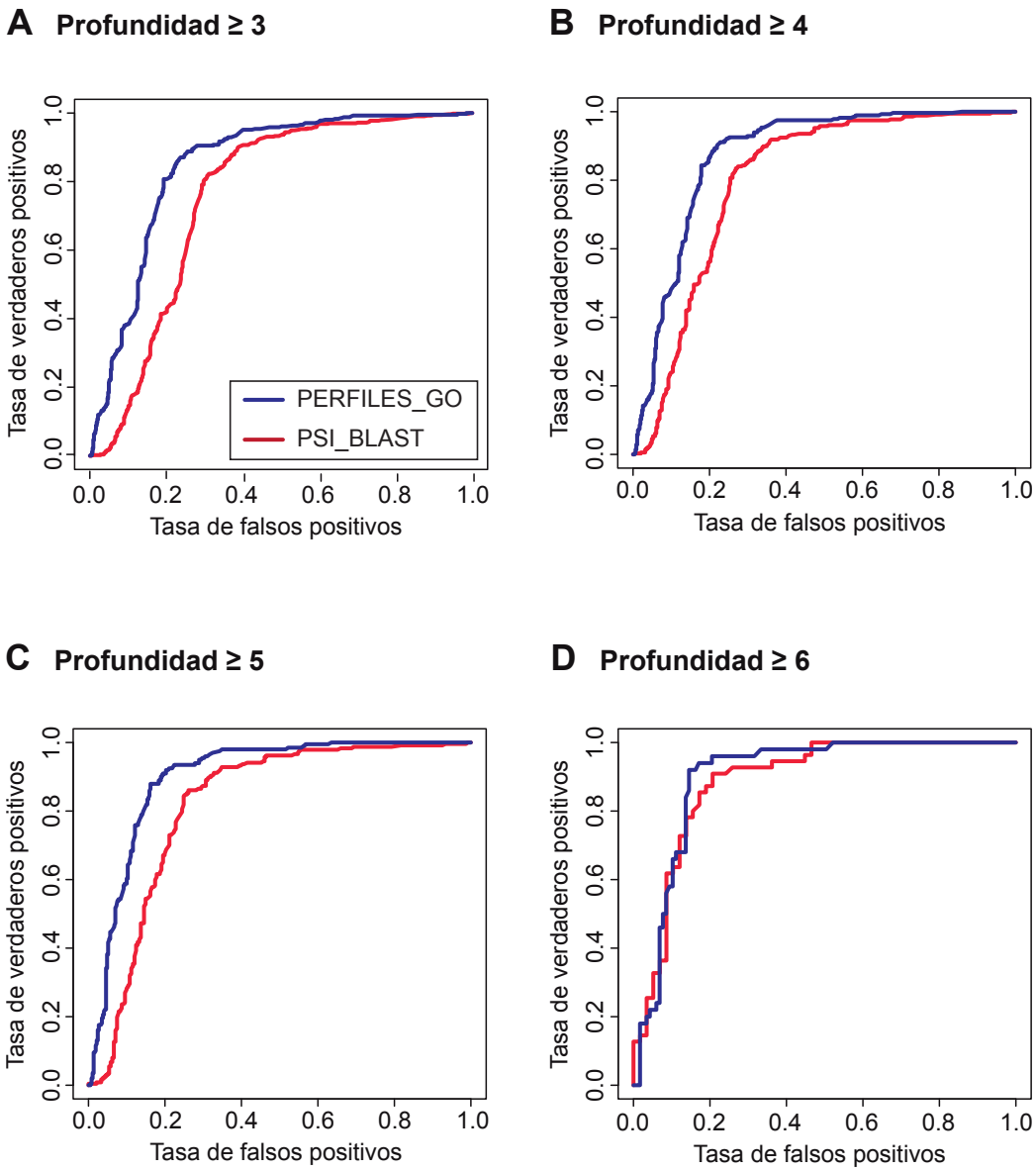


Figura 4.6: ***Evaluación a gran escala para términos GO de distinta profundidad (especificidad).*** *Las curvas ROC muestran la capacidad de ambos métodos para discriminar predicciones funcionales correctas de incorrectas para el conjunto de proteínas de prueba. Se tiene en cuenta únicamente la mejor predicción (menor valor-E). Se consideran predicciones correctas aquellas que recuperan el término GO originalmente asociado en SCOP2GO al dominio correspondiente. En cada caso se consideran sólo los términos GO con la profundidad mínima indicada, siendo la profundidad la distancia mínima a la raíz del árbol de GO.*

controla el umbral mínimo de conservación a considerar.

De esta forma, la lista colapsada de paneles da una idea general de los posibles dominios de la secuencia problema y de la posible función/estructura de cada uno de ellos. En el ejemplo mostrado en la figura 4.7, los resultados claramente indican la presencia de dos dominios: el N-terminal, asociado con el plegamiento c.37 y términos FM-GO relacionados con el GTP, y un dominio medio asociado al plegamiento b.43 y términos FM-GO relacionados con la unión de RNA. Estas predicciones están en concordancia con las características de la proteína del ejemplo (EF-Tu).

Al expandir un panel de resultados se obtiene información adicional del correspondiente perfil (fig. 4.7), como la descripción completa del termino FM-GO y el plegamiento de *SCOP*, así como enlaces a las correspondientes bases de datos. En la parte inferior del panel se muestra la secuencia problema con la región (dominio) que alinea con el perfil marcada en negrita, y los residuos correspondientes a las posiciones conservadas marcados en rojo. El panel también incluye una representación interactiva de la estructura tridimensional del miembro representativo del perfil. Esta vista, que se puede manipular (p.e. rotar o agrandar), también presenta remarcados los residuos conservados del perfil y los que alinean con la secuencia problema. Finalmente, la parte inferior del panel incluye un enlace para explorar el MSA asociado al perfil y la secuencia problema utilizando *Jalview* (Waterhouse et al., 2009) (fig. 4.7). Al abrirlo también se muestra el modelo tridimensional implícito de la región (dominio) de la secuencia problema en Jmol. Las posiciones conservadas se colorean en ambas representaciones.

La principal ventaja del servidor COPRED es que es sencillo de usar y proporciona una manera rápida de obtener las características funcionales y estructurales de los dominios de una proteína. A pesar de que muchos recursos actuales están empezando a adoptar un punto de vista orientado a dominios (de Lima Morais et al., 2011), COPRED es el primer recurso específicamente desarrollado para llevar a cabo esta tarea. Sin embargo, su principal limitación es el relativamente escaso número de perfiles en los que

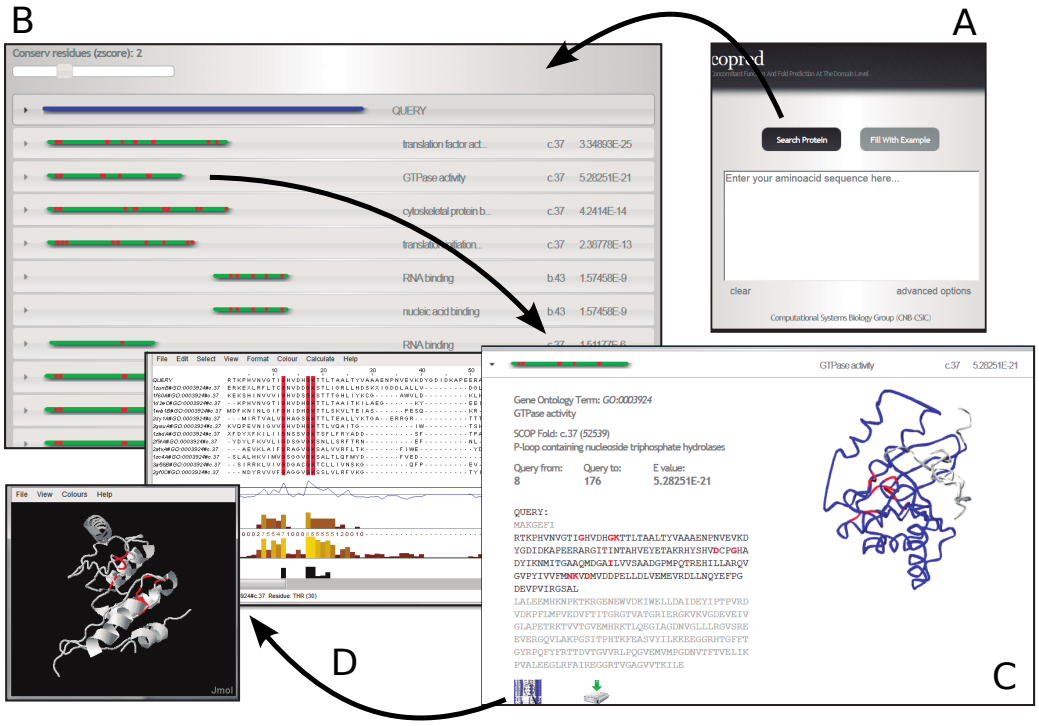


Figura 4.7: ***Capturas de pantalla del servidor COPRED (http://csbg.cnb.csic.es/copred/).*** *(**A**) Página de inicio, con enlaces para insertar un ejemplo, acceder a la página de ayuda y recuperar trabajos anteriores. (**B**) lista de resultados con los paneles colapsados. Los perfiles (en verde) alinean en diferentes regiones de la secuencia problema (en azul). La manipulación del botón deslizante superior permite mostrar posiciones conservadas en los perfiles (puntos rojos). (**C**) vista detallada de un panel de resultados. El panel contiene detalles de la región de la proteína problema (dominio) que alinea con el perfil, los posibles residuos funcionales, una vista interactiva tridimensional del representante del perfil, un enlace para explorar el resultado con Jalview (Waterhouse et al., 2009) y otro para descargar la información. (**D**) al abrir el resultado con Jalview se carga el MSA asociado al perfil junto con la secuencia problema. Adicionalmente se muestra el modelo tridimensional implícito de la secuencia problema en Jmol.*

está basado comparado con otros recursos. Esto se debe en parte a que, en lugar de usar bases de datos de secuencias, los PERFILES_GO se generan utilizando información estructural, que es mucho más escasa. De hecho, esta es una de las mayores diferencias con otros recursos que utilizan perfiles derivados de familias o superfamilias de dominios (Wilson et al., 2007). Estos otros recursos tienen la ventaja de que la similitud de una secuencia problema con alguno de sus perfiles proporciona información adicional sobre los miembros de su familia/superfamilia y su origen evolutivo. En este sentido, todos estos recursos se complementan unos a otros para la caracterización funcional, estructural y evolutiva de las proteínas a nivel de dominio. Además, en principio COPRED puede usar cualquier base de datos de anotaciones funcionales de dominios estructurales. Esto hace que el servidor pueda ser fácilmente ampliado para incluir un mayor conjunto de dominios anotados.

La aplicación web *COPRED* se puede acceder desde la dirección: http://csbg.cnb.csic.es/copred/

4.3. Características funcionales de las proteínas codificadas en el mRNA de manera independiente de los aminoácidos

Trabajos anteriores han estudiado la relación entre diferentes propiedades locales del mRNA relacionadas con la VE y aspectos estructurales de las proteínas. Sin embargo, nunca se ha hecho un estudio a gran escala que incluya anotaciones estructurales y funcionales. En esta tesis doctoral se analizó la relación entre tres propiedades del mRNA relacionadas con la VE obtenidas experimentalmente (estructura secundaria del mRNA, perfiles de densidad ribosómica y concentración de tRNA) y el conjunto de anotaciones locales estructurales y funcionales de *Uniprot* (UniProt Consortium, 2010) para las correspondientes proteínas (véase Materiales y Métodos, sec. 3.3). Para cada anotación de *UniProt*, se

calcularon los valores medios de las propiedades del mRNA en las proximidades de la anotación (patrón promedio) y se realizaron pruebas de significación estadística de la asociación entre un determinado patrón y la región funcional.

4.3.1. Análisis de los patrones promedio del mRNA

Algunas anotaciones estructurales de *UniProt* presentan cambios abruptos en los patrones promedio de las propiedades del mRNA analizadas. Por ejemplo, para "hélice α" y "lámina β", el patrón promedio de las tres propiedades del mRNA aumenta justo antes del inicio del elemento de estructura secundaria (posición 0) y desciende a continuación (fig. 4.8). Asumiendo que estas propiedades del mRNA son buenas aproximaciones de la VE, los resultados indican una deceleración justo antes y una aceleración en los primeros aminoácidos del motivo. De hecho, estos resultados son coherentes con estudios anteriores que utilizan otras propiedades del mRNA para aproximar la VE (Li et al., 2012).

Otro ejemplo se observa en la propiedad de *Uniprot* "metionina inicial", que marca el inicio de la región codificante (fig. 4.8). En este caso, los patrones medios de concentración de tRNA y ocupación ribosomal sugieren que hay una ralentización en los primeros codones de la anotación (fig. 4.8). Esta ralentización podría explicarse por la "rampa traduccional" descrita en estudios anteriores (Tuller et al., 2010; Dana y Tuller, 2012). Por el contrario, el patrón promedio de estructura secundaria del mRNA no muestra la rampa traduccional. A pesar de que está ampliamente aceptado que la estructura secundaria del mRNA y la disponibilidad de tRNA son factores clave determinantes para iniciación de la traducción y la VE, el mecanismo por el cual se logra tal efecto sigue siendo controvertido (Shabalina et al., 2013). En este sentido, la VE se aproxima mejor mediante la cuantificación de los perfiles de ocupación ribosomal, ya que se trata de una medida experimental directa de la velocidad a la que se traduce el mRNA.

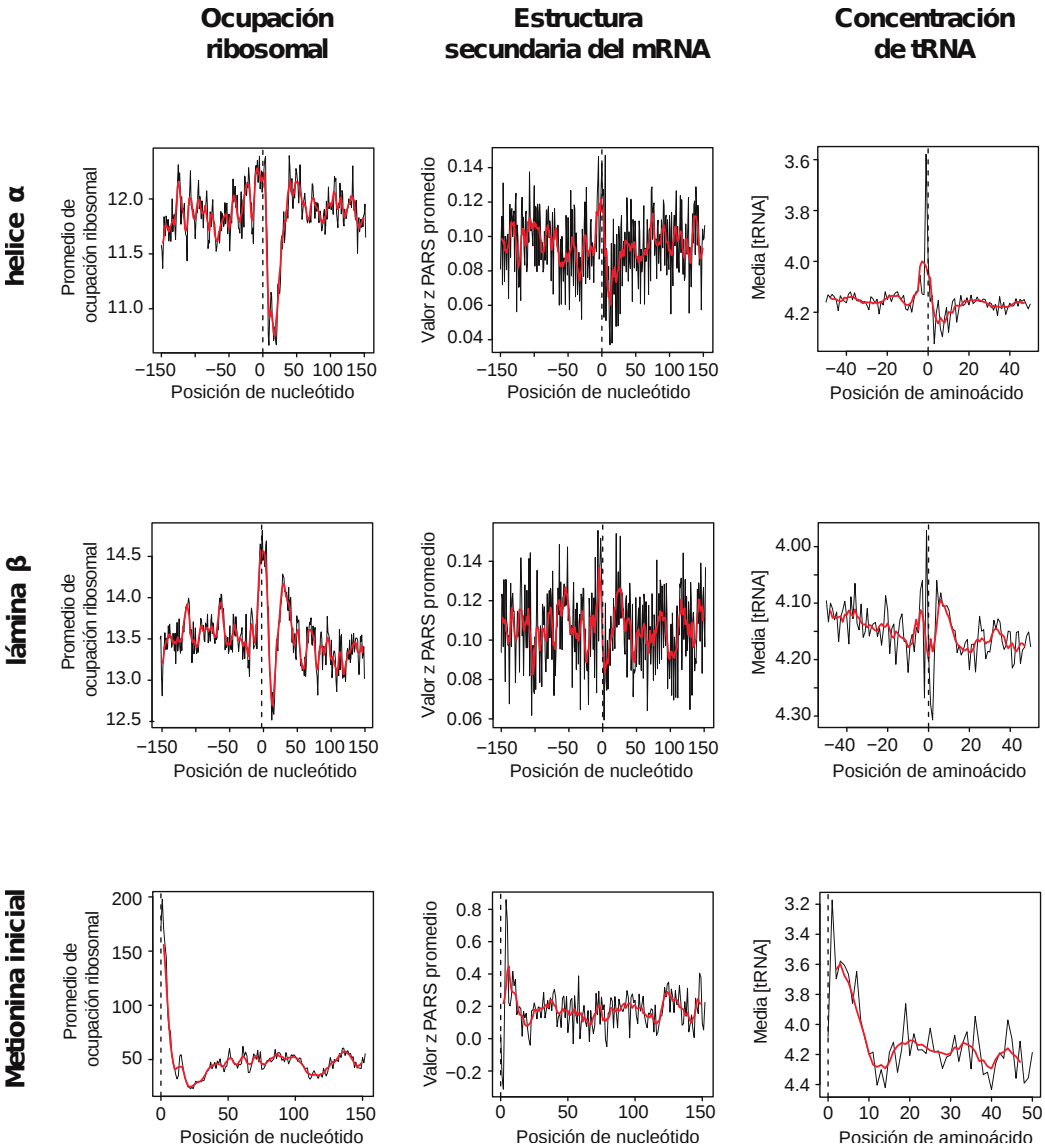


Figura 4.8: ***Patrón promedio de tres propiedades del mRNA para algunas anotaciones estructurales de UniProt****. Cada gráfica representa el patrón promedio de ocupación ribosomal (izquierda), estructura secundaria del mRNA (centro) y concentración de tRNA (derecha) para tres tipos de anotaciones estructurales de Uniprot: "hélice α", "lámina β" y "metionina inicial". La posición indica la distancia hasta el inicio de la anotación, donde los números negativos representan la región del mRNA antes de la anotación (aguas arriba) y el "0" el primer nucleótido del motivo estructural. La línea roja es un suavizado de los datos (media móvil con una ventana de 6 nucleótidos). La concentración de tRNA muestra los valores invertidos por su relación "inversa" con bajadas de la VE.*

Además de las anotaciones estructurales, también se analizaron los patrones promedio para las anotaciones funcionales de *UniProt* (véase Apéndice C, Patrones promedio para anotaciones funcionales y estructurales de proteínas). Por ejemplo, la anotación "unión a nucleótido fosfato" es probablemente la característica funcional mejor reflejada en los patrones promedio de las propiedades del mRNA analizadas (fig. 4.9). Para la ocupación ribosomal y la estructura secundaria del mRNA, los patrones promedio muestran una deceleración de unos 25 nucleótidos inmediatamente después del inicio del motivo. De hecho, esta región coincide con la longitud media de estas anotaciones (8 aminoácidos). El patrón promedio de concentración de tRNA, a pesar de que también muestra fuertes fluctuaciones en esa zona, difiere considerablemente de los otros dos. En parte, esto podría deberse al sesgo que produce el tipo de aminoácido en el cálculo de este patrón medio. Por ejemplo, la Cisteína es un aminoácido común en sitios activos enzimáticos (fig. 4.9). Este aminoácido está codificado por dos codones cuyos aa-tRNA se encuentran aproximadamente en la misma concentración ($\sim$ 2,46 %). Si al alinear una propiedad de *Uniprot*, una determinada posición estuviese enriquecida en ese aminoácido, la concentración media de tRNA tendería al mismo valor con independencia del codón utilizado (véase Materiales y Métodos, sec. 3.3.2). En cambio, para otros aminoácidos (p.e. Arginina) la diferencia de concentración de sus aa-tRNA puede ser de más de un orden de magnitud. Debido a esta dificultad para normalizar los valores de concentración de tRNA no se pudieron abordar las mismas pruebas estadísticas aplicadas a los patrones de estructura secundaria del mRNA y densidad ribosomal comentadas más adelante.

4.3.2. Análisis estadístico de los patrones promedio

Determinar si una anotación de *Uniprot* presenta un patrón promedio característico en las proximidades del mRNA que lo codifica no es trivial. Al fin y al cabo, el patrón promedio sólo muestra la tendencia general, pero no refleja la dispersión de los datos ni la especificidad de la relación entre el

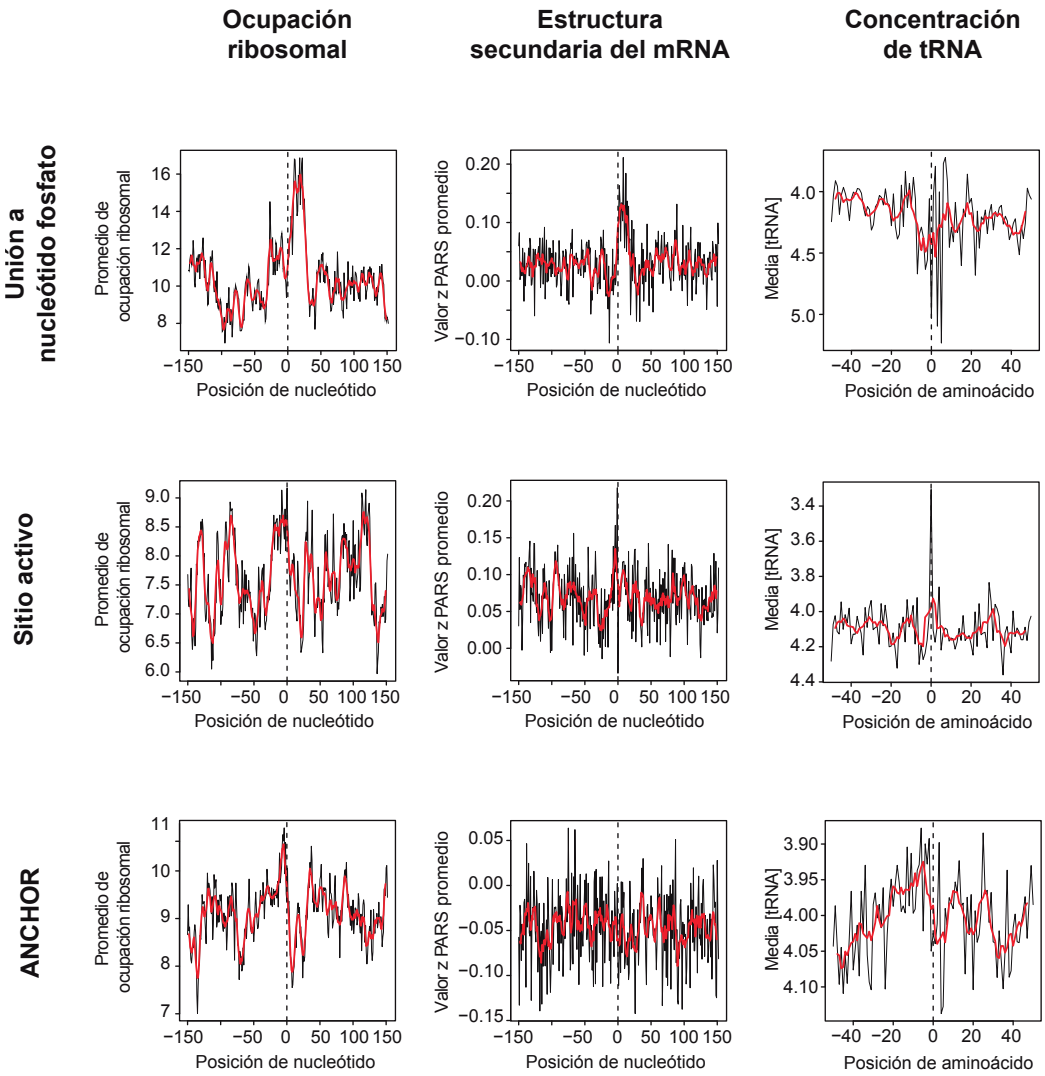


Figura 4.9: ***Patrón promedio de tres propiedades del mRNA para algunas anotaciones funcionales de UniProt.*** *Cada gráfica representa el patrón promedio de ocupación ribosomal (izquierda), estructura secundaria del mRNA (centro) y concentración de tRNA (derecha) para dos tipos de anotaciones funcionales de Uniprot y para las predicciones realizadas con ANCHOR. La posición indica la distancia hasta el inicio de la anotación, donde los números negativos representan la región del mRNA antes de la anotación (aguas arriba) y el "0" el primer nucleótido del motivo estructural. La línea roja es un suavizado de los datos (media móvil con una ventana de 6 nucleótidos). La concentración de tRNA muestra los valores invertidos por su relación "inversa" con bajadas de la VE.*

patrón y la anotación de *UniProt*. Más aún, los patrones claros, aquellos con fuertes subidas o bajadas, pueden ser detectados fácilmente, pero otros más complejos podrían no ser evidentes a simple vista. Por ejemplo, mientras que el patrón promedio de ocupación ribosomal para las predicciones de *ANCHOR* presenta una fuerte bajada justo al inicio de la anotación, el de estructura secundaria parece mantenerse constante a lo largo de la ventana analizada (fig. 4.9). En este sentido, la prueba estadística de diferencia de medias de correlaciones se hace necesaria para determinar qué patrones son significativos (sec. 3.3.2). En el caso de *ANCHOR*, ambos patrones promedio presentan resultados significativos (valor-$P \leq 0{,}05$) (tabla 4.6). Estos resultados indican que, tanto los valores de ocupación ribosomal como los de estructura secundaria del mRNA en torno a las regiones predichas por *ANCHOR* están más correlacionados entre sí de lo que cabría esperar por azar (tomando ventanas de valores aleatoriamente del conjunto de mRNA).

La tabla 4.6 muestra los resultados de la prueba estadística para todas las anotaciones de *UniProt* para las que se pudo realizar la prueba, junto con las regiones desordenadas predichas por *ANCHOR* y *IUPred*. Los resultados muestran que muchas de estas anotaciones, especialmente las relacionadas con elementos de estructura secundaria, presentan patrones promedio de estructura secundaria del mRNA y ocupación ribosomal característicos y significativamente diferentes del resto de las secuencias codificantes (tabla 4.6).

En algunos casos, esto podría reflejar un mecanismo sutil de regulación local mediante la variación de la VE. Esta regulación podría incluso ser necesaria para la adquisición de determinadas características regionales en las proteínas. Asumiendo que las proteínas comienzan a plegarse conforme van siendo sintetizadas (plegamiento co-traduccional), la regulación de la VE podría permitir, por ejemplo, la ralentización en una determinada región para que un ligando se inserte en ella antes de que se termine de plegar o le afecten otras regiones de la proteína. En estos casos, el mecanismo de regulación de la VE podría jugar un papel clave en la determinación funcional de algunas proteínas.

Tabla 4.6: ***Resultado de la prueba estadística para evaluar la relación entre distintas anotaciones protéicas y el patrón de dos propiedades del mRNA que las codifica****. Valor-P de la prueba estadística para los patrones promedio de las anotaciones de Uniprot y las predicciones de ANCHOR e IUPred. Los guiones indican que no se pudo realizar la prueba estadística por un número insuficiente de instancias de la anotación.*

Anotación protéica	**Oc. rib.**	**Estr. sec.**
hélice α	**3,34E-036**	**4,23E-068**
unión a nucleótido fosfato	**0,00E+000**	**1,54E-059**
péptido señal	**0,00E+000**	**9,04E-050**
lámina β	**2,79E-026**	**4,67E-046**
giro	**4,64E-024**	**2,55E-041**
región transmembrana	**9,67E-009**	**6,47E-028**
dominio topológico	**1,37E-015**	**9,70E-021**
IUPred	**4,97E-033**	**4,63E-017**
sitio de unión	**5,27E-017**	**2,21E-014**
dominio	**7,89E-058**	**6,84E-012**
repetición	**4,02E-004**	**8,43E-012**
sitio	5,59E-002	**6,42E-009**
sitio activo	**1,83E-034**	**7,36E-009**
hélice superenrollada (*coiled-coil*)	**1,92E-002**	**5,17E-008**
región de interés	**1,51E-003**	**4,09E-007**
péptido	**7,40E-009**	**2,19E-006**
sitio de unión a ion metálico	**2,06E-036**	**4,40E-006**
residuo modificado	**2,50E-017**	**3,18E-005**
ANCHOR	**3,84E-126**	**2,25E-003**
metionina inicial	**0,00E+000**	**2,60E-003**
región de unión a motivo lipídico	**1,67E-002**	**3,19E-003**
motivo de secuencia corta	**7,19E-017**	**4,50E-003**
región de unión a calcio	-	**5,53E-003**
reticulación (*Cross-link*)	7,47E-001	**3,49E-002**
variante de corte	**1,84E-006**	2,63E-001
región de composición sesgada	**4,09E-002**	3,01E-001
enlace disulfuro	**2,40E-003**	3,71E-001
región dedo de zinc	**1,68E-005**	4,90E-001
propéptido	8,33E-002	7,08E-001
región de unión a DNA	**1,14E-051**	8,49E-001
región intramembrana	1,39E-001	-
aminoácido no estándar	5,66E-001	-

Los resultados también apoyan la idea de que el mRNA codifica más que una simple cadena de aminoácidos. Estos resultados ayudan a esclarecer la relación entre función protéica y su codificación en el genoma. Además, las relaciones encontradas entre las propiedades del mRNA y algunas características estructurales y funcionales de las proteínas podrían ayudar a concebir nuevas formas de diseño de proteínas. Mientras que los cambios en los aminoácidos tienen un efecto drástico en la función, jugar con las propiedades del mRNA relacionadas con la VE podría ser una forma sutil de modificar la función. Además, estos resultados podrían tener también implicaciones prácticas en la expresión de proteínas. Actualmente, la estrategia general para la expresión de proteínas de forma heteróloga suele consistir en usar los codones "óptimos" del organismo huésped para aumentar la eficiencia global de la traducción. Sin embargo, se ha visto que la alteración de la secuencia del mRNA, aun codificando los mismos aminoácidos, puede tener resultados inesperados, como el descenso en la producción de proteína o cambios en su actividad que pueden incluso afectar a la viabilidad global del organismo (Agashe et al., 2012). En algunos casos, esto podría deberse a la alteración del patrón local de propiedades del mRNA relacionadas con la VE. De acuerdo con los resultados obtenidos, el gen insertado debería diseñarse tratando de respetar los patrones originales.

A pesar de que en la mayoría de los casos los patrones encontrados son muy sutiles, para determinadas características funcionales con patrones claros éstos podrían usarse para predecir sitios funcionales en proteínas a partir de su secuencia genética. En la actualidad, casi todos los métodos de predicción de sitios funcionales requieren numerosas secuencias de la misma familia para, por ejemplo, detectar posiciones conservadas en un MSA. En cambio, esta técnica permitiría caracterizar funcionalmente nuevas proteínas con independencia de su similitud con otras proteínas disponibles ya anotadas. Además, la posibilidad de usar técnicas de alto rendimiento, como los perfiles de ocupación ribosomal, permitirían aplicarla de forma masiva a genomas completos.

4.3.2.1. Ejemplo

La figura 4.10 muestra una representación lineal de una región de la proteína FtsH de *E. coli* con distintas anotaciones funcionales y estructurales extraídas de *UniProt*. La figura muestra los patrones de ocupación ribosomal y concentración de tRNA para el mRNA que codifica esa región de la proteína. El patrón de ocupación ribosomal presenta cuatro picos claros en las proximidades de la región desordenada predicha por *IUPred* y las anotaciones funcionales de *UniProt* "unión a nucleótido fosfato" y "sitio". En este ejemplo, el patrón de ocupación ribosomal para "unión a nucleótido fosfato" coincide con el patrón promedio para esta propiedad del mRNA (fig. 4.9). Sin embargo, el patrón para concentración de tRNA no es muy claro en este caso.

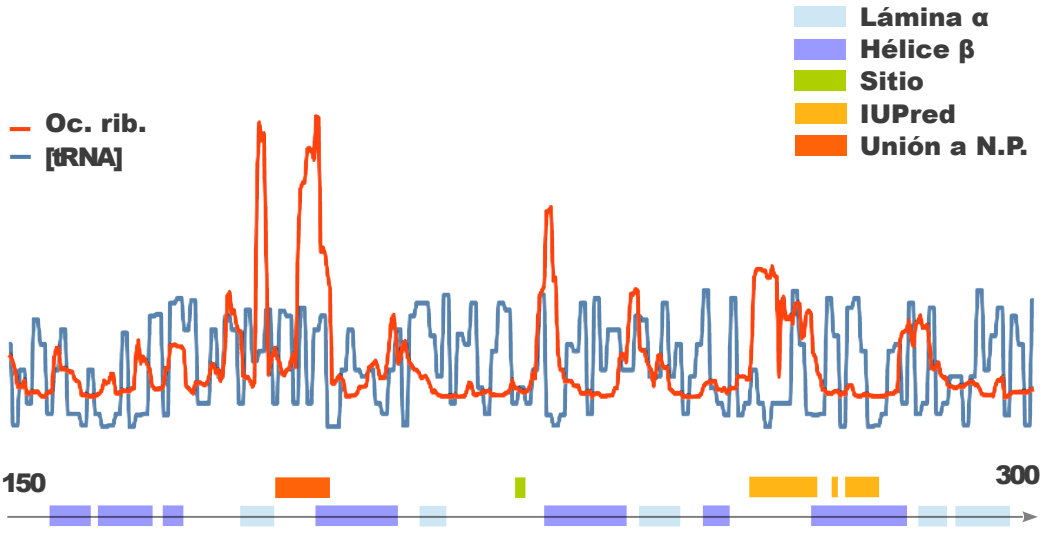


Figura 4.10: ***Patrones de propiedades del mRNA que codifica la FtsH.*** *La figura muestra el patrón de ocupación ribosomal y el de concentración de tRNA para la región 150-300 del mRNA que codifica la metaloproteasa dependiente de zinc FtsH. Las cajas de colores representan anotaciones de Uniprot para esa proteína.*

Capítulo 5

Conclusiones

1. Los métodos de predicción funcional basados en la anotación de cadenas completas no deben ser usados para predecir la función molecular de dominios individuales.

2. La metodología y el recurso desarrollado para la anotación funcional de dominios de proteínas, *SCOP2GO*, permite el desarrollo de métodos de predicción funcional orientados a dominios.

3. La predicción funcional de dominios utilizando perfiles basados en alineamientos estructurales de dominios anotados con la misma función mejora sustancialmente comparada con un método estándar de predicción funcional basado en homología.

4. La relación encontrada entre propiedades del mRNA relacionadas con la VE y determinadas características funcionales y estructurales locales de las proteínas ayuda a esclarecer la relación entre función proteica y su codificación en el genoma.

5. Estos resultados sugieren que, en algunos casos, esta relación podría ser el reflejo de un mecanismo sutil de regulación de aspectos funcionales locales mediante la variación local de la VE.

Apéndice A

Recursos utilizados

A.1. Herramientas bioinformáticas

Dalilite (Holm y Park, 2000). Versión autónoma del programa de alineamiento estructural de pares de proteínas *DALI* (Holm y Sander, 1995). Para realizar el alineamiento, *DALI* calcula la matriz de distancias entre fragmentos (hexapéptidos) de las proteínas.

T-coffee (Notredame et al., 2000). Programa de alineamiento múltiple de secuencias basado en alineamientos binarios globales.

BLAST (Altschul et al., 1997). Algoritmo heurístico diseñado para hacer comparaciones locales de secuencias biológicas. *BLAST* suele hacer referencia a la familia de programas que implementan este algoritmo. El valor-E describe el número de secuencias que se espera encontrar simplemente por azar.

PSI-BLAST. Variante de BLAST (Altschul et al., 1997) usada para localizar posibles proteínas homólogas remotas. Al inicio, utiliza *Blastp* para encontrar proteínas cercanas (similares) en la base de datos. Compila las proteínas en un perfil PSSM y busca el perfil contra la base de datos para encontrar más proteínas relacionadas que se añaden al perfil. El proceso se repite tantas veces como haya especificado el usuario o hasta que no se encuentren más proteínas.

RPS-BLAST. Variante de *BLAST* (Altschul et al., 1997) usada para buscar una secuencia contra una base de datos de perfiles PSSM utilizando

un algoritmo similar a *BLAST*.

IUPred (Dosztányi et al., 2005). Programa de predicción de regiones proteicas intrínsecamente desordenadas. El método de predicción está basado en la estimación de la capacidad de los polipéptidos de formar contactos estabilizadores entre residuos, usando como conjunto de entrenamiento proteínas globulares de estructura conocida.

ANCHOR (Mészáros et al., 2009). Programa de predicción de regiones desordenadas potencialmente involucradas en la interacción entre proteínas. El algoritmo busca regiones desordenadas predichas por *IUPred* que pueden tener la capacidad de formar contactos estabilizadores entre residuos al interaccionar con otra proteína globular.

A.2. Bases de datos

SCOP (Andreeva et al., 2004). Base de datos de dominios estructurales clasificados manualmente en base a su similitud de estructura y secuencia. Además su objetivo es describir las relaciones evolutivas entre todas las proteínas con estructura conocida. Los dominios se clasifican en una jerarquía de 7 niveles. Por ejemplo, los dominios con un origen evolutivo común se clasifican dentro de la misma categoría "superfamilia", y aquellos con clara similitud de secuencia dentro de la misma "familia". La categoía "plegamiento" únicamente implica similitud general de estructura terciaria entre dominios.

GO (Harris et al., 2004). Proyecto cuyo objetivo es describir las propiedades de genes y productos de genes mediante un vocabulario unificado de términos englobados en tres dominios independientes: *componente celular* (partes de una célula y su ambiente extra-celular), *función molecular* (la actividad elemental de un producto genético a nivel molecular) y *proceso biológico* (colección de eventos moleculares con un principio y un fin definido). Asimismo, GO define códigos de evidencia para las anotaciones distinguiendo la fuente de la anotación, según sea experimental, computacional, revisada por expertos o inferida automáticamente.

PDB (Berman et al., 2000). Base de datos de estructuras tridimensionales de moléculas biológicas, incluyendo proteínas y ácidos

nucleicos. La mayoría de estas estructuras han sido obtenidas mediante cristalografía de rayos-X o resonancia magnética nuclear.

UniProtKB/Swiss-Prot (UniProt Consortium, 2010). Base de datos de anotaciones funcionales de proteínas de alta calidad revisadas manualmente por expertos.

GOA-PDB. Base de datos de anotaciones de cadenas *PDB* con términos GO. Estas anotaciones se transfieren de la entrada de *UniprotKB* asociada (aquella con al menos un 90 % de identidad de secuencia). Como todas las anotaciones de GO, vienen acompañadas de un código de evidencia que hace referencia al origen y, por ende, la calidad de dichas anotaciones (por ejemplo anotaciones manuales e inferidas electrónicamente entre otras).

InterPro (Hunter et al., 2012). Recurso que recopila motivos funcionales de proteínas de 11 bases de datos entre las que se incluyen algunas clasificaciones de familias de dominios como Pfam (Punta et al., 2012) o SUPERFAMILY (de Lima Morais et al., 2011).

InterPro2go (Hunter et al., 2012). Base de datos de asociaciones entre entradas de *InterPro* y términos GO. Para generarla, primero agrupa las entradas de *InterPro* pertenecientes a la misma familia de proteínas o al mismo dominio. Este recurso se cruza con *UniprotKB* para detectar las proteínas cuya secuencia alinea con dominios de *InterPro*. Cuando un dominio de *InterPro* alinea con un conjunto de proteínas funcionalmente similares, éste se anota con los términos GO que describen la función o la localización de las proteínas. Las anotaciones son revisadas manualmente por expertos.

GOtcha (Martin et al., 2004). Servidor diseñado para predecir la función de una secuencia problema mediante la asignación de términos GO. El proceso se basa en la transferencia de anotaciones funcionales desde secuencias similares encontradas mediante *BLAST*.

Astral (Chandonia et al., 2004). Conjunto de bases de datos y herramientas para analizar la estructura y secuencia de las proteínas. La información deriva en gran parte de *SCOP* (Andreeva et al., 2004). Entre otras cosas, facilita la obtención de la secuencia de aminoácidos de cada dominio de *SCOP*.

Superligands (Michalsky et al., 2005). Base de datos de estructuras de pequeñas moléculas etiquetadas como "ligandos" en *PDB*. El servidor permite además buscar moléculas similares a un determinado compuesto basándose en el coeficiente de similitud de Tanimoto (Holliday et al., 2002).

GEO (Barrett et al., 2013). Repositorio público principalmente de datos de *microarrays* y NGS enviados por la comunidad científica.

A.3. Pruebas estadísticas y normalizaciones

Coeficiente de correlación de Pearson. Medida del grado de dependencia lineal entre dos variables. Las variables fuertemente correlacionadas presentan valores cercanos a 1 o -1 mientras que en las independientes el valor tiende a 0. Sin embargo, el coeficiente de Pearson no caracteriza completamente la relación entre dos variables: una alta correlación no implica necesariamente una relación de linealidad y una baja correlación no es necesariamente debida a la independencia entre las variables (p.e. pueden tener una dependencia no lineal). La correlación tampoco implica causalidad. Para una muestra, el coeficiente de correlacón de pearson r viene dado por:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \tag{A.1}$$

Prueba t de Student. Cualquier prueba en la que el estadístico utilizado sigue una distribución t de Student cuando se acepta la hipótesis nula. Se puede usar para determinar si dos conjuntos de datos difieren de forma significativa mediante la comparación de sus medias. Si los datos de cada población siguen una distribución normal, se puede asumir que tiene la misma desviación estándar (σ) y se han obtenido de forma independiente para cada población, el estadístico t viene dado por:

$$t = \frac{\bar{X_1} - \bar{X_2}}{\sigma_{X_1X_2}\sqrt{\frac{2}{n}}} \tag{A.2}$$

donde

$$\sigma_{X_1X_2} = \sqrt{\frac{1}{2}(\sigma_{X_1}^2 + \sigma_{X_2}^2)} \tag{A.3}$$

Distribución hipergeométrica. Distribución discreta con muestreos aleatorios y sin reemplazo donde los elementos pueden ser clasificados en dos categorías. Suponiendo una población de N elementos de los cuales K pertenecen a la categoría A, la distribución hipergeométrica mide la probabilidad de obtener k elementos de la categoría A de una muestra de n elementos de la población original. La función de probabilidad de una variable aleatoria X que sigue dicha distribución viene dada por la ecuación

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \tag{A.4}$$

Curva ROC. Ilustra la variación de eficacia de un sistema de clasificación binaria en función del umbral de discriminación (puntuación). Cada punto de una curva ROC representa la fracción de verdaderos positivos ($\frac{VP}{VP+FN}$) frente a la fracción de falsos positivos ($\frac{FP}{FP+VN}$) para un determinado umbral de puntuación, donde VP, VN, FP, FN hace referencia a verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos. El AUC (Área bajo la curva ROC) se utiliza comúnmente para cuantificar de manera global la capacidad del sistema de puntuación de separar predicciones correctas de incorrectas.

Valor-Z. También llamado valor típico (Z), indica el número de desviaciones típicas (σ) que un dato está por encima o por debajo de la media (μ). Es un tipo de normalización usado comúnmente para comparar datos de diferentes poblaciones. Para un valor x éste viene dado por

$$Z_x = \frac{x - \mu_x}{\sigma_x} \tag{A.5}$$

Valor-P. En una prueba estadística de contraste de hipótesis, el valor-P se define como la probabilidad de obtener un resultado al menos tan extremo como el observado cuando se asume que la hipótesis nula es cierta. Normalmente, la hipótesis nula "se rechaza" cuando el valor-P es menor que un nivel de significación predeterminado (p.e. 0,05).

Apéndice B

Conjunto de proteínas multidominio

Multidomain Protein	One-Domain Protein(s)	domain in scop2go	domain matching	GO term	GO description
Endo/exocellulase:cellobiose E-4 1js4 3.2.1.4	Glucoamylas 1ayx 3.2.1.3	d1js4a1	superfamily	GO:0004553	hydrolase activity, hydrolyzing O-glycosyl compounds
	b.2.2				
β-amylase 1b90 3.2.1.2	β-amylase 1bfn 3.2.1.2	d1b90a2	family	GO:0016161	beta-amylase activity
	b.3.1				
α-toxin 1ca1 3.1.4.3	Bacterial phospholipase C 1ah7 3.1.4.3	d1ca1_1	family	GO:0004629	phospholipase C activity
	b.12.1				
Adenylate kinase 1ak2 2.7.4.3	Guanylate kinase 1ex7 2.7.4.8	d1ak2_1	family	GO:0004017	adenylate kinase activity
	g.41.2				
Methionyl-tRNAfmet formyltransferase 2fmt 2.1.2.9	Glycinamide ribonucleotide transformylase (GART) 1jkx 2.1.2.2	d2fmta2	family	GO:0016742	hydroxymethyl-, formyl- and related transferase activity
	3-methyladenine DNA glycosylase 1ewn 3.2.2.21	?			
Aspartyl-tRNA synthetase (AspRS) 1asy 6.1.1.12	Asparagine synthetase 12as 6.3.1.1	d1asya2	family	GO:0004812	aminoacyl-tRNA ligase activity
	ssDNA-binding protein 1eyg	?			
Chemotaxis receptor methyltransferase 1af7 2.1.1.79	RNA methyltransferase FtsJ 1ej0 2.1.1.-	d1af7_2	superfamily	GO:0008757	S-adenosylmethionine-dependent methyltransferase activity
	a.58.1				
Lytic transglycosylase Slt70 1qsa 3.2.1.-	Lysozyme 1lys 3.2.1.17	d1qsaa2	superfamily	GO:0004553	hydrolase activity, hydrolyzing O-glycosyl compounds
	a.118.5				
DNA polymerase β 1bpd 2.7.7.7	Kanamycin nucleotidyltransferase (KNTase) 1kny 2.7.7.-	d1bpd_4	superfamily	GO:0016779	nucleotidyltransferase activity
	a.60.6				
Carbamoyl phosphate synthetase small subunit 1a9x 6.3.5.5 (Whole complex)	γ-glutamyl hydrolase 1l9x 3.4.19.9				
	c.8.3				
Prolyl oligopeptidase 1e5t 3.4.21.26	Proline iminopeptidase 1azw 3.4.11.5	d1e5ta2	superfamily	GO:0004287	prolyl oligopeptidase activity
	b.69.7				
Methionyl-tRNA synthetase 1a8h 6.1.1.10	Tryptophanyl-tRNA synthetase 1m83 6.1.1.2	d1a8h_2	family	GO:0004812	aminoacyl-tRNA ligase activity
	a.27.1				
Glutaminyl-tRNA synthetase 1euq 6.1.1.18	Tryptophanyl-tRNA synthetase 1m83 6.1.1.2	d1euqa2	family	GO:0004818	glutamate-tRNA ligase activity
	Ribosomal protein L25 1dfu	?			

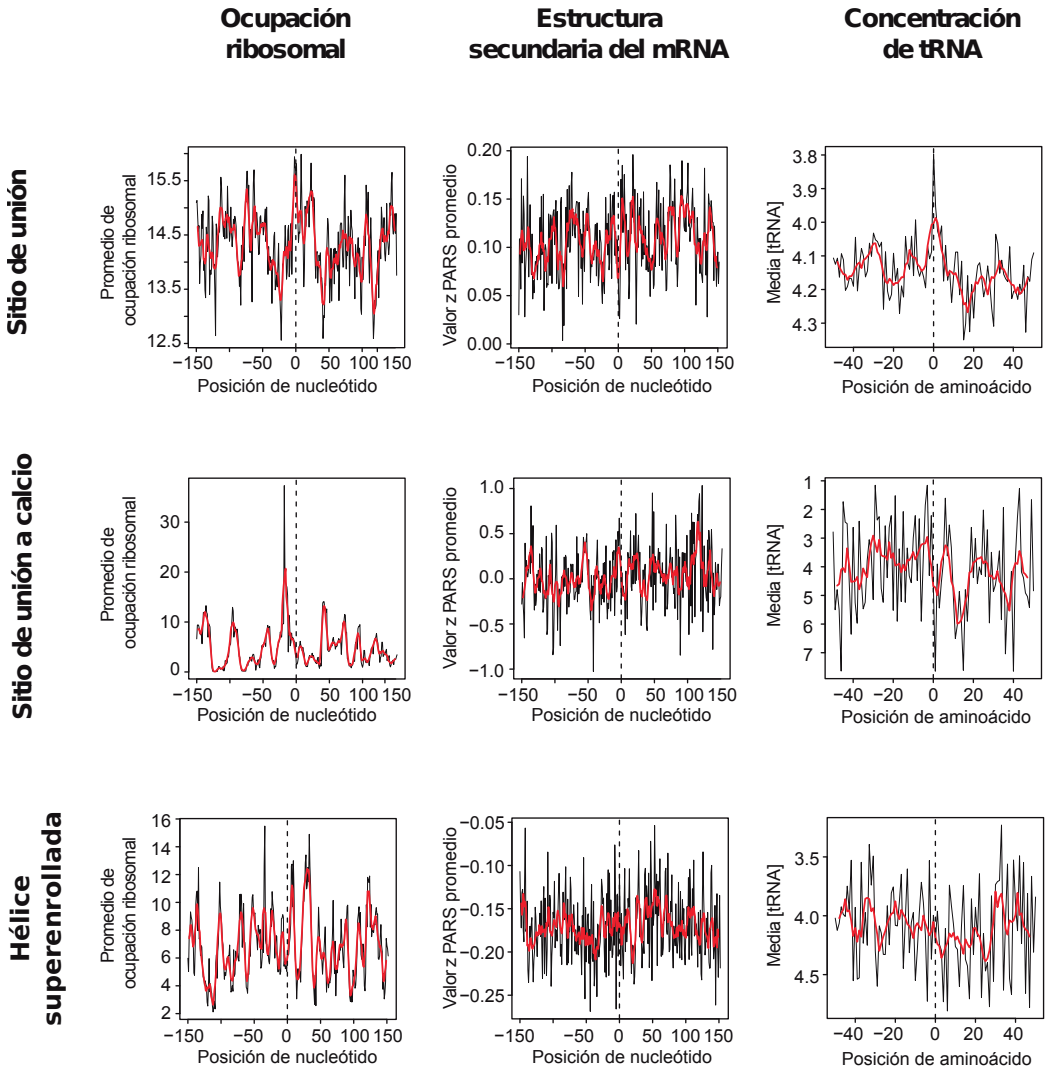
Multidomain Protein	One-Domain Protein(s)	domain in scop2go	domain matching	GO term	GO description
Cytochrome cd1 1aof 1.9.3.2	Mitochondrial cytochrome c6 1c75	d1aofa1	superfamily	GO:0005506	iron ion binding
				GO:0009055	electron carrier activity
				GO:0020037	heme binding
	b.70.2				
Naphthalene 1,2-dioxygenase α subunit 1eg9 1.14.12.12	Soluble, respiratory-type Rieske protein 1nyk	?			
	Phoshatidylinositol transfer protein (PITP) 1fvz	?			
Nitrous oxide reductase 1qni 1.7.99.6	Plastocyanin 1pcs	d1qnia1	superfamily	GO:0004129	cytochrome-c oxidase activity
	b.69.3				
Ruberdoxin: oxygen oxidoreductase 1e5d	Flavodoxin 1ag9	d1e5da1	family	GO:0010181	FMN binding
	Zn metallo-β-lactamase 1dxk 3.5.2.6	?			
Iron hydrogenase large (catalytic) subunit 1hfe 1.18.99.1	Ferredoxin II 1fxd	?			
	c.96.1				
Rubrerythrin 1b71	Ruberdoxin 1rb9	?			
	Bacterioferritin (cytochrome b1) 1bcf	d1b71a1	family	GO:0046914	transition metal ion binding
				GO:0009055	electron carrier activity
Peptidase T 1fno 3.4.11.4	Carboxypeptidase A 1f57 3.4.17.1	d1fnoa4	superfamily	GO:0008235	metalloexopeptidase activity
	d.58.19				
β-glucuronidase 1bhg 3.2.1.31	β-glucanase 1ghs 3.2.1.39	d1bhga3	family	GO:0004553	hydrolase activity, hydrolyzing O-glycosyl compounds
	Fucose-binding lectin 1k12	?			
β-galactosidase 1jz7 3.2.1.23	Glactose mutarotase 1nsx 5.1.3.3	?			
	b.1.4				
Penicillin-recognizing enzyme 1ei5 3.4.11.19	β-lactamase 1erm 3.5.2.6	d1ei5a3	family	GO:0008800	beta-lactamase activity
	b.61.3				
Tyrosine phosphatase 2shp 3.1.3.48	Dual-specificity protein phosphatase VHR 1vhr 3.1.3.48	d2shpa1	superfamily	GO:0004725	protein tyrosine phosphatase activity
	d.93.1				
R1 subunit of ribonucleotide reductase. Catalyzes the synthesis of deoxyribonucleotides 1r1r 1.17.4.1	B12-dependent (class II) ribonucleotide reductase 1l1l 1.17.4.2	d1r1ra2	superfamily	GO:0004748	ribonucleoside-diphosphate reductase activity
	a.98.1				
Molybdate-dependent transcriptional regulator ModE 1b9m	MARR antibiotic-resistance repressor 1jgs	d1b9ma1	superfamily	GO:0003700	transcription factor activity
	Molybdate/tungstate-binding protein II 1gug	d1b9ma3	superfamily	GO:0030151	molybdenum ion binding
		d1b9ma4	superfamily	GO:0030151	molybdenum ion binding

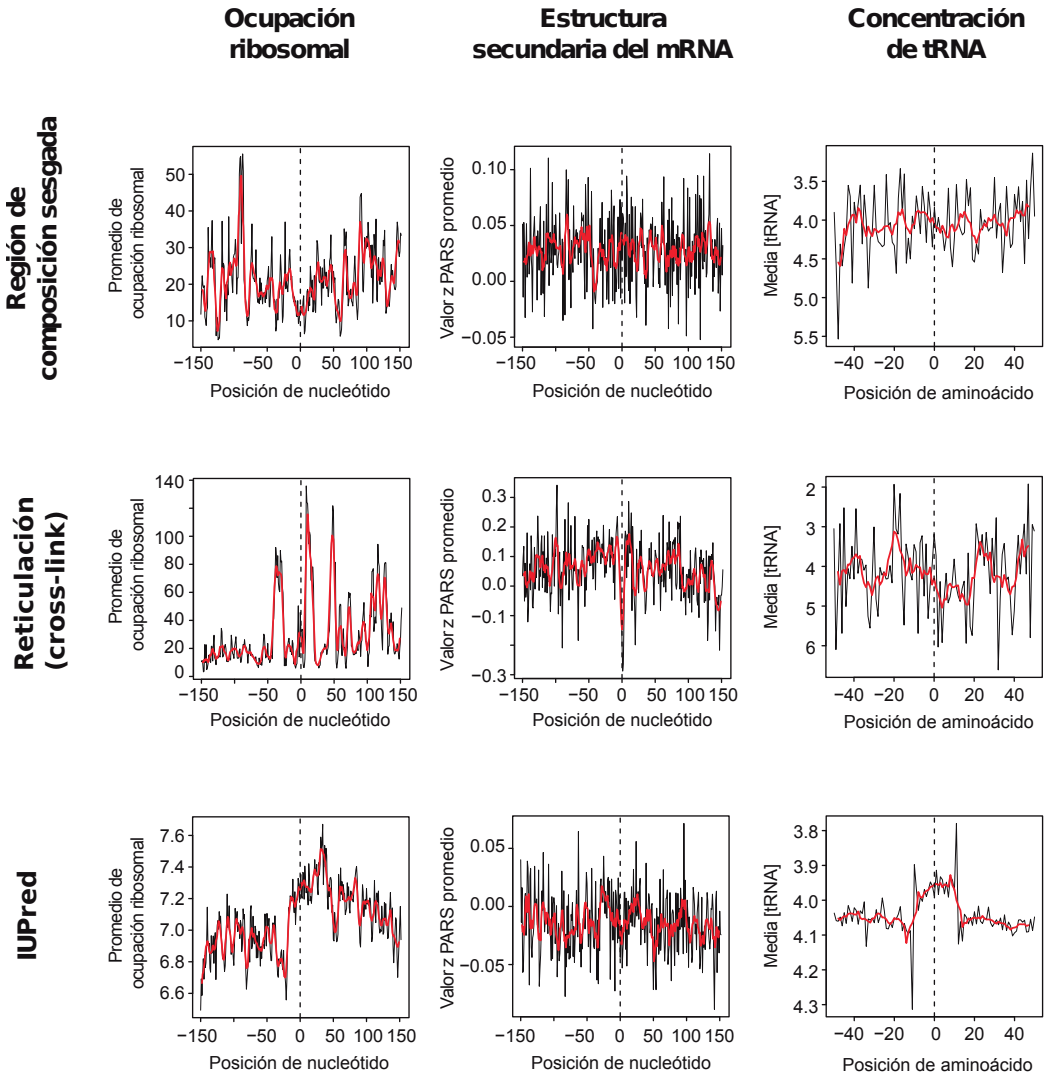
Multidomain Protein	One-Domain Protein(s)	domain in scop2go	domain matching	GO term	GO description
Asparagine synthetase B 1ct9 6.3.5.4	Penicillin V acylase 2pva 3.5.1.11				
	NH3-dependent NAD+-synthetase 1kqp 6.3.5.1	d1ct9a1	family	GO:0016879	ligase activity, forming carbon-nitrogen bonds
Glutamine 5-phospho-ribosyl-1-pyrophosphate (PRPP) amidotransferase 1ecc 2.4.2.14	Xanthine-guanine PRTase 1nul 2.4.2.22	d1ecca1	family	GO:0016763	transferase activity, transferring pentosyl groups
	Putative glutamine amidotransferase 1te5	?			
NS3 protease 1cu1 3.4.21.-	Trypsin 1 1trn 3.4.21.4	d1cu1a1	superfamily	GO:0008236	serine-type peptidase activity
	Guanylate kinase 1ex7 2.7.4.8	d1cu1a2	superfamily	GO:0005524	ATP binding
		d1cu1a3	superfamily	GO:0005524	ATP binding
N-Acetylglucosamine-1-PO4 uridyltransferase (GlmU) 1hv9 2.3.1.157 and 2.7.7.23	UDP-N-acetylglucosamine acyltransferase 1lxa 2.3.1.129	d1hv9a1	superfamily	GO:0008415	acyltransferase activity
	Uridyl transferase 1jvd 2.7.7.23	d1hv9a2	family	GO:0016779	nucleotidyltransferase activity
DNA polymerase II 1d5a 2.7.7.7	RNase H 1ril 3.1.26.4	not found in scop2go			
	Bacteriophage T7 RNA Polymerase 1msw 2.7.7.6				
6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase 1bif 2.7.1.105 and 3.1.3.46	Phosphoribulokinase 1a7j 2.7.1.19	d1bif_1	superfamily	GO:0005524	ATP binding
	Prostatic acid phosphatase 1nd6 3.1.3.2	?			
NitFhit fusion protein 1ems 3.6.1.29	FHIT (fragile histidine triad protein) 1fit 3.6.1.29	?			
	N-carbamoyl-D-amino acid amidohydrolase 1uf5 3.5.1.77	d1emsa2	superfamily	GO:0016810	hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds
PI-SceI 1dfa	I-CreI 1bp7	d1dfaa2	superfamily	GO:0004519	endonuclease activity
		d1dfaa3	superfamily	GO:0004519	endonuclease activity
	GyrA intein am2	d1dfaa1	family	GO:0046961	hydrogen ion transporting ATPase activity, rotational mechanism
				GO:0046933	hydrogen ion transporting ATP synthase activity, rotational mechanism
SinR repressor 1b0n	Cro lambda repressor 5cro	d1b0na2	superfamily	GO:0043565	sequence-specific DNA binding
	SinI antirepressor 1b0n	?			
Alcohol dehydrogenase 1a71 1.1.1.1	Drosophila (short chain) alcohol dehydrogenase 1b16 1.1.1.1	?			
	Chaperonin-10 (GroES) 1aon				
Restriction endonuclease (FokI) 1fok 3.1.21.4	EcoRV 1rva 3.1.21.4	d1foka4	superfamily	GO:0009036	Type II site-specific deoxyribonuclease activity
	MARR antibiotic-resistance repressor 1jgs				

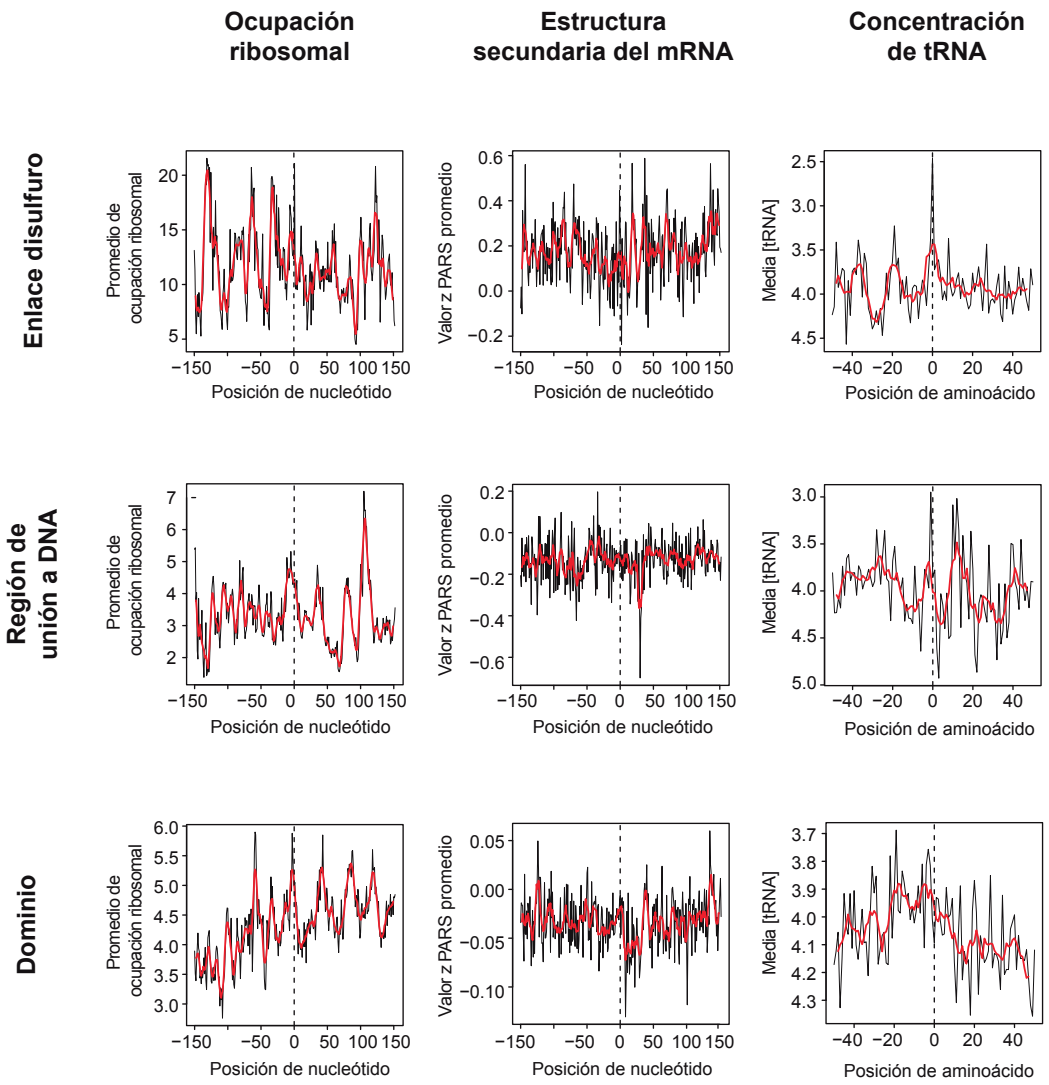
Multidomain Protein	One-Domain Protein(s)	domain in scop2go	domain matching	GO term	GO description
Copper chaperone of superoxide dismutase (CCS) 1qup	Hah1 Metallochaperone 1fe0	d1qupa2	family	GO:0008324	cation transmembrane transporter activity
	Cu, Zn superoxide dismutase (SOD) 1cbj	d1qupa1	family	GO:0004785	copper, zinc superoxide dismutase activity
Iron protein from quinol-fumarate reductase. (Binds iron clusters in the fumarate reductase complex) 1fum 1.3.99.1	[2Fe-2S] ferredoxin 1czp				
	a.1.2				
Flavocytochrome b2 1fcb 1.1.2.3	Glycolate oxidase 1al8 1.1.3.15	d1fcba1	family	GO:0046914	transition metal ion binding
				GO:0004460	L-lactate dehydrogenase (cytochrome) activity
	Cytochrome b5 1cyo	d1fcba2	family	GO:0020037	heme binding
Ribosomal protein L2 1ffk	Ribosomal protein L24 1jj2				
	Single-stranded DNA-binding protein 1eyg				
GreA transcript cleavage factor 1grj	FK-506-binding protein (FKBP12) 1fkh 5.2.1.8	?			
	a.2.1				
Formate dehydrogenase H 1aa6 1.2.1.2	Pyruvoyl-dependent aspartate decarboxylase (ADC) 1aw8 4.1.1.11	?			
	c.81.1				
Allosteric threonine deaminase 1tdj 4.2.1.16	Threonine synthase 1e5x 4.2.3.1	d1tdj_1	family	GO:0030170	pyridoxal phosphate binding
	Putative glycine cleavage system repressor 1u8s	?			

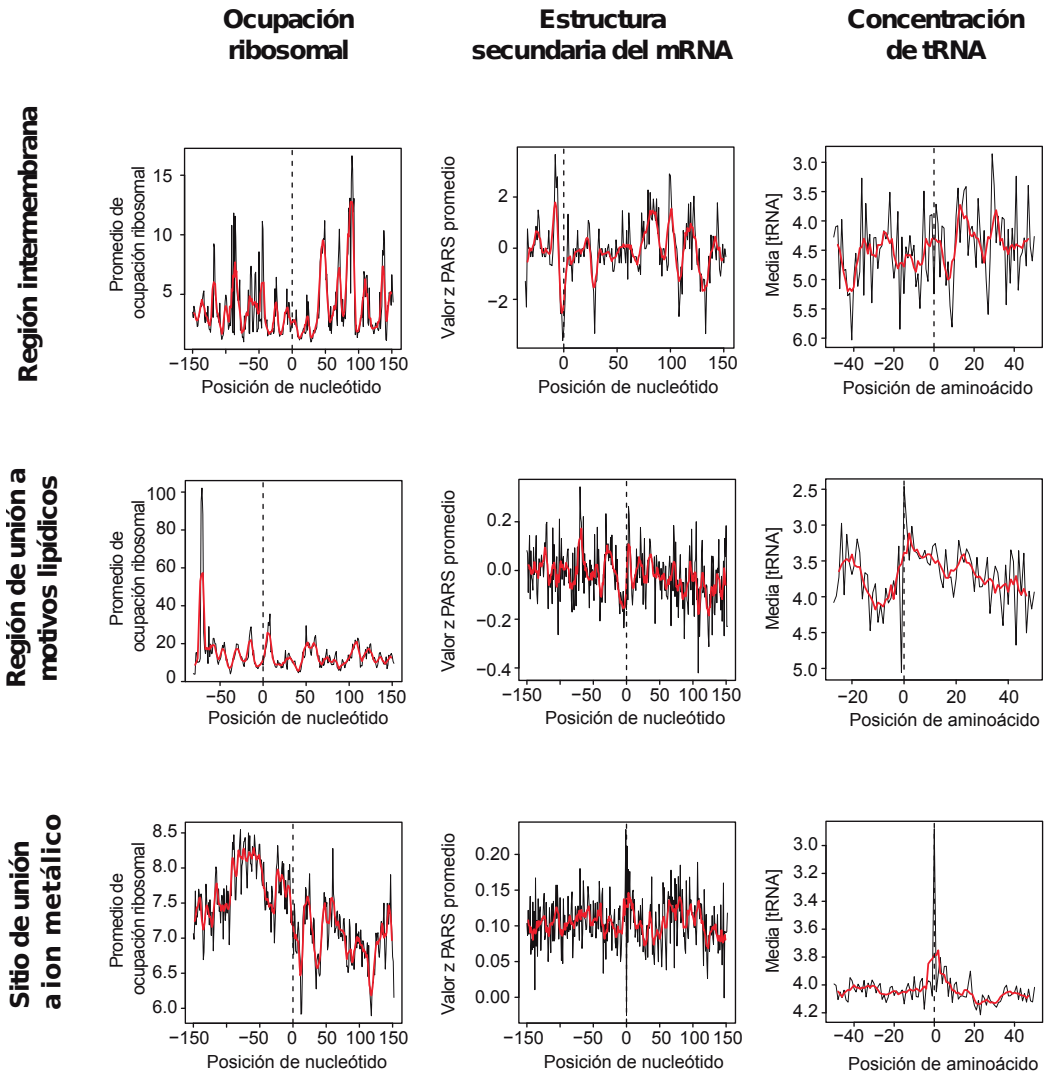
Apéndice C

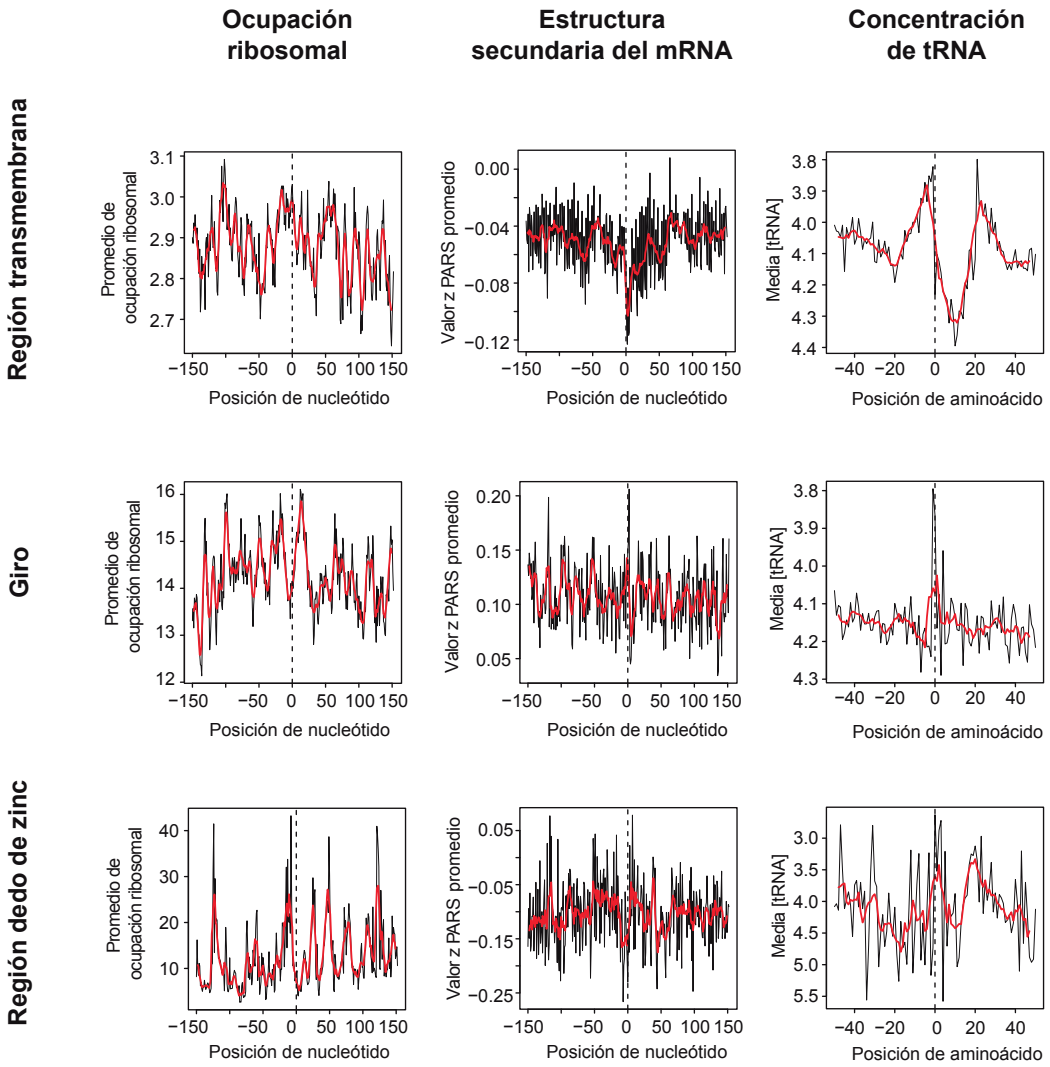
Patrones promedio para anotaciones funcionales y estructurales de proteínas

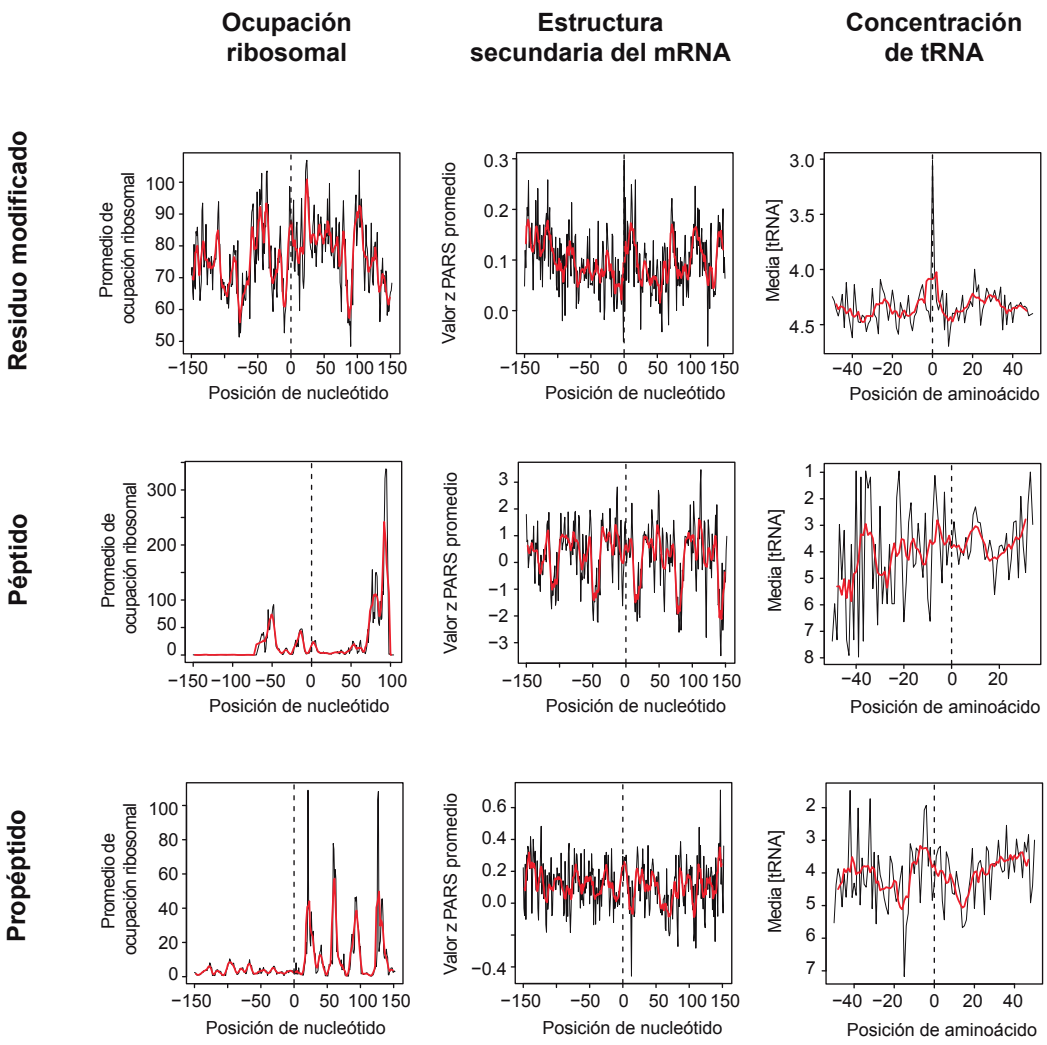

Ocupación ribosomal
Estructura secundaria del mRNA
Concentración de tRNA
Sitio de unión
Sitio de unión a calcio
Hélice superenrollada
Promedio de ocupación ribosomal
Valor z PARS promedio
Media [tRNA]
Posición de nucleótido
Posición de aminoácido

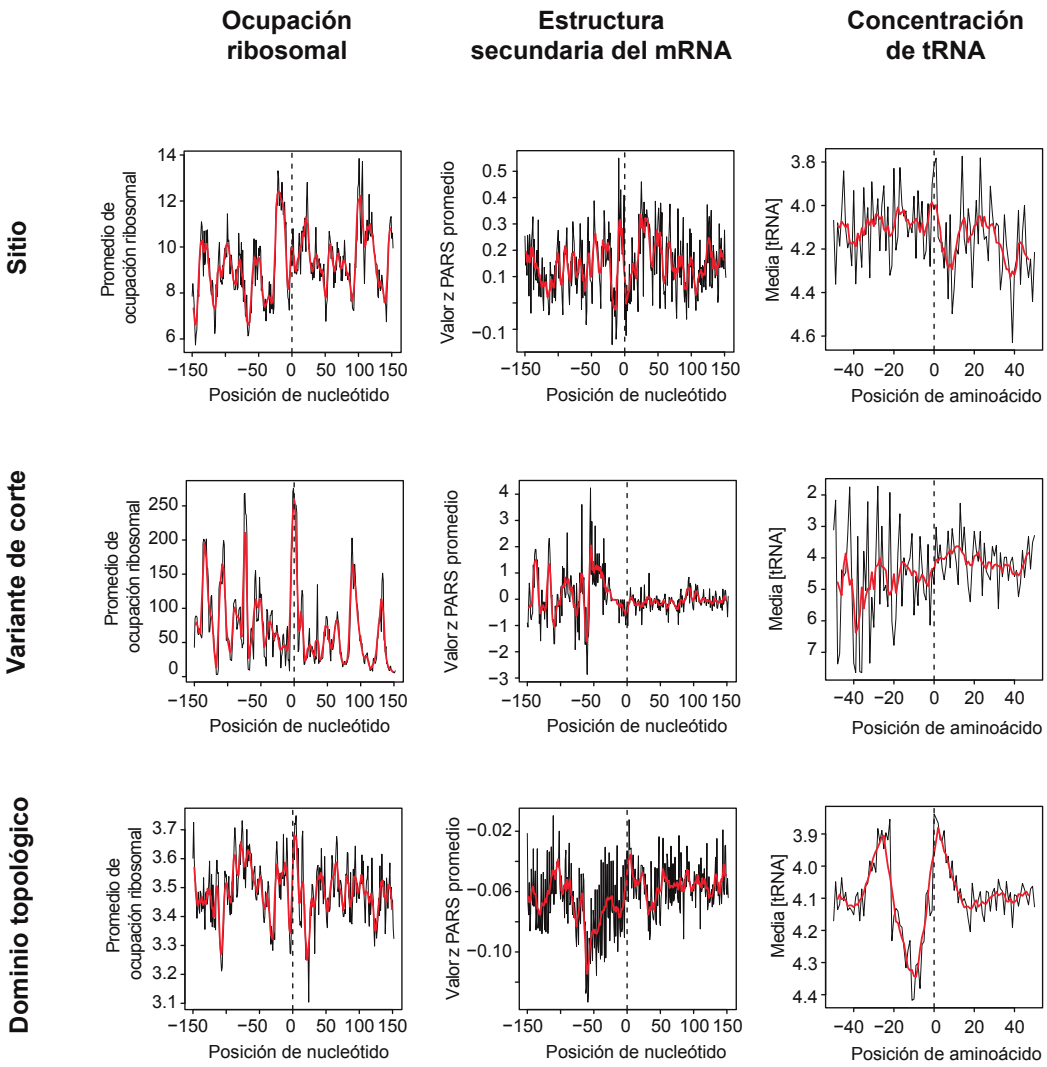

Ocupación ribosomal
Estructura secundaria del mRNA
Concentración de tRNA
Región de composición sesgada
Reticulación (cross-link)
IUPred
Promedio de ocupación ribosomal
Valor z PARS promedio
Media [tRNA]
Posición de nucleótido
Posición de aminoácido

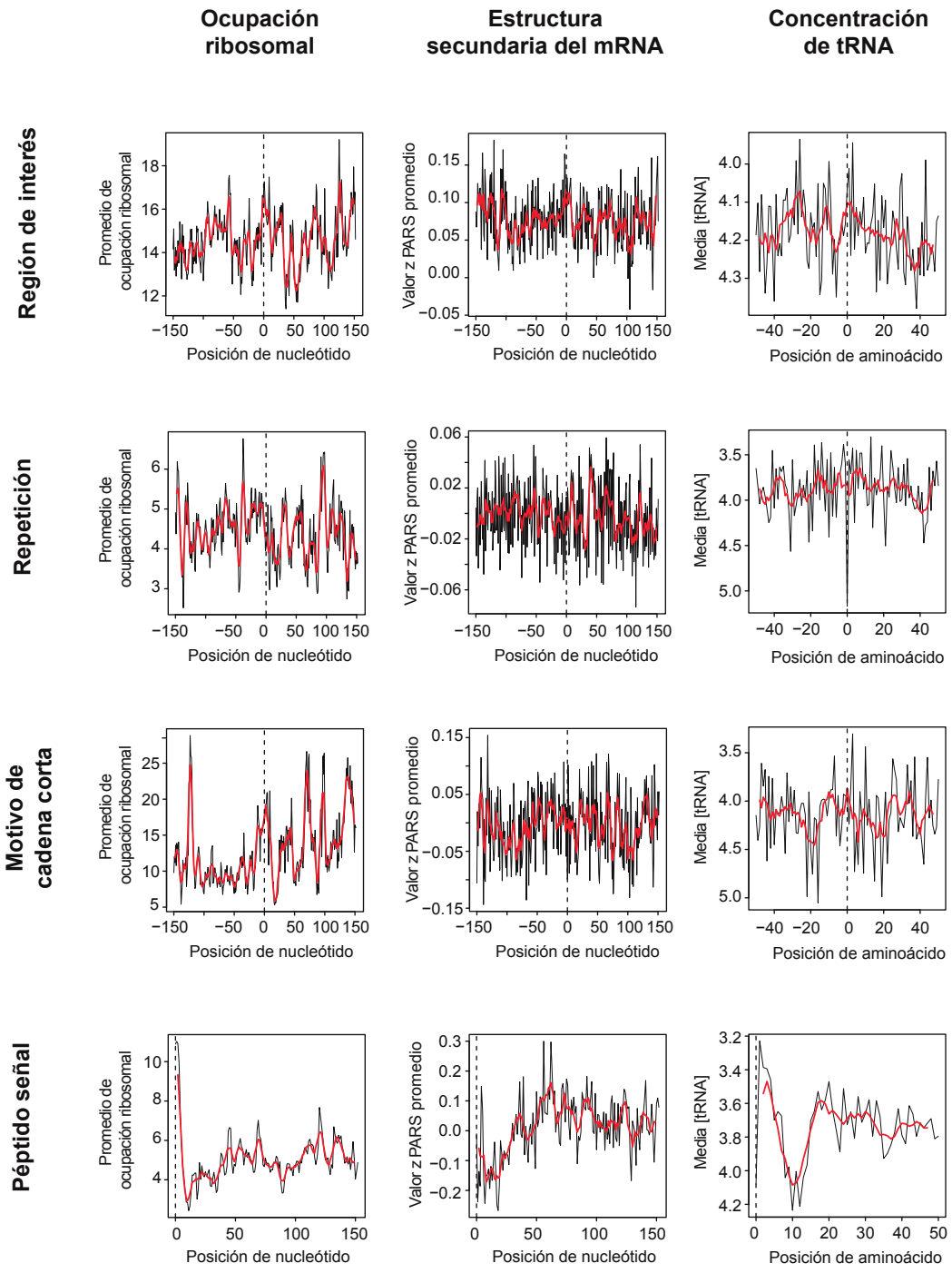

Ocupación ribosomal
Estructura secundaria del mRNA
Concentración de tRNA
Enlace disulfuro
Región de unión a DNA
Dominio
Promedio de ocupación ribosomal
Valor z PARS promedio
Media [tRNA]
Posición de nucleótido
Posición de aminoácido


Ocupación ribosomal
Estructura secundaria del mRNA
Concentración de tRNA
Región intermembrana
Región de unión a motivos lipídicos
Sitio de unión a ion metálico
Promedio de ocupación ribosomal
Valor z PARS promedio
Media [tRNA]
Posición de nucleótido
Posición de aminoácido


Ocupación ribosomal
Estructura secundaria del mRNA
Concentración de tRNA
Región transmembrana
Giro
Región dedo de zinc
Promedio de ocupación ribosomal
Valor z PARS promedio
Media [tRNA]
Posición de nucleótido
Posición de aminoácido


Ocupación ribosomal
Estructura secundaria del mRNA
Concentración de tRNA
Residuo modificado
Péptido
Propéptido
Promedio de ocupación ribosomal
Valor z PARS promedio
Media [tRNA]
Posición de nucleótido
Posición de aminoácido


Ocupación ribosomal
Estructura secundaria del mRNA
Concentración de tRNA
Sitio
Variante de corte
Dominio topológico
Promedio de ocupación ribosomal
Valor z PARS promedio
Media [tRNA]
Posición de nucleótido
Posición de aminoácido


Ocupación ribosomal
Estructura secundaria del mRNA
Concentración de tRNA
Región de interés
Repetición
Motivo de cadena corta
Péptido señal
Promedio de ocupación ribosomal
Valor z PARS promedio
Media [tRNA]
Posición de nucleótido
Posición de aminoácido

Apéndice D

Publicaciones

Este trabajo ha dado lugar a tres publicaciones y una revisión en revistas internacionales:

- **D. López**, F. Pazos. Gene ontology functional annotations at the structural domain level. *Proteins 2009*, **76**(3):598-607
- **D. López** and F. Pazos. Concomitant prediction of function and fold at the domain level with GO-based profiles. *BMC bioinformatics 2013*, **14**(Suppl 3):S12
- N. Pietrosemoli, **D. López**, A. Segura and F. Pazos. Computational prediction of important regions in protein sequences. *IEEE signal processing magazine 2012*, **29**(6):143-147
- **D. López** and F. Pazos. COPRED: Prediction of fold, GO molecular function and functional residues at the domain level. *Bioinformatics 2013*, doi:10.1093/bioinformatics/btt283

PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Gene ontology functional annotations at the structural domain level

Daniel Lopez and Florencio Pazos*

Computational Systems Biology Group, National Centre for Biotechnology (CNB-CSIC), Madrid 28049, Spain

ABSTRACT

Most proteins are organized in domains which can be seen as independent modular units in terms of molecular function (MF). Nevertheless, current functional annotations are done on a "whole-chain" basis without associating specific functions to the individual domains. We present here an automatic method for discerning which particular structural domain within a protein is responsible for a given MF originally attributed to the whole protein. By annotating the SCOP structural domains with gene ontology terms using this method, we obtained the first large-scale functional annotation at the domain level. We performed a large-scale comparison of these annotations with the ones implicit in the functional annotations of Interpro signatures, showing that the performance of this method is globally better. We also discuss in detail some particular examples. Generated automatically and available online, this resource could be the basis for future manually curated annotations.

Proteins 2009; 76:598–607.

Key words: SCOP; gene ontology; functional annotation; protein domain; protein function.

INTRODUCTION

Domains are the structural, evolutionary, and functional units of proteins. Multidomain proteins are very abundant, especially in eukaryotic organisms where they can make up 66–80% of the proteome.[1] In many cases the molecular functions (MFs) of the domains of a given protein are highly independent, to the extreme that these domains appear as independent polypeptides in other organism(s).[2] For example, the two domains of *Aspergillus* tryptophan synthase perform two different and independent catalytic activities, whilst the corresponding ortholog domains in *Escherichia coli* appear as two separate protein chains (tryptophan synthases α and β). Even in less extreme cases where the domains do not appear in isolation, particular MFs can be differentially assigned to them, that is, an SH2 domain for protein interaction, a GTPase domain for hydrolyzing GTP, and so forth. In these cases, additional "higher order" MFs appear as a result of the combination of the MFs of the domains.[3]

Protein function is a very complex phenomenon which requires intricate multidimensional ontologies to be represented. gene ontology (GO) has become the *de facto* standard for representing protein functions.[4] GO defines a controlled vocabulary composed of a set of terms that describe different functional aspects of proteins. These terms are related by parenthood relationships forming a directed acyclic graph (DAG). The DAG can be navigated from the very general (i.e., "enzyme activity") to the highly specific functions (i.e., "2-furoate-CoA ligase activity"). In fact, GO defines three independent ontologies (characterized by their corresponding DAGs) to represent three orthogonal aspects of the complex phenomenon of protein function: "MF," "biological process" (BP), and "cellular component." A gene product (protein) is functionally annotated by assigning to it a set of GO terms coming from these three ontologies.

Although domains are, in many cases, independent in functional terms, proteins are normally annotated on a whole-chain basis, without reference to the particular domain responsible for a given MF. This is acceptable for many applications, but it can sometimes cause serious problems, such as the automatic transfer of specific functions between multidomain proteins. It also complicates tasks which might look trivial, such as simply obtaining the set of structural domains associated with a given GO term in order to study structure–function relationships. Even in domain-oriented databases such as Pfam[5] or Interpro,[6] the GO annotations are not intended to be interpreted in terms of physical domains, although it may appear so at first sight. These resources use a consensus procedure to transfer to an entry (domain) the GO annotations of the proteins to which it is linked. This procedure leads to many entries being associated with functions that are not physically located in the segment of the protein the domain is representing. These annotations, wrongly associated with a domain, generally come

Additional Supporting Information may be found in the online version of this article.
Contract grant sponsor: Spanish Ministry for Education and Science; Contract grant numbers: BIO2006-15318, PIE 200620I240
*Correspondence to: Florencio Pazos, Computational Systems Biology Group, National Centre for Biotechnology (CNB-CSIC), Madrid 28049, Spain. E-mail: pazos@cnb.csic.es.
Received 5 August 2008; Revised 9 December 2008; Accepted 11 December 2008
Published online 20 January 2009 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.22373

from those which occur together with it in different proteins. Functional annotations at the domain level could also be inferred from resources such as Superfamily[7] which contain semimanual annotations for SCOP superfamilies that could be transferred to their domain members. Nevertheless, these resources can only work with coarse-grained functional vocabularies (general functional classes mixing BP and MF) since they have to ensure that all members of the superfamily have the same function, which would not be the case for detailed MFs (i.e., "GTP-" and "ATP-binding" proteins may belong to the same superfamily). On the other hand, these general function classes make sense when one appreciates that the main goal of this resource is to predict function for new sequences by assigning them to superfamilies.

Hence, despite being recognized as a key resource,[8] a dedicated database with explicit functional annotations at the domain level using detailed functional vocabularies is still lacking.

In this work, we describe a fully-automatic procedure for distinguishing which structural domain within a chain is responsible for a given GO "MF" term (GO-MF) originally associated with the whole chain. We carried out a large-scale evaluation of the accuracy of the method for some GO-MF terms for which the assignment to structural domains can be inferred from other data sources. We also compared that accuracy with the one obtained by transferring the GO annotations of Interpro entries to the corresponding structural domains and showed that the method presented here performs better. Additionally, some examples of GO-MF terms that would be wrongly associated with physical domains in Pfam/Interpro or by prediction methods "trained" in annotations done in a whole-chain basis are discussed in detail. We show how these cases are correctly associated by this method. We applied the method to the GOA-PDB[9] whole-chain-based annotation of protein structures, generating the scop2go domain-based annotation. This resource is available on-line through a web server.

MATERIALS AND METHODS

Scop2go method

The scop2go method for taking functional annotations to the structural domain level is based on the fold distribution of the set of PDB chains associated with a given GO-MF term.

We used the functional annotation for PDB chains provided by the GOA consortium,[9] which is done on a whole-chain basis. For a given GO-MF term, initially a matrix of PDB chains × SCOP folds for the set of chains associated with that term is generated. This matrix indicates which fold(s) a given chain possesses (see Fig. 1). The method has two assumptions: (i) the most populated fold in this matrix (the one present in the largest number of chains) corresponds to the structural domains of these chains which are responsible for that function; and (ii) only one domain within a chain is responsible for a "specific" (see below) GO-MF term, unless other domain(s) within that chain has (have) the same or a very similar fold distribution.

Once the most populated fold is located in the matrix, the corresponding domains in all the chains possessing this fold are associated with the GO-MF term. The other structural domains within these chains are "marked" as nonassociated with that term. The second most populated fold is located but excluding the domains marked in the previous step for the counts. If two folds have a highly similar chain distribution, both are associated with the term. Two folds are considered to have a similar chain distribution when 97% of the chains containing the fold with the highest frequency (in a given iteration of the procedure) also contain the other fold. One can then detect cases where more than one domain is required for a given function (i.e., triangle–rhombus in Fig. 1). Using 97% instead of 100% means one can detect cases where a domain is missing in some PDB structures due to fragment crystallization.

This process is iterated until no more folds remain (see Fig. 1). The whole process can be regarded as searching for the minimum set of folds that covers all the chains, that is, the minimum set of domains necessary to explain the (observed) fact that all these chains have that function.

For each fold, we calculated a *P*-value based on a hypergeometric distribution:

$$P(k, N, m, n) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}$$

where N is the total number of chains in SCOP (constant); m is the number of chains in SCOP associated with the fold (having some domain belonging to that fold); n is the number of chains annotated with the function; and K is the number of chains associated with the fold annotated with the function.

For example, for the "GTP-binding" GO-MF term (GO:0005525) there are 24 different SCOP folds in the 523 PDB chains annotated with this term in GOA-PDB. The most populated fold is the "P-loop" (c.37 -SCOP fold nomenclature-; *P*-value = 0.0). After associating the domains having this fold with the GO:0005525 term, other domains of these chains are marked as nonassociated. The excluded domains comprise, for example, domains of elongation factors not involved in GTP-binding (folds b.43, b.44) and domains inserted in the GTP-binding domains of Gα proteins (a.66). The fold with the next highest frequency (after excluding those linked to c.37) is the pair c.32–d.79, which corresponds to tubu-

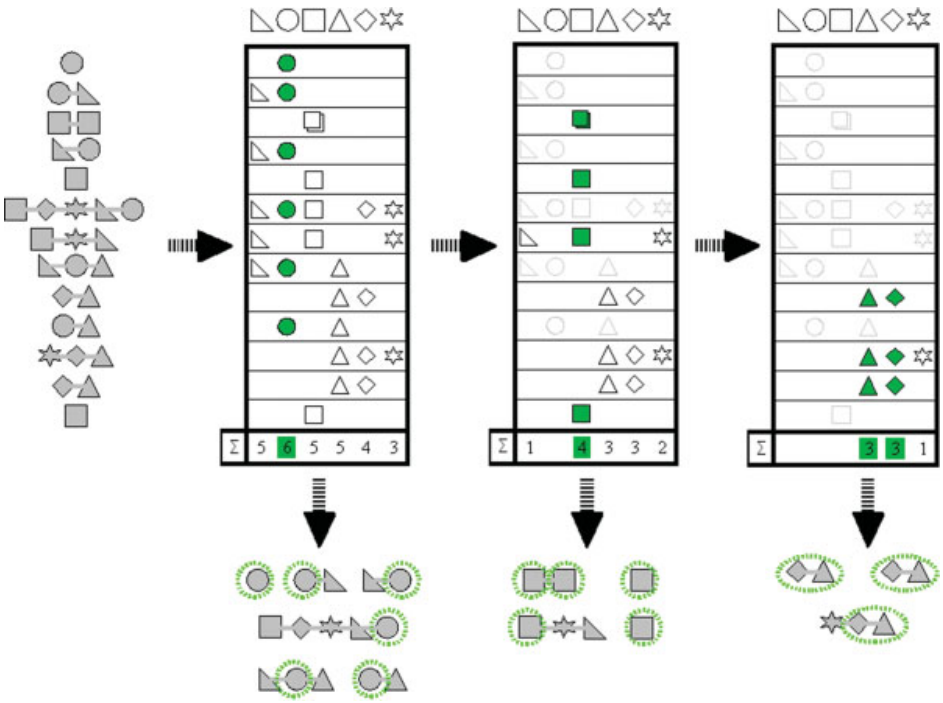


Figure 1

Schema of the scop2go method. The starting point is the whole set of protein chains annotated (in a whole chain-basis) with a given GO-MF term (left). The shapes represent different folds, as annotated in SCOP. A matrix of protein chains against SCOP folds is constructed. The predominant fold in this matrix is located (circle). The GO-MF term is assigned to the domains of the proteins with this predominant fold. The other domains of these proteins are marked as nonrelated to that term and excluded in the forthcoming counting. The next fold with the highest frequency is located (square), and the whole process is repeated. When two folds have the same (or very similar) frequency, both are considered responsible for that function (triangle–rhombus—right).

lin/FtsZ (*P*-value: 2.8E-39). The GTP-binding domain of tubulin/FtsZ is the c.32. Nevertheless, these two domains are tightly associated and c.32 never appears, either alone or as a GTP-binding module of a domain other than d.79 (see Results section).

At this point, all the structural domains are annotated with a given GO-MF term originally associated with the whole chains. The whole process is repeated for the other GO-MF terms and the annotations are accumulated in the domains.

It is important to remark that this method does not assume that a given fold only hosts one function, or that a given function is hosted in only one fold.

Additionally, the method is not affected by over-representation of proteins. It is easy to follow the example in Figure 1 "replicating" any of these chains many times (i.e., an important protein crystallized many times in different forms) to ascertain that the result is the same.

The assumption that only one domain is responsible for a given GO-MF (unless more domains have folds with a similar distribution) holds true for the highly specific MFs, but not necessarily for the general ones. For this reason, this procedure is only applied to GO-MF terms at a distance of 2 or higher from the root of the MF DAG, as an approximation of the specific ones. When all the structural domains associated with this set of specific functions are obtained, their annotations are expanded with less specific functions (ancestors) by following the GO hierarchy.

Generation of functional annotations at the domain level from interpro2go

The annotations obtained with this method are compared with the ones implicit in the Interpro2go annotation, although, strictly speaking, the objective of Interpro2go is not to "physically" locate the domains responsible for a given GO term (see Results section).

Interpro2go associates GO terms with Interpro entries (A.K.A. signatures or domains) whilst Interpro entries are associated with one or more SCOP families. We first obtain all the Interpro entries associated with a given GO-MF term (ATP, GTP, and heme-binding, in this case – see the next point). Following the association between Interpro entries and SCOP families, all the SCOP structural domains belonging to these families are

associated with the original GO-MF term. Redundancy is removed from this list because different Interpro entries (associated with the same GO-MF term) can be linked to the same SCOP family, members of which would consequently appear more than once. In this way we end up with a list of SCOP domains (single or groups of domains) associated with GO-MF terms equivalent to the one of scop2go. The two equivalent lists can be evaluated by the same procedure (next section).

Large scale semiautomatic evaluation of "binding" GO-MF terms

This is the first large-scale functional annotation of protein domains with GO-MF terms. There are no similar resources or gold standards with which to make a comparison. Apart from presenting the detailed results for some examples (see Results section) we did our best to perform a large-scale evaluation of some functions whose assignment to domains can be evaluated from other independent sources (binding of prosthetic groups associated with these functions).

Here we describe the semiautomatic procedure for evaluating the accuracy of the functional annotations of structural domains for GO-MF terms related to the binding of prosthetic groups (see Fig. 2). In this work, we applied this procedure to three prosthetic groups: GTP, ATP, and heme. This procedure was applied both to the scop2go predictions and to the ones inferred from interpro2go (previous point).

For each prosthetic group, a list of PDB[10] heteroatom identifiers which represent variations of it (equivalent molecules) is generated. For example, for the GO-MF term "GTP-binding" (GO:0005525), the list would contain GTP and other natural or artificial variations of this nucleotide which bind with the same proteins (GDP, GNP, GXP, etc. in PDB nomenclature). This list is obtained by interrogating the "Superligands" server[11] for ligands in PDB structurally similar to the given one. "Superligands" evaluate the chemical similarity using the Tanimoto coefficient.[12] We considered those equivalent ligands with a Tanimoto similarity $\geq$95% with the starting one. For ATP, we manually include in the list AMP, whose similarity with ATP is below the threshold due to the relatively large structural difference (two phosphate groups). Supporting Information (Additional File 1) contains these lists of similar ligands for ATP, GTP, and heme. For a given structural domain predicted to have a given binding function, all the other entries with the same "protein domain" annotation are retrieved from SCOP, with the idea of including different structures of the same protein (i.e., crystallized in different conditions, with different ligands, mutants, orthologs in different species, etc.). This is carried out because the particular structural domain being asessed may be in its *apo* form (without ligand) in that particular PDB (see Fig. 2), and would produce a false negative. Then, we retrieve from MSD[13] the list of residues contacting our ligand (or any of its structural neighbors–see above) in these PDB structures. Finally, we calculate the fraction of these residues which are within the structural domain we predicted as associated with that binding function (overlap). If our predicted domain (in any of its crystallized forms) is not in contact with that ligand (or a structural variation of it), this fraction would be zero or close to it (see Fig. 2). This continuous overlap value can be used as such as a performance figure, or it can be made discrete by imposing a cut-off (i.e., considering correct predictions to be those with overlap $\geq$70%, otherwise they are incorrect).

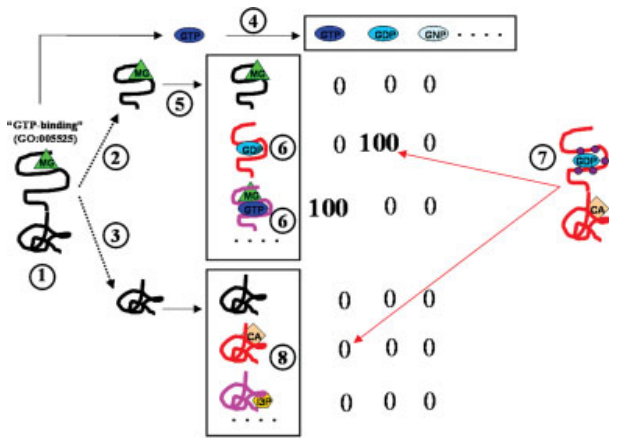


Figure 2

Illustrations of the procedure followed for the automatic evaluation of binding-related functional assignments. GO-MF terms related to binding functions can be assessed by looking for structures with the corresponding ligands bound to them. A protein chain with two domains is annotated (on a whole-chain basis) with a binding-related GO-MF term ("GTP-binding," in this example) (1). Scop2go is used to predict which domain within the chain is responsible for that function. It could point to the right domain (2) or to a wrong one (3). The first step is to retrieve a list of prosthetic groups which can be representing that binding function (4). Since the predicted domain could correspond to a particular instance of the protein for which the structure was determined in absence of ligand (i.e., a mutant unable to bind to it), all the SCOP entries in the same "domain" category are retrieved to have a representation of all the forms of that domain available in PDB (5). If our predicted domain does bind that prosthetic group, one or more of these SCOP domains are probably binding some of the (similar) prosthetic groups representing that function (6). The percentage of residues contacting the prosthetic group (as annotated in MSD (7)) that are within our predicted domain is calculated. For a wrong prediction (3), none of the domains would be contacting these prosthetic groups (8).

RESULTS AND DISCUSSION

Since this is, to our knowledge, the first global annotation of SCOP[14] structural domains with GO-MF terms, it has no "gold standard" to which it can be compared for global accuracy. Actually, the long-term goal of this

work is to generate such a resource. Nevertheless, for some particular GO-MF terms, such as the ones related to binding of prosthetic groups, their association with particular structural domains can be indirectly evaluated in a relatively automatic way. In the first part we present global accuracy values for some of these "binding" functions. In the second part we compare that accuracy with the one obtained following a similar procedure with the interpro2go annotations. Finally, in the third part some examples of annotations of SCOP domains are discussed in detail and compared with the corresponding Pfam/Interpro annotations, as well as with those obtained by prediction methods trained with whole-chain annotations, in order to illustrate the novelties of this approach.

Accuracy of the annotations for some "binding" GO-MF terms

The annotation of a SCOP structural domain with a binding-related function can be semiautomatically evaluated by assessing whether the position of the "binder" in the structure of the protein corresponds to that particular domain. Three binding-related functions, "GTP-binding," "ATP-binding," and "heme-binding", have been used. For each structural domain predicted to be involved in one of these binding functions, the percentage of the annotated binding residues falling within that domain (the overlap) was obtained, as detailed in Methods section.

Table I summarizes the results. Starting with the whole-chain annotations of GOA-PDB, the method automatically assigns the "GTP-binding" function to 524 structural domains, "ATP-binding" to 2962 and "heme-binding" to 1945. Most of the predicted domains have an overlap >70% with annotated GTP, heme, and ATP-binding sites (92.8, 94.1, and 68.5%, respectively). Although reasonably good (especially for GTP and heme), these results contain some "false negatives": these are domains predicted to bind the expected prosthetic group (correct predictions), but which are not seen as positives by the automatic protocol (see Methods section). A detailed inspection of the results shows that there are many reasons for these negatives. Some of them come from the lack of representative structures complexed with that particular ligand in the version of SCOP we have used (1.71). Newer versions of SCOP, however, do contain structures with that particular prosthetic group. In other cases, the structures are complexed with a ligand which is clearly equivalent to the one of interest, but has a Tanimoto coefficient below our threshold (95%). For example, "siroheme," "heme-C," and "heme-D" fall below this threshold when compared with "heme." We manually inspected the results looking for false negatives owing to these and other reasons and reassessed the accuracy taking them into account (Table I). Supporting Information (Additional file 1) contains information on these manually assessed domains along with notes giving the reasons. The accuracy rises to 95.2, 96.4, and 82.4% (for GTP, heme, and ATP, respectively) when these cases are taken into account.

There are also some "false positives" in this dataset that are due to errors in the original GOA annotation for PDB chains. For example, the GroES cochaperone is wrongly annotated as ATP-binding in GOA, possibly due to its close interplay with GroEL, a chaperone whose functional cycle is driven by ATP-binding/hydrolysis. More than errors, these cases can be considered "complex-based" annotations; that is, transferring to all the members of a protein complex the MF(s) of one of the members. These cases (not discarded in the manual assessment mentioned above) artificially reduce the accuracy reported above. For example, there are 188 GroES and similar proteins (that probably suffered the same interaction-based GO-MF annotation) which are clearly contributing to the relatively low accuracy for ATP-binding (82.4%) compared with GTP and heme (95.2 and 96.4%, respectively).

Finally, we used these datasets to evaluate the power of the *P*-values discriminating correct from incorrect assignments. The lists of domains were sorted in descending order by the logarithm of the corresponding *P*-values, and an ROC analysis[15] carried out with the lists containing correct and incorrect assignments. The "area under the ROC curve" (AUC) value quantifies the capacity of a given score (the *P*-value in this case) separating correct from incorrect predictions. It ranges from 1.0 (perfect

Table I
Overlap Between Domains Predicted to have a "Binding" Function and Annotated Ligand-Contacting Sites

GO-MF ``Binding'' term	Domains	Folds	Average overlap	Binding domains (overlap >70%)	Binding domains (including manually assessed)	P-values AUC
GTP	524	11	92.00%	486 (92.75%)	499 (95.23%)	0.991
HEME	1945	15	94.00%	1831 (94.14%)	1875 (96.40%)	0.934
ATP	2962	64	67.00%	2028 (68.47%)	2442 (82.44%)	0.988

For each GO-MF term, the columns show the number of structural domains predicted to have that function, the number of different SCOP folds they cover, the average overlap with binding sites, the number of domains which are actually binding that prosthetic group (overlap higher than 70% with the annotated binding site), the number of overlapping domains after a manual assessment (see Results section), and finally the "area under the ROC curve" as a measure of the discriminative power of the *P*-values separating good from bad assignments. See Supporting Information (Additional File 1) for additional data about these results.

discrimination) to 0.5 (random discrimination: good and bad cases uniformly distributed through the sorted list). Table I contains the AUC values for GTP-, ATP-, and heme-binding. These values, being very close to 1.0, show that the *P*-value can be used as an indicator of confidence in the predictions, since high (bad) values tend to be associated with wrong assignments, and the other way around. Nevertheless, these *P*-values have to be used with caution since for any GO annotation we can be sure about the "positives" but not about the "negatives" (proteins lacking a given annotation). The fact that a given chain is not associated with a GO term does not necessarily mean that the protein lacks that function. It could be either not known or simply not targeted by annotators. This fact could affect the frequencies used in the calculation of the *P*-values (see Methods section).

In summary, these results show that the method can identify the structural domain responsible for these binding functions in most cases, and that the *P*-values associated with the assignments give a measure of confidence [Supporting Information (Additional file 1) contains more data on these results].

The performance figures obtained for this set of thousands of domains with three predicted functions related to the binding of small compounds cannot be taken as the global performance of the method. Nevertheless there is no reason to think that the performance of the method for other specific functions not related to binding of prosthetic groups would be markedly different (i.e., neither better nor worse).

Comparison with interpro2go annotations

The goal of the interpro2go annotation[9] is not specifically to assign the GO-MF terms to the segment of the proteins represented by a given Interpro entry (see below). However, we generated such assignments in order to compare them with the ones generated by the method presented here, the goal being to map the functions physically.

For the same three GO terms related to binding, we obtain all SCOP structural domains that would be associated with them following the GO annotations of Interpro domains[9] and the association of the latest with SCOP structural domains. These domain functional assignments were assessed by the same procedure used in the previous section, based on the binding of prosthetic groups (see "Materials and Methods" for a more complete description of the procedure).

Figure 3 shows the comparison of the accuracies obtained for the scop2go and interpro2go domain annotations for the GTP, ATP, and heme-binding MF-GO terms. For scop2go, the original accuracies obtained with the fully automatic evaluation are given before the manual inspection of some false negatives (see previous point). In both the cases, a prediction was deemed to be correct if the overlap was 70% or higher.

While the accuracy for heme-binding is virtually identical, scop2go clearly out performs interpro2go for ATP and GTP-binding. This is probably due to the fact that many "wrong" annotations in Interpro are due to the copresence of domains in the same protein (see the elongation factor example given the next point). Many heme-binding domains are isolated and hence they do not lead to this problem. Actually, for this kind of mono-domain proteins, we do not need any method for transferring functional annotations from the chain to the domain level.

The way in which we evaluate these predictions gives some advantage to predictions involving more than one domain, since we evaluate how well the binding site is covered by the predicted domain/s (sensitivity) but not whether other domains are "over-predicted" (specificity). In the limit, predicting all the domains in the chain as responsible for the function would lead to perfect predictions since they will necessarily cover the binding site. While poly-domain predictions in scop2go are rare and normally only one domain is associated with the function, interpro2go leads to many functions associated with more than one structural domain. This is due to the many Interpro signatures covering more than one structural domain, that is, in the elongation factor example (next point) there is one Interpro entry (domain) for each of the three structural domains (see Fig. 4), and an additional Interpro entry covering the whole protein. This tendency to poly-domain predictions gives some advantage to interpro2go, since these predictions have a greater chance of being correct. To assess how this affects the results, we evaluate the accuracy of scop2go and interpro2go for the mono-domain predictions alone [Fig. 3(B)]. While the accuracy of scop2go remains virtually the same, the accuracy of interpro2go drops considerably.

It is important to bear in mind that the goal of Interpro GO annotations is not necessarily to locate functions in physical domains of proteins. The fact that a given Interpro entry (domain) is associated with a GO term should not be interpreted as the segments of the proteins corresponding to that domain being the ones responsible for that function. GO annotations for a given Interpro entry are obtained by a consensus procedure from the ones of the sequences linked to that entry. This method is designed so that any new protein matching the sequence signature of that Interpro entry can be confidently associated with the GO terms linked to that entry, regardless of whether or not the sequence segments responsible for that function(s) coincide with the ones responsible for the match. This is well illustrated in the elongation factor example (next point). In this work, we "interpret" the interpro2go annotations in physical terms in order to compare with the method presented here, the goal being in fact to map MFs at the physical domain level. Such a comparison was done to show how important a resource

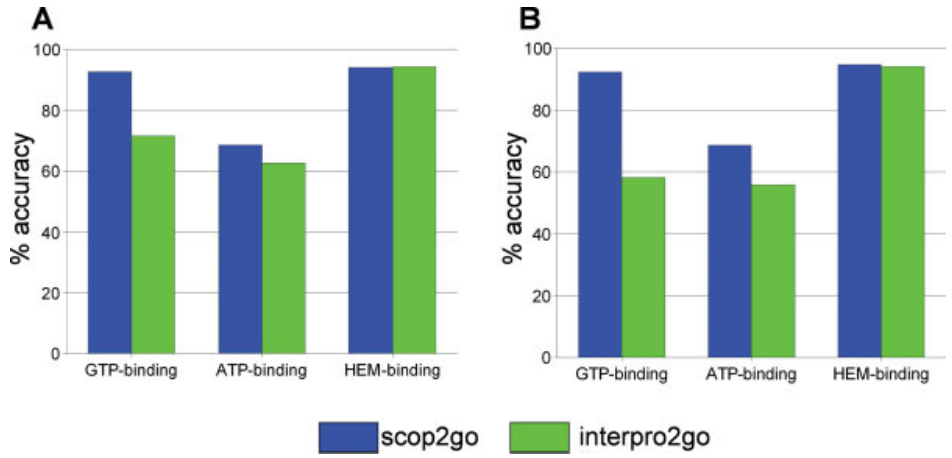


Figure 3

Performance comparison of scop2go and interpro2go for three GO-MF terms related to binding of prosthetic groups. The *y*-axis represents the percentage of domains predicted as responsible for a given binding function which are correct, that is, which are actually binding a prosthetic group associated with that function. (**A**) All the predictions. (**B**) Predictions comprising one domain only. Supporting Information (Additional file 1) contains more details on this datasets.

such as the one presented here can be, since "misinterpreting" other resources as functional annotations at the domain level would lead to wrong results.

Supporting Information (Additional file 1) contains detailed information on the interpro2go predictions and their comparison with scop2go. The next point discusses

Domain	Pfam/Interpro	scop2go	
d1jnyA3	(PF00009) GTP-binding	GTP-binding	elongation factor
d1jnyA1	(PF03144) GTP-binding		
d1jnyA2	(PF03143) GTP-binding		

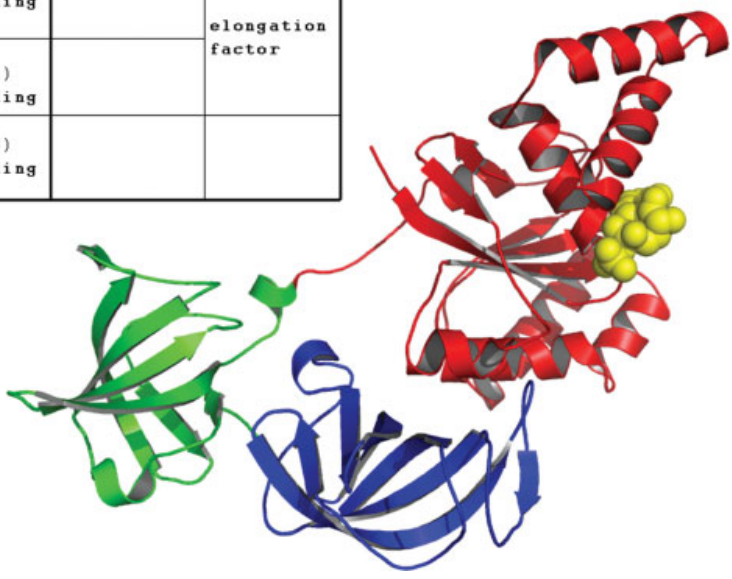


Figure 4

Comparing the Pfam/Interpro and scop2go annotations for the domains of an elongation factor. Pfam/Interpro and scop2go annotations for the three structural domains of the *Sulfolobus solfataricus* elongation factor 1α (PDB: 1jny_A). The GTP bound to the N-terminal (red) domain (SCOP:d1jnyA3) is shown in yellow. Figure generated with Pymol (http://pymol.sourceforge.net/)

in detail some examples with different predictions of scop2go and interpro2go.

Examples of functional annotations at the structural domain level and comparison with other resources

In this section, we discuss in detail some examples to illustrate the power and limitations of this approach and its differences with respect to existing annotations.

The first example is the *Sulfolobus solfataricus* Elongation Factor 1α (PDB:1jny, chain A). This protein has three structural and functional domains: the N-terminal GTP-binding domain, the middle domain involved in tRNA-binding, and the C-terminal domain involved in binding the tRNA and EF-Ts (see Fig. 4). This protein (the whole chain) is annotated with three GO-MF terms in GOA-PDB: "GTP-binding," "GTP-ase activity," and "elongation factor." In Pfam, all three domains signatures (PF00009, PF03143, and PF03144) are automatically annotated as "GTP-binding," in spite of the fact that the textual information available in the same Pfam entries clearly states their different roles. The corresponding Interpro entries share these "wrong" annotations with Pfam. Our method automatically recognizes that the "GTP-binding" function is associated with only the N-terminal domain (SCOP: d1jnya3) with a *P*-value of 0.0. "GTP-ase activity" is also associated with the Nt domain alone (*P*-value = 4.4E-135). The method also assigns the "translation elongation factor activity" to domains N-terminal and middle together (d1jnya3-d1jnya1; *P*-value = 2.2E-81). These two domains are present in some elongation factors without the C-terminal domain, indicating that they are the main ones responsible for this function.

As discussed in the previous point, the goal of Interpro2go is not to associate GO terms with physical segments of the protein, but to be sure that any new protein matching the sequence signature of the entry can be confidentially associated with the corresponding GO terms. In this sense, the results for these elongation factor domains (middle and Ct) are not wrong because a sequence matching these domains would be a elongation factor and hence it would bind GTP (although not specifically in these particular domains). We stress again that here we "interpret" Interpro2go in terms of physical domains to compare it with our method, and show that functional annotations at the domain level are required, since "misinterpreting" these annotations from existing resources can lead to wrong results.

In contrast, our method is not restricted to sequences close in the sequence-space. Since it considers all chains annotated as GTP-binding, regardless of their sequence similarities, it can easily discard these two β domains (see Fig. 4) because it recognizes that the c.37-fold can be present in many GTP-binding chains without them, and hence this structural domain alone must adequately explain the observed fact that these proteins are GTP-binding.

To illustrate other problems associated with whole-chain functional annotations, we submitted separately the sequence segments of these three domains to the GOtcha server.[16] This system predicts GO terms for an unannotated sequence based on the ones associated with its homologs. The server assigned the GTP-binding function to the sequences of the three domains with virtually identical scores [see Supporting Information (Additional file 1)], similarly to what happened with the annotations in Pfam/Interpro. GOtcha also associates "elongation factor activity" with the three domains indistinguishably. This shows that systems "trained" in whole-chain annotations cannot be used for domain-oriented functional predictions, and highlights the importance of developing a domain-oriented functional annotation.

For elongation factor Tu (PDB:1eft), our methods also correctly assigns the "GTP-binding" function to the N-terminal domain but, contrary to EF-1α, it does not associate "elongation factor". In this case, this is simply due to the protein (whole-chain) not being annotated as "elongation factor" in the original GOA-PDB. This highlights one important characteristic of this method – it can take the whole-chain annotations to the domain level, but it cannot "invent" or predict annotations "*de novo*".

The tyrosine protein kinase src (PDB:2src) comprises three domains: an N-terminal SH3, a middle SH2, and a C-terminal Tyr-kinase (Uniprot:P12931) [Fig. 5(A)]. It is annotated as "protein kinase activity," "protein-tyrosine kinase activity," and "ATP-binding" in GOA-PDB. Our method automatically assigns all these annotations to the C-terminal domain (*P*-values = 0.0 and 1.3E-162), discarding the involvement of the SH2 and SH3 modules in these functions. This is in agreement with the Interpro[6] domain assignment for this protein, which places the "prot kinase core" and the "tyr pkinase" domains (and hence ATP-binding) in the C-terminal domain [Fig. 5(A)].

For FtsZ (PDB:1w5b, chains A or B), the method transfers the "GTP-binding" annotation to both domains together (SCOP:d1w5a1-dw5a2; *P*-value = 2.8E-59). Only the N-terminal domain (d1w5a1) is annotated as GTP-binding in SCOP. These two domains are tightly associated [Fig. 5(B)], and d1w5a1 homologs never appear alone or as a GTP-binding module of another domain (except with d1w5a2 homologs). For this reason, the method considers these two domains as necessarily together for binding GTP, although strictly this has to be regarded as a false positive. This C-terminal domain of tubulin/FtsZ is also "incorrectly" associated with GTP-binding in Pfam/Interpro.

These cases are not anecdotic, since the global comparison between scop2go and interpro2go (previous point) shows that scop2go is better in assigning functions to structural domains.

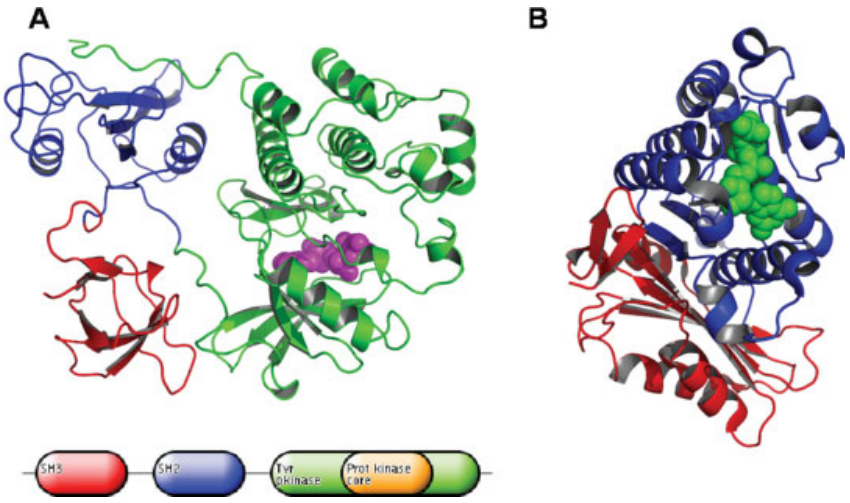


Figure 5

Other examples of functional annotations at the structural domain level. (**A**) Tyrosine-protein kinase Src. The figure shows the 3D structure of this protein (with the ATP analog in magenta) and its Interpro domain organization.[6] (**B**) FtsZ. The two structural domains, as described in SCOP, are marked. GTP is colored green. Figures generated with Pymol (http://pymol.sourceforge.net/).

Other examples of functional assignments at the domain level

As an additional dataset, we evaluated the functional assignments of our method for the 45 multidomain proteins discussed by Bashton and Chothia.[3] These authors undertook a comparative study of the functions of a set of multidomain proteins and the functions of the corresponding homologous mono-domain protein chains. They constitute a perfect dataset for the method presented here, since the accuracy of the functional assignments for the domains of the multidomain proteins can be manually and qualitatively assessed through the functions of the corresponding homologous mono-domain proteins. For example, the molybdate-dependent transcriptional regulator ModE (PDB: 1b9m) has three domains, one being the homolog of the MARR antibiotic-resistance repressor (PDB: 1jgs) and the other two of the molybdate/tungstate-binding protein II (PDB: 1gug). Our method assigns "transcription factor activity" (GO:0003700) to the first domain, and "molybdenum ion-binding" (GO:0030151) to the other two domains, in perfect agreement with the functions of the mono-domain homologs. For almost all domains in the Bashton and Chothia dataset annotated in scop2go, the functional assignment assessed using the monodomain proteins looks correct according to this qualitative assessment. The detailed results for the 45 proteins are available in Supporting Information (Additional File 1).

The scop2go web server

The scop2go website contains GO-MF functional annotations for 39804 SCOP domains (and groups of domains) belonging to 33584 PDB chains. These domains are annotated using 948 specific GO-MF terms. The ancestors of these specific terms are also included in the annotation, but marked as such. This resource is available at http://pdg.cnb.uam.es/scop2go. The user can interogate the server for a given PDB, SCOP or GO IDs, as well as for parts of them. In the results pages, the PDB ids, SCOP domain ids, and GO-MF term ids are active hyperlinks to the corresponding entries in pdbsum,[17] SCOP[14] and GO,[4] respectively. The complete annotation is available as a parseable file upon request.

CONCLUSIONS

We have developed an automatic method for taking functional annotations from the chain to the structural domain level. The method can discern which structural domain is responsible for a MF originally associated with the whole chain. This is the first approach intended to generate large-scale detailed functional annotations at the domain level in a fully automatic way.

This procedure produces better results than those obtained by mapping GO terms to SCOP domains via the interpro2go associations. Resources such as Interpro/Pfam or Superfamily could also be used for inferring functional annotations at the domain level, although that is not their main goal. These resources are intended for being used in function prediction: to annotate new proteins based on sequence matching with the signatures in these databases. With regard to Interpro, we show (both in a large scale test set and in detailed particular examples) how "interpreting" the interpro2go annotations in

terms of physical domains responsible for the GO terms can give wrong results. Contrary to these resources, the goal of scop2go is not to "predict" functions, but to take functional assignments from the chain to the domain level. Thus these resources complement each other.

We think that our automated procedure can be the basis of a future manually curated database of functional annotations at the structural domain level. Such a resource is clearly necessary for carrying on studies related to the complex phenomenon of protein MF.[8]

In the future we hope to extend this method to proteins without 3D structure, and also explore whether a similar method can be used to predict new functions.

ACKNOWLEDGMENTS

The authors acknowledge Alfonso Valencia, Angela Pozo, and David Juan (CNIO) for interesting and inspiring discussions. We also thank J. C. Triviño (CNB-CSIC) for help with the *P*-value analysis.

REFERENCES

1. Apic G, Gough J, Teichmann SA. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. J Mol Biol 2001;310:311–325.
2. Marcotte EM, Pellegrini M, Ho-Leung N, Rice DW, Yeates TO, Eisenberg D. Detecting protein function and protein-protein interactions from genome sequences. Science 1999;285:751–753.
3. Bashton M, Chothia C. The generation of new protein functions by the combination of domains. Structure 2007;15:85–99.
4. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, Richter J, Rubin GM, Blake JA, Bult C, Dolan M, Drabkin H, Eppig JT, Hill DP, Ni L, Ringwald M, Balakrishnan R, Cherry JM, Christie KR, Costanzo MC, Dwight SS, Engel S, Fisk DG, Hirschman JE, Hong EL, Nash RS, Sethuraman A, Theesfeld CL, Botstein D, Dolinski K, Feierbach B, Berardini T, Mundodi S, Rhee SY, Apweiler R, Barrell D, Camon E, Dimmer E, Lee V, Chisholm R, Gaudet P, Kibbe W, Kishore R, Schwarz EM, Sternberg P, Gwinn M, Hannick L, Wortman J, Berriman M, Wood V, de la Cruz N, Tonellato P, Jaiswal P, Seigfried T, White R. The gene ontology (GO) database and informatics resource. Nucleic Acids Res 2004;32 (Database issue):D258–D261.
5. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C, Eddy SR. The Pfam protein families database. Nucleic Acids Res 2004;32 (Database issue):D138–D141.
6. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Barrell D, Bateman A, Binns D, Biswas M, Bradley P, Bork P, Bucher P, Copley RR, Courcelle E, Das U, Durbin R, Falquet L, Fleischmann W, Griffiths-Jones S, Haft D, Harte N, Hulo N, Kahn D, Kanapin A, Krestyaninova M, Lopez R, Letunic I, Lonsdale D, Silventoinen V, Orchard SE, Pagni M, Peyruc D, Ponting CP, Selengut JD, Servant F, Sigrist CJ, Vaughan R, Zdobnov EM. The InterPro Database, 2003 brings increased coverage and new features. Nucleic Acids Res 2003;31:315–318.
7. Wilson D, Madera M, Vogel C, Chothia C, Gough J. The SUPERFAMILY database in 2007: families and functions. Nucleic Acids Res 2007;35 (Database issue):D308–313.
8. Riley M. Searchlight on domains. Structure 2007;15:1–2.
9. Camon E, Magrane M, Barrell D, Binns D, Fleischmann W, Kersey P, Mulder N, Oinn T, Maslen J, Cox A, Apweiler R. The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT. Tr EMBL, and Inter Pro. Genome Res 2003;13:662–672.
10. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
11. The Superligands Server. Available at http://bioinf.charite.de/superligands.
12. Holliday JD, Hu CY, Willett P. Grouping of coefficients for the calculation of inter-molecular similarity and dissimilarity using 2D fragment bit-strings. Comb Chem High Throughput Screen 2002;5:155–166.
13. Boutselakis H, Dimitropoulos D, Fillon J, Golovin A, Henrick K, Hussain A, Ionides J, John M, Keller PA, Krissinel E, McNeil P, Naim A, Newman R, Oldfield T, Pineda J, Rachedi A, Copeland J, Sitnov A, Sobhany S, Suarez-Uruena A, Swaminathan J, Tagari M, Tate J, Tromm S, Velankar S, Vranken W. E-MSD: the European Bioinformatics Institute Macromolecular Structure Database. Nucleic Acids Res 2003;31:458–462.
14. Andreeva A, Howorth D, Brenner SE, Hubbard TJ, Chothia C, Murzin AG. SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res 2004;32 (Database issue):D226–D229.
15. Wikipedia. ROC Analysis. Available at http://en.wikipedia.org/wiki/Receiver_operating_characteristic.
16. Martin DM, Berriman M, Barton GJ. GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes. BMC Bioinformatics 2004;5:178.
17. Laskowski RA, Chistyakov VV, Thornton JM. PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids. Nucleic Acids Res 2005;33:D266–D268.

Lopez and Pazos *BMC Bioinformatics* 2013, **14**(Suppl 3):S12
http://www.biomedcentral.com/1471-2105/14/S3/S12

BMC Bioinformatics

PROCEEDINGS Open Access

Concomitant prediction of function and fold at the domain level with GO-based profiles

Daniel Lopez, Florencio Pazos*

From Automated Function Prediction SIG 2011 featuring the CAFA Challenge: Critical Assessment of Function Annotations
Vienna, Austria. 15-16 July 2011

Abstract

Predicting the function of newly sequenced proteins is crucial due to the pace at which these raw sequences are being obtained. Almost all resources for predicting protein function assign functional terms to whole chains, and do not distinguish which particular domain is responsible for the allocated function. This is not a limitation of the methodologies themselves but it is due to the fact that in the databases of functional annotations these methods use for transferring functional terms to new proteins, these annotations are done on a whole-chain basis. Nevertheless, domains are the basic evolutionary and often functional units of proteins. In many cases, the domains of a protein chain have distinct molecular functions, independent from each other. For that reason resources with functional annotations at the domain level, as well as methodologies for predicting function for individual domains adapted to these resources are required.

We present a methodology for predicting the molecular function of individual domains, based on a previously developed database of functional annotations at the domain level. The approach, which we show outperforms a standard method based on sequence searches in assigning function, concomitantly predicts the structural fold of the domains and can give hints on the functionally important residues associated to the predicted function.

Background

Proteins are the key players of the cellular processes. Obtaining information on the structure, function and important residues for the protein repertory of a given organism (proteome) is crucial not only for getting insight into its biology, but also to foresee possible ways for modifying it in our benefit. Nevertheless, obtaining experimentally this kind of information is very slow and expensive. On the contrary, obtaining the raw sequences of complete proteomes or part of them is nowadays relatively fast and inexpensive, and this is getting even better with "next generation sequencing" technologies [1]. For these reasons, developing computational techniques for assigning structural and functional features to protein sequences is an active area of research.

Methods for predicting protein three-dimensional structure from sequence generally are based on the known relationship between sequence similarity and structural similarity. Most of these methods look for homologous proteins of known structure and model the problem sequence based on them. This search is either based on simple sequence matching methods for cases of close homology, or profile-based methods for remote homology.

Similarly, most methods for computationally assigning function to proteins ("annotation") are also based on the observed relationship between sequence similarity and functional similarity [2-4]. Functions of unknown proteins are inferred (transferred) from those of their homologs. This relationship between sequence similarity and functional similarity is far more complex than that between sequence and structure, in part due to the problem of precisely defining and quantifying "protein function" [5]. Contrary to what happens with protein structures, which can be univocally defined, quantified and compared, protein functions are more difficult to define. Many functional schemas and vocabularies co-existed in the past and still do, a lack of consensus which actually reflects this

* Correspondence: pazos@cnb.csic.es
Computational Systems Biology Group, National Centre for Biotechnology (CNB-CSIC), C/ Darwin 3, 28049 Madrid, Spain

problem of lacking a precise definition of the concept "protein function". The de-facto standard nowadays for representing protein function is that generated and maintained the Gene Ontology (GO) consortium [6]. GO defines a set of functional terms (vocabulary) related by parenthood relationships. These relationships form a partially hierarchical structure which can be navigated from terms representing very general to those representing highly specific functional aspects of proteins. Additionally, GO terms can be divided in three classes created to represent three independent aspects of the complex phenomenon of protein function: 'molecular function', 'biological process', and 'cellular component'. A given protein is annotated by assigning to it one or more terms from these three sets. In the following, we will focus on the 'molecular function' aspect of proteins (GO:MF) since that is the one used in this work.

The basic concepts and methodologies for transferring function from homologous sequences have evolved with time and, at the same time, adapted to these new structured vocabularies (for recent reviews describing in detail the field see [7-10]). The evolution consisted mainly of incorporating more sensitive methods, based on profiles, and phylogenetic approaches for locating distant homologs from which to transfer function. Some methods also consider the GO:MF functional terms associated with all the homologs and their underlying hierarchical relationships to come up with a final set of terms for the problem sequence [11-13]. Another tendency is to concentrate on motifs or groups of residues, defined based on sequence and/or structural criteria, indicative of function, instead of relying on global sequence similarities spread through the whole length of the protein [14-17].

Most of these methods, specially those based on global sequence matches against individual proteins or profiles, are intended to assign function at the whole chain level, without distinguishing which individual domain is associated to a given GO:MF term. In most cases, this is not a problem of the methodologies themselves but of the annotations contained in the resources they search against. In these resources, functions are associated to whole chains, not to particular protein domains, and as such they are transferred to the problem sequences. Nevertheless, domains are the structural, evolutionary, and often functional units of proteins. In many cases, individual molecular functions can be assigned to them. Even the functional annotations in domain-oriented databases such as Pfam or Intepro suffer from this problem when these annotations are interpreted in terms of physical domains [18].

In this work, we present a method for annotating proteins with GO:MF terms at the domain level. The method is based on matching against a library of "position specific scoring matrix" (PSSM) profiles [19] derived from structural alignments of domains annotated with the same GO:MF functional term. These annotations are taken from the first resource specifically devoted to assigning GO:MF terms at the domain level, SCOP2GO [18]. Since all the domains within a profile share the same fold, the method also implicitly assigns fold to the domains of the query proteins, although that is not its main goal. Moreover, the pattern of positional conservation within these profiles can give clues on the functional sites of the query sequence. We show that a psi-blast [19] search against this library of profiles renders better results than an equivalent search against a database containing the original sequences of the domains, demonstrating the added value of constructing the profiles guided by the GO:MF functional annotations at the domain level. So, this method allows to concomitantly obtain information on function, fold and functional sites at the domain level for unknown proteins.

Methods

Figure 1 illustrates the methodology used for building the library of profiles and searching against it, as well as the protocol used for benchmarking the method and comparing with psi-blast.

Library of GO:MF profiles at the domain level

The idea is that each entry in this library represents an alignment of all domain sequences known to have a given GO:MF function (a non-redundant representation of them, actually) that can be related in an alignment, i.e. belonging to the same fold and hence amenable of structural alignment.

The starting point is the SCOP2GO resource, which contains GO:MF annotations at the structural domain level [18]. SCOP2GO uses an automatic method for discerning which particular domain of a protein chain is responsible for a GO:MF annotation originally assigned to the chain as a whole. Starting with the fold distribution of all the chains associated to a given GO:MF term, the method looks for the minimum set of structural folds necessary for explaining the (observed) fact that all these chains have that function. The GO:MF term is assigned to the domains with these folds. The process is iterated for the other GO:MF terms and the annotations accumulated in the domains [18].

Multidomain entries, as well as those corresponding to PDB chains annotated as "mutant" and "circular permutation" are excluded. The resulting domains can be seen as arranged in a matrix of Fold X Function (GO:MF) (Figure 1). Folds are structural folds as defined in SCOP [20]. Each entry in this matrix (i.e. set of structural domains with the same fold and the same GO:MF term) is made non-redundant at 40% identity with T-coffee [21]. Entries with fewer than 3 domains are discarded. The next step is to generate a multiple structural

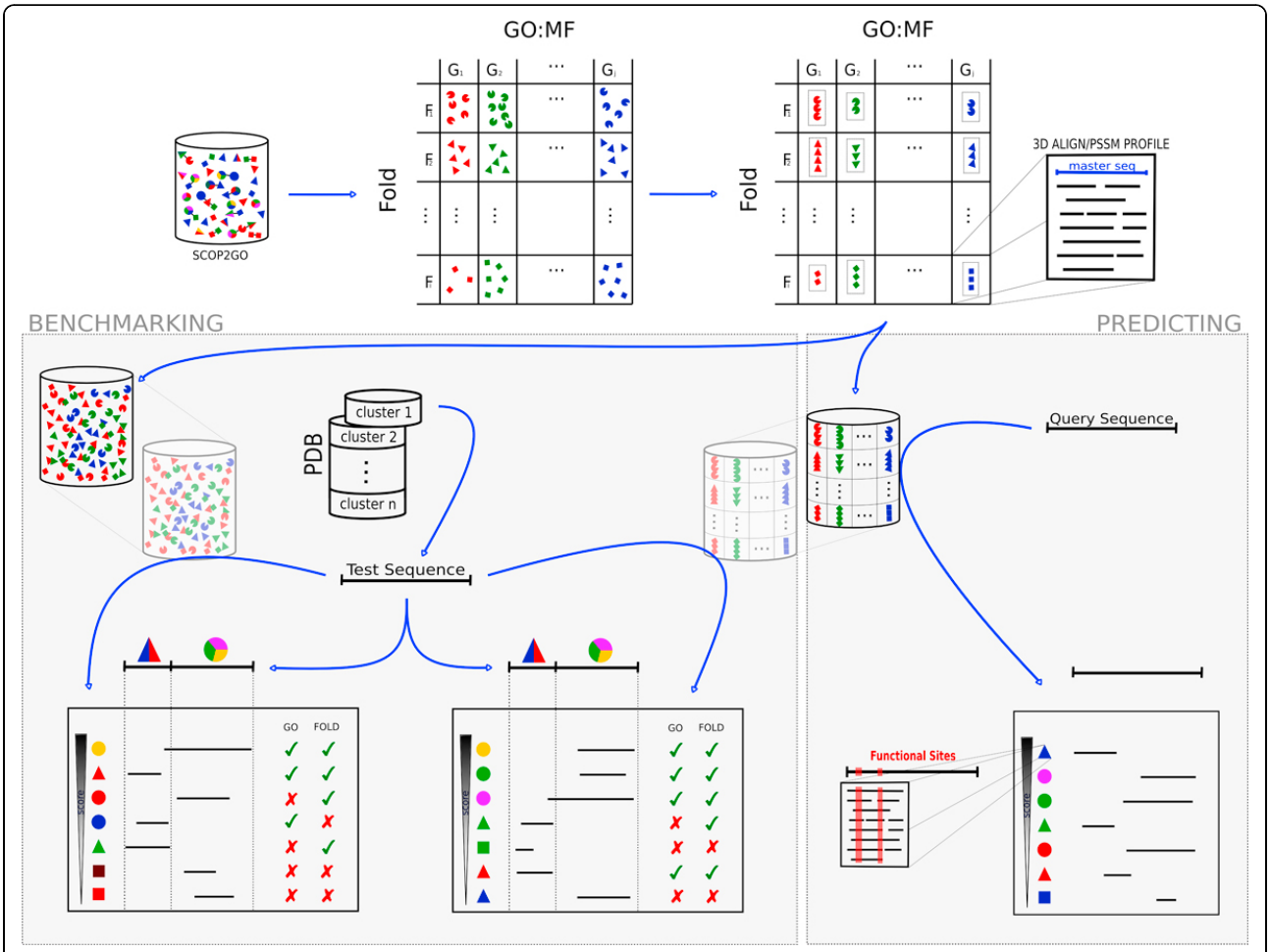


Figure 1 Schema of the method. The starting point is the SCOP2GO functional annotation of chains at the domain level (top left). The shapes represent the folds of the domains, while the colors represent assigned GO:MF functions. Structural alignments for the domains with the same fold and the same function are generated and PSSM profiles are derived from them. For benchmarking (left), an equivalent database is constructed merging all the domain sequences involved in these profiles. For assessing the performance of both resources for a given test sequence, new versions of both databases are built by excluding this sequence and all its homologs (transparent cylinders in the figure). Querying the test sequence against both resources produces list of hits which can be interpreted as predictions of folds and functions (colored shapes) associated to its domains. These predictions of both resources for the domains of the test sequence can be contrasted against its original SCOP2GO annotations (multi-colored triangle and circle). For predicting (right), a sequence of unknown domain characteristics in terms of fold and function is queried against the database of PSSMs. The hits can be interpreted as predictions of fold and function at the domain level. Additionally, the conservation pattern of the structural alignments associated to the matched PSSMs can give clues about functionally important residues.

alignment with the resulting domains (Figure 1). Most programs for generating "real" multiple structural alignments are limited to a relatively small number of structures, which is exceeded in many cases in our dataset. For that reason, we used Dali_lite [22] to generate individual binary alignments of each domain against a "master", and generated a pseudo-multiple structural alignment by piling up these binary alignments. As the master domain, we choose that with the length closest to the mean length of all domains within the subset. The same procedure is repeated for each entry in the matrix. As explained above, each entry in the matrix represents a non-redundant subset of domains with the same structural fold and annotated in SCOP2GO with a specific MF:GO term. Finally, psi-blast PSSM profiles are generated for all these alignments. A PSSM ("position specific scoring matrix") is a representation of the aminoacid distributions of the positions of a multiple sequence alignment [19].

The current version of the library contains 338 entries covering 115 different GO:MF terms and 150 SCOP folds.

Querying a sequence against the library

Since each entry in the library is associated to a fold and a GO:MF term, querying a whole-length sequence

against the alignments/profiles within this library with "reverse psi-blast" (rpsblast) produces a list of hits each representing a concomitant prediction of fold and function for a particular segment (i.e. a domain) of the query sequence (Figure 1). Additionally, due to the way in which these structural alignments are generated, explained above, their conserved positions are expected to correspond to sites with some functional importance for that GO:MF function hosted in that fold, although positions conserved due to purely structural reason would also show up here. For this reason, inspecting the alignment of the query sequence against these conserved positions can give clues on its functional residues as well (Figure 1).

Benchmarking

One of the added values of the method presented here is that the profiles are constructed "informed" by GO:MF annotations, instead of relying on the domain groupings that would result from sequence relationships alone (e.g. families and superfamilies) To evaluate the effect of this, we compared the results of searches against this database of pre-computed profiles, with those obtained by the same method (psi-blast) against an equivalent database with exactly the same domain sequences but not grouped according with GO:MF terms. In order to do that, all the domain sequences after the 40% ID filtering (just before performing the structural alignment) are mixed together in a large database which is formatted for psi-blast (Figure 1). Each sequence retains information on the Fold/GO:MF it comes from in order to later evaluate the results of querying against this database.

We constructed a test set for evaluating the performance of these two resources form the entire PDB clustered at 30% ID downloaded form the RCSB site [23]. The test set is constructed by taking one representative chain per cluster. The first sequence of the cluster having some domain annotated in SCOP2GO is taken. Note that, even if the two resources described above are based on domains, this test set is composed of whole-length chains, since that is the real-world scenario for applying the method presented here. The final test set contains 1017 chains. We have used the largest possible dataset taking into account the requirements of the sequences (known SCOP and SCOP2GO domain annotations) and the sequence redundancy filter.

For each chain in our test set, we carry out the following procedure. First we re-construct the two databases as described before but removing from the very beginning any domain corresponding to a chain within the same 30% ID PDB cluster as the test chain (Figure 1). This is to simulate a scenario in which predictions are going to be generated for sequences without clear homologs in the databases. In the case of the library of profiles, this obviously involves re-building the PSSM profiles which contain any of these chains homologous to the test chain without them. Then, the sequence of the test chain is queried against the two resources (single sequences and profiles) resulting in two lists of hits with their associated scores (e-values), each hit representing a Fold/GO:MF pair (Figure 1). Since the annotations of the domain(s) of the test sequence are known (in SCOP2GO), each hit can be labeled as "true" or "false" in terms of function and fold (Figure 1). The region of the test sequence aligning with a given hit is taken into account when deciding whether that hit is correct or not. I.e. a case in which the test sequence has the same fold/function as the hit but not in the aligned domain is not considered a match (Figure 1). This is done by "blasting" the region of the test sequence aligned with the hit against the sequences of all its domains, taken from ASTRAL [24], to confirm/discard that the alignment is in the correct domain.

So, for a given chain in the test set we obtain two sorted lists of hits, one for each method/resource, called "GO_PROFILE" and "PSI_BLAST" hereafter (Figure 1). Each hit can be labeled as correct or incorrect in terms of fold and function as explained before. In order to base the comparison on the same number of cases, only the top hit of each list (highest score) is evaluated. For that, a single list of "top hits" and their associated scores is generated for each method.

A ROC (receiver operator characteristics) analysis [25] is performed on these lists in order to evaluate the capacity of both resources to discriminate correct from incorrect hits. The ROC analysis generates a plot of "true positives rate" (TPR) against "false positives rate" (FPR) when varying the classification threshold (score of the method). A random method, without discriminative capacity, would produce a list with positives and negatives uniformly distributed through it that would render a diagonal, from [0,0] to [1,1], in the ROC plot. Curves above the diagonal represent methods with some discriminative power. This discriminative capacity is better as the curve gets closer to the top-left corner of the plot ([0,1]). So, a ROC curve is generated by cutting the sorted list of scores at different thresholds and plotting the resulting TPR's against the FPR's, calculated as

$$\mathbf{TPR = Tp/(Tp + Fn) = sensitivity}$$
$$\mathbf{FPR = Fp/(Fp + Tn) = 1 - specificity}$$

where Tp, Fn, Fp and Tn are the "true positives, "false negatives", "false positives" and "true negatives" resulting from a given threshold.

Results

In the first part of this section we show the results of the large-scale evaluation of performances for GO_PROFILE and PSI_BLAST based on a test set of 1017 protein chains as explained in "Methods". In the second part, we show some examples of cases where one of the methods finds a

right match while the other fails and vice versa, and where these failures are due to different reasons, to illustrate the advantage and drawbacks of this methodology as well as its complementarity with others. Another example allows to illustrate an additional advantage of this method: the possibility of obtaining a prediction of functionally important residues associated to the predicted GO:MF term.

Large-scale evaluation

Figure 2 shows the ROC plots generated for the lists of top hits of each method. Figures 2a and 2b show the performance of the methods in detecting the right GO:MF term, while Figure 2c shows the performance in detecting the right fold. The difference between Figures 2a and 2b is that in the last the evaluation is restricted to GO:MF terms far apart from the root of the hierarchy, i.e. those at distance 4 or higher from that root, in an attempt to evaluate only specific GO:MF terms. Although the distance to the root is not a perfect criteria to separate broad (e.g. "enzyme") from specific (e.g. "thymidylate synthase") GO:MF terms due to the uneven distribution of terms in the GO graph and the fact that it is not a perfect hierarchy, is a very convenient and easy way to have a first quantification of the level of broadness/specificity of a term. Actually very broad terms (distance to the root ≤ 2) are never used in this work since they are not included in the original SCOP2GO annotation of domains used for building the profiles [18]. From the original 115 different GO:MF terms contained in the library, 89 end up in the results used for generating Figure 2a, while 78 (more specific, distance > = 4) are used for generating Figure 2b.

It can be seen that both approaches present a very good discriminative power. Nevertheless, GO_PROFILE outperforms PSI_BLAST in assigning the right functions to the right domains (Figures 2a and 2b). When evaluating only more specific GO terms (Figure 2b) the difference in performance is lower and the results of a psi-blast search get closer to those obtained with the methodology presented here. In the Additional File 1 there are additional ROC plots for other levels of "functional specificity" (distances to the GO:MF root) which support this conclusion. This is probably due to the fact that, as we go to more specific functions, these are better reflected at the sequence level and hence they can be captured with standard sequence-based methods. On the contrary, proteins sharing a broad function (i.e. "hydrolase") might have been diverged largely at the sequence level or even lack a common evolutionary origin, and hence the landmarks they share in their sequences can only be captured with "supervised" profiles such as those presented here.

The ROC plots in the Additional File 1 include also the results of an hmmer search against HMM models [26] derived for the same alignments as the PSSMs. They are very similar and both are better than the psi-blast search against single sequences, highlighting the added value of the GO-based profiles which are able to capture subtle sequence landmarks of distant (or even evolutionary unrelated) proteins, as commented in the previous paragraph.

For the case of fold prediction, it can be seen that the performance of both approaches is very high and very similar (Figure 2c).

Examples

The first example is the mitochondrial precursor of the ATP-synthase beta chain ([PDB:1w0k]D). The top hit of our method is the profile GO:0005524/c.37 (function "ATP binding" in fold "P-loop nucleoside phosphate hydrolases"), matched against the central domain of that protein. Nevertheless, an equivalent search with psi-blast

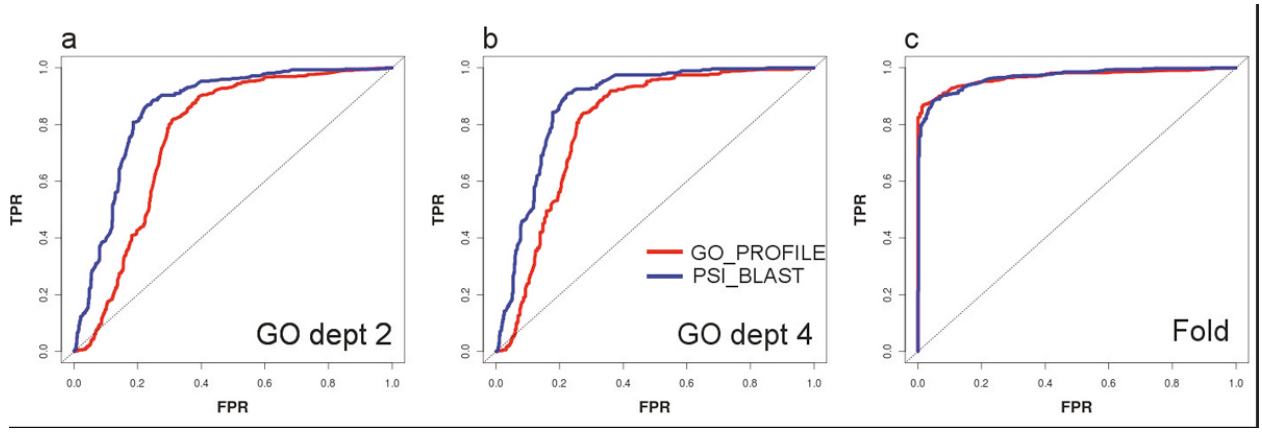


Figure 2 Large-scale evaluation results. ROC plots illustrating the discriminative capacity of the highest scoring hits detected by both methods in detecting the right function (a and b) and fold (c) in the correct domain. a) Only the GO terms at distance 2 or higher from the root of the GO:MF graph are evaluated; b) the same for distance 4 or higher (more specific terms).

finds as top hit a domain with function GO:0004156 ("dihydropteroate synthase activity", an enzymatic activity which is not even ATP-dependent) and fold c.1 ("TIM-barrel").

Another example is the periplasmic cytochrome C551I ([PDB:2mta]C). While GO_PROFILES correctly predicts GO:0020037 ("heme binding") in fold a.3 ("cytochrome C fold"), psi-blast's top hit is a DNA-binding protein (GO:0003677) with fold d.218.

The next example illustrates a problem of this method: the quality of the GO:MF domain annotations it relies on. For the Aspartyl-tRNA synthase [PDB:1b8a]B, the method matches its N-terminal domain with the profile GO:0005524/b.40 (function "ATP-binding" in fold "all-β/OB-fold". Such profile should not exist since there are not proteins with domains of that particular fold hosting that function. Nevertheless, there are examples of such domains wrongly annotated with that function in SCOP2GO. The reason for these wrong annotations is that these domains (responsible for anticodon binding in tRNA synthases) are frequently linked to the ATP-binding domains of these proteins, and there are many instances of them crystallized in isolation (as fragments) in PDB. The problem arises because these fragments are annotated with the function of the complete chain (ATP-binding) and consequently confound the methodology used in SCOP2GO (see [18] for details). On the contrary, psi-blast correctly matches this domain against the correct ATP-binding domain of a protein. This kind of errors due to problems in the SCOP2GO annotations would be alleviated as the SCOP2GO annotation is improved (e.g. by manual curation, etc) or future functional annotations at the domain level are used.

In the case of the mono-domain tyrosine phosphatase [PDB:1l8k]A, our method "correctly" matches it against the GO:0004725/c.45 profile ("protein tyrosine phosphatase activity" in fold "α/β phosphotyrosine phosphatases"). Nevertheless, this protein is annotated with a less specific term in GO (GO:0004721, "phosphoprotein phosphatase activity", the "ancestor" of GO:0004725). For this reason this counts as a failure in the automatic large-scale evaluation discussed in the previous point, even if our method is providing a more detailed (and correct) annotation. In this case, psi-blast matches against a protein with that less specific annotation (GO:0004721) and hence it counts as a true match.

The last example illustrates an additional advantage of this method: the fact that it can provide clues about possible functional sites, concomitantly with the prediction of fold and function. The casein kinase 1 ([PDB:2csn]A) is correctly matched against the GO:0004672/d.144 profile ("protein kinase activity" in fold "protein kinase like"). Figure 3 shows the positions conserved (95%) in this profile mapped on the 3D structure of this kinase, together with the residues annotated in the "catalytic site atlas" [27] for the same protein. It can be seen that all but one conserved residues either are annotated as catalytic, are very close to them, or are involved in binding cofactors (Figure 3).

Discussion

It is well known that most proteins, especially in eukaryotic organisms, are multidomain [28]. In most cases, these domains perform distinct and quite independent molecular functions, to the extreme that some of these domains exist as independent proteins in other organisms (This is actually the basis of the "Rosetta Stone" method for predicting protein relationships [29].)

As commented in the Introduction, almost all methods and resources for predicting protein function are intended to work at the whole-chain level. Even the functional annotations of entries domain-oriented databases such as Pfam or Intepro are not intended to be interpreted in terms of physical domains. In [18] we show many examples of errors obtained when these resources and databases are used to infer annotations at the domain level. Obviously, this problem only applies to the "molecular function" aspect of the proteins, since the other two GO functional aspects ("cellular component" and "biological process") apply to complete chains and not domains.

The main methodological novelty of the procedure presented here is the usage of profiles derived from structural alignments of all domains associated to a given GO molecular function. This association of GO:MF terms to structural domains is taken from the first resource specifically devoted to this task [18]. Including all domains associated to a given function (those that can be structurally aligned, actually), and not only those with a common evolutionary origin, ensures that the molecular signatures within these profiles comprise the information of the whole sequence-space associated to a particular function (within a fold), and not only that restricted to a particular family or superfamily of proteins. This is actually one of the major differences, together with the GO:MF annotations at the domain level, with other resources intended to search against profiles derived for families or superfamilies [30]. In turn, these resources have the advantage that a match against their profiles provides additional information on family/superfamily membership and evolutionary origin. In this sense, all these resources complement each other in the evolutionary, structural and functional characterization of proteins at the domain level.

We compare this method with a base-line methodology for predicting protein function (psi-blast) in order to illustrate the added value of these novelties. Actually, only the added value of the GO-informed profiles, since the GO:MF annotations at the domain level are also provided to psi-blast in this benchmark. The large-scale evaluation

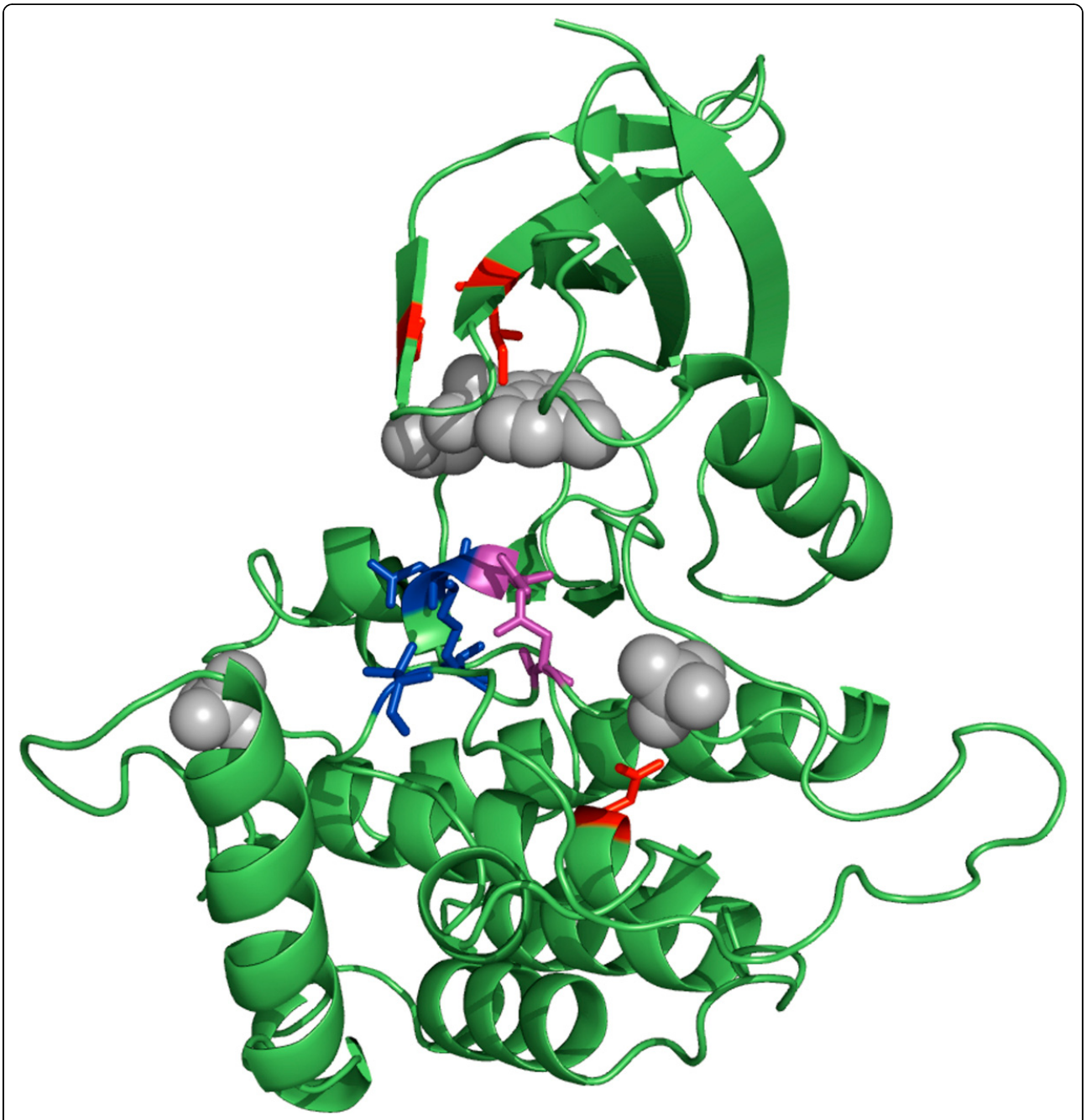


Figure 3 Example of predicted functional residues. Residues predicted by the method as associated to the GO:0004672 ("protein kinase activity") function for the casein kinase 1 ([PDB:2csn]A) mapped in the structure of this protein. Red and purple: predicted residues. Blue and purple: catalytic residues annotated in CSA. The prosthetic groups are shown in grey and spacefill. Figure generated with PyMOL (http://www.pymol.org).

based on GO annotations is not perfect due to many reasons (unspecific annotations, etc), some of them illustrated in the examples shown. Nevertheless, all these factors affect both methodologies due to the parallel evaluation procedure followed, based on the same dataset. An exhaustive comparison with more sophisticated methods for function prediction is outside the scope of this work due to the different functional vocabularies and databases used and, more importantly, the domain orientation of this method: almost all other existing resources for function prediction assign function at the whole-chain level. We also show some examples to illustrate the advantages of this method in particular situations and highlight its complementariness with existing approaches. Indeed, the

method presented here is not intended to compete with the plethora of methods designed for predicting function at the whole chain level, but to fill a very particular niche: function prediction at the domain level. Nevertheless, this method can be also used to infer molecular functions for whole chains in two ways: i) although excluded from the benchmark presented here for simplicity, the SCOP2GO resource also contain functional assignments for groups of domains, and ii) the function of the whole chain can be manually inferred from the annotations of the individual domains although this requires some expert knowledge.

The domain orientation of this methodology also makes that it can be only applied to the "molecular function" category of GO (GO:MF) and not to the "biological process" category (GO:BP). As commented above, only the molecular functions can be differentially assigned to particular domains, while biological processes are properties of whole chains.

This resource will be improved as the GO:MF annotations at the domain level it is based on, which right now are generated with an automatic procedure, are extended and manually curated. Moreover, the method presented here can be implemented with any other domain-based functional annotation.

Conclusions

We present here a method and a resource for the concomitant prediction of fold and molecular function at the domain level, using a single sequence as input. The method outperforms standard sequence-based methods. Functionally important sites may also be identified although this feature has not been exhaustively benchmarked so far and we only show illustrative examples.

Additional material

Additional file 1: Additional results of the large-scale evaluation. Additional ROC plots for other levels of functional specificity (distance to the root of the GO:MF graph) including also the results of HMM searches against the profiles.

Authors' contributions

FP conceived the original idea. FP and DL designed the experiments. DL implemented and performed all the experiments. FP and DL contributed to the writing of the manuscript.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

The authors want to thank the members of the Computational Systems Biology Group (CNB-CSIC) and Dr. Mark Wass (CNIO) for interesting discussions and support.

Funding: This work was funded by the Spanish Ministry for Science and Innovation [projects numbers **BIO2009-11966** and **BIO2010-22109**].

This article has been published as part of *BMC Bioinformatics* Volume 14 Supplement 3, 2013: Proceedings of Automated Function Prediction SIG 2011 featuring the CAFA Challenge: Critical Assessment of Function Annotations. The full contents of the supplement are available online at URL. http://www.biomedcentral.com/bmcbioinformatics/supplements/14/S3

Published: 28 January 2013

References

1. Schuster SC: **Next-generation sequencing transforms today's biology.** *Nat Methods* 2008, **5(1)**:16-18.
2. Devos D, Valencia A: **Practical limits of function prediction.** *Proteins* 2000, **41**:98-107.
3. Tian W, Skolnick J: **How well is enzyme function conserved as a function of pairwise sequence identity?** *J Mol Biol* 2003, **333(4)**:863-882.
4. Rost B: **Enzyme function less conserved than anticipated.** *J Mol Biol* 2002, **318**:595-608.
5. Chagoyen M, Pazos F: **Quantifying the Biological Significance of Gene Ontology Bio-logical Processes - Implications for the Analysis of Systems-wide data.** *Bioinformatics* 2010, **26**:378-384.
6. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25(1)**:25-29.
7. Rentzsch R, Orengo C: **Protein function prediction - the power of multiplicity.** *Trends Biotech* 2009, **27(4)**:210-219.
8. Valencia A: **Automatic annotation of protein function.** *Curr Opin Struct Biol* 2005, **15(3)**:267-274.
9. Watson JD, Laskowski RA, Thornton JM: **Predicting protein function from sequence and structural data.** *Curr Opin Struct Biol* 2005, **15(3)**:275-284.
10. Rost B, Liu J, Nair R, Wrzeszczynski KO, Ofran Y: **Automatic prediction of protein function.** *Cell Mol Life Sci* 2003, **60(12)**:2637-2650.
11. Gotz S, Garcia-Gomez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talon M, Dopazo J, Conesa A: **High-throughput functional annotation and data mining with the Blast2GO suite.** *Nucleic Acids Res* 2008, **36(10)**:3420-3435.
12. Hawkins T, Luban S, Kihara D: **Enhanced automated function prediction using distantly related sequences and contextual association by PFP.** *Protein Sci* 2006, **15(6)**:1550-1556.
13. Martin DM, Berriman M, Barton GJ: **GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes.** *BMC Bioinformatics* 2004, **5**:178.
14. Pal D, Eisenberg D: **Inference of protein function from protein structure.** *Structure (Camb)* 2005, **13(1)**:121-130.
15. Pazos F, Sternberg MJE: **Automated prediction of protein function and detection of functional sites from structure.** *Proc Natl Acad Sci USA* 2004, **101(41)**:14754-14759.
16. Wass MN, Sternberg MJ: **ConFunc–functional annotation in the twilight zone.** *Bioinformatics* 2008, **24(6)**:798-806.
17. Xie L, Bourne PE: **Detecting evolutionary relationships across existing fold space, using sequence order-independent profile-profile alignments.** *Proc Natl Acad Sci USA* 2008, **105(14)**:5441-5446.
18. Lopez D, Pazos F: **Gene Ontology functional annotations at the structural domain level.** *Proteins* 2009, **76**:598-607.
19. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucl Acids Res* 1997, **25**:3389-3402.
20. Andreeva A, Howorth D, Brenner SE, Hubbard TJ, Chothia C, Murzin AG: **SCOP database in 2004: refinements integrate structure and sequence family data.** *Nucleic Acids Res* 2004, **32(Database)**:D226-229.
21. Notredame C, Higgins DG, Heringa J: **T-Coffee: A novel method for fast and accurate multiple sequence alignment.** *J Mol Biol* 2000, **302(1)**:205-217.
22. Holm L, Kaariainen S, Wilton C, Plewczynski D: **Using Dali for structural comparison of proteins.** *Curr Protoc Bioinformatics* 2006, vol. Chapter 5, Unit 5 5.
23. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank.** *Nucleic Acids Res* 2000, **28**:235-242.
24. Chandonia JM, Hon G, Walker NS, Lo Conte L, Koehl P, Levitt M, Brenner SE: **The ASTRAL compendium in 2004.** *Nucl Acids Res* 2004, **32**:D189-D192.
25. Fawcett T: **An introduction to ROC analysis.** *Pattern Recogn Lett* 2006, **27(8)**:861-874.

26. Wistrand M, Sonnhammer EL: **Improved profile HMM performance by assessment of critical algorithmic features in SAM and HMMER.** *BMC Bioinformatics* 2005, **6**:99.
27. Porter CT, Bartlett GJ, Thornton JM: **The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data.** *Nucleic Acids Res* 2004, **32(Database)**:D129-133.
28. Apic G, Gough J, Teichmann SA: **Domain combinations in archaeal, eubacterial and eukaryotic proteomes.** *J Mol Biol* 2001, **310(2)**:311-325.
29. Marcotte EM, Pellegrini M, Ho-Leung N, Rice DW, Yeates TO, Eisenberg D: **Detecting protein function and protein-protein interactions from genome sequences.** *Science* 1999, **285**:751-753.
30. de Lima Morais DA, Fang H, Rackham OJ, Wilson D, Pethica R, Chothia C, Gough J: **SUPERFAMILY 1.75 including a domain-centric gene ontology method.** *Nucleic Acids Res* 2011, **39(Database)**:D427-434.

doi:10.1186/1471-2105-14-S3-S12
Cite this article as: Lopez and Pazos: Concomitant prediction of function and fold at the domain level with GO-based profiles. *BMC Bioinformatics* 2013 **14**(Suppl 3):S12.

Natalia Pietrosemoli, Daniel López, Aldo Segura-Cabrera, and Florencio Pazos

Computational Prediction of Important Regions in Protein Sequences

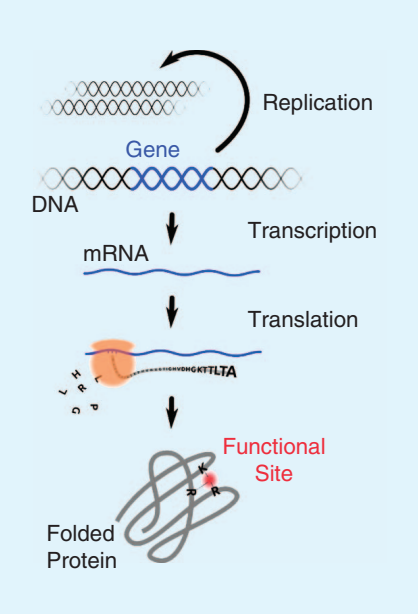
If the genomic era was characterized by the massive determination of genomic sequences, the so-called postgenomic era is, among other things, characterized by a lack of methods for obtaining functionally relevant information from these raw sequences. As the number of known protein sequences grows exponentially, it is impossible to experimentally determine their biological functions and the particular regions of these proteins responsible for such functions. For this reason, computational methods that are able to process this genomic information for extracting protein functional features are sought after.

This column summarizes the main approaches for predicting protein functional sites from sequence information. Extracting functional regions from linear protein sequences has many similarities with prototypical problems in signal processing, and some of the methodologies used are taken from this discipline.

PROTEINS

Proteins are the key players in all biological processes. It is difficult to think of a biological process in which these macromolecules are not involved. Proteins carry out the most diverse functions within a cell such as catalysis of the chemical transformations of the metabolism (enzymes), formation of structures and frameworks, communication, and signaling.

Proteins are synthesized as linear chains of subunits called amino acids. There are 20 different natural amino acids with different physico-chemical properties, which are the basic building blocks of proteins. The type, number, and order of amino acids in a given protein chain, known as protein sequence, are coded by the corresponding gene (Figure 1). Within the long chain of deoxyribonucleic acid (DNA) that comprises an organism genome, a gene can be defined as the piece of DNA that carries the required information for synthesizing a particular protein. The human genome, for instance, performs coding for around 25,000 proteins. After being synthesized as a linear chain of amino acids, a protein chain folds into a particular three-dimensional (3-D) structure (Figure 1), due to interactions between its amino acids. In general, the building blocks of proteins are named *amino acids* when they are free (not linked within a protein chain) and *residues* when they are linked to form a protein.



[FIG1] "Central dogma" of molecular biology, representing the flow of genetic information. The genetic information is coded in a double helix of DNA, which can be replicated to pass identical copies to the offspring. The segment of DNA that carries the required information for building a protein is called a gene. A gene is transcribed into a messenger ribonucleic acid (mRNA), which goes to the ribosome (orange), a macromolecular structure in charge of building proteins. The ribosome builds a protein by linking instances of the 20 possible building blocks (amino acids; see the letters in the figure) in the order and amount specified by the mRNA. This linear string of amino acids folds into a 3-D structure leading to a functional protein. In this 3-D arrangement, certain residues lay in the right conformation to form the active site of the protein.

In a simplified view, proteins can be considered as nanomachines that perform different tasks (functions) within living systems and the genes would be the "plans" describing how to build these machines. Most proteins have to be folded in specific 3-D conformations (Figure 1) to perform their biological functions. Only in this conformation a particular set of residues is placed in the required 3-D layout to specifically interact with other biological molecules and perform the function of the protein. This set of residues is known as the active site or functional site of the protein.

Protein sequences change over time (evolve) in a process of Darwinian evolution. New protein sequences that have arisen from random changes in existing proteins (ultimately due to mutations in the corresponding genes) can be either discarded if they lead to a nonfunctional protein or maintained if the protein is functional. This evolutionary process is responsible for the emergence of proteins with new or improved functions. All proteins that have arisen from a common ancestor through this evolutionary process are known as *homologous*. Accordingly, homologous proteins tend to have similar amino acid sequences and functions and have the same global 3-D structure [4].

Digital Object Identifier 10.1109/MSP.2012.2211476
Date of publication: 15 October 2012

Recent decades have seen a tremendous improvement in DNA sequencing techniques, which has allowed determining the complete genomes of more than 1,700 organisms. The corresponding proteomes (set of protein sequences coded by given genomes) can be obtained by translating the genes in these genomes. Because of this, the number of known protein sequences grows exponentially, doubling approximately every two years. Consequently, today we know the sequences of more than 20 million proteins, which are available in different Web-accessible databases. On the contrary, determining protein 3-D structures experimentally is not an easy task, and we only know the 3-D structures of around 50,000 different proteins.

PREDICTION OF PROTEIN FUNCTIONAL SITES

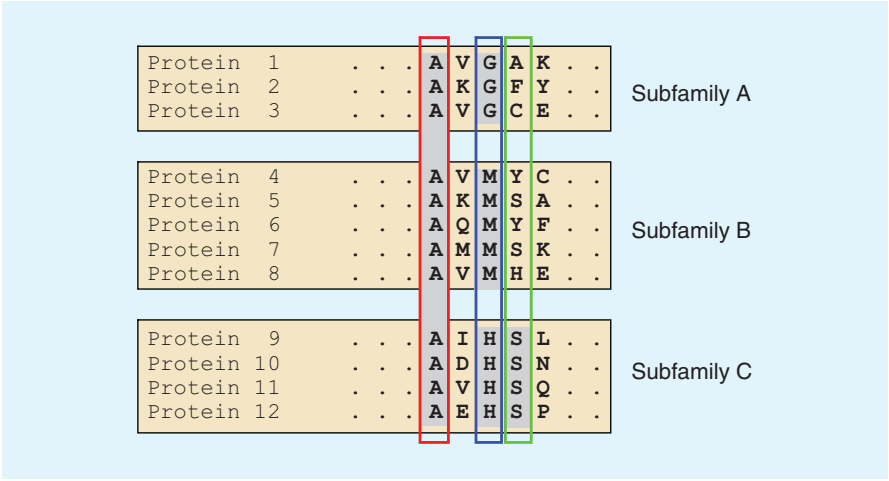
Determining which residues within a protein are functional is important not only to understand its function at the atomic level, but also to devise ways for modifying it to our benefit.

Experimental determination of protein functional sites is both very expensive and time-consuming. This experimental determination consists in changing a specific residue(s) of a protein sequence by another residue(s) and evaluating whether these changes produce a nonfunctional protein or a protein with an altered function. To improve the efficiency of this process, a number of computational techniques have been developed that use information on protein amino acid sequences (and in some cases 3-D structures) to predict functional sites. These methods can help experimental techniques, for example by providing a number of candidate positions to mutate.

Most computational methods are based on the fact that functional residues tend to be preserved in the evolutionary process, i.e., the amino acid types within these positions are not allowed to change since they are required to perform a particular function. Accordingly, when a set of homologous proteins is compared, the functional residues tend to be conserved among them.

PROTEIN SEQUENCE ALIGNMENTS AND FUNCTIONAL SITES

Homologous proteins are proteins sharing a common ancestor. These proteins can be grouped and their sequences aligned in such a way that evolutionary equivalent residues can be compared. This layout is known as multiple sequence alignment (MSA). In its most common form, an MSA is represented as a matrix, in which proteins are arranged in rows and the columns represent equivalent residues (Figure 2). Columns are called *positions* of the MSA. MSAs are rich sources of structural and functional information because they can be seen as a representation of the amino acid changes, allowed by evolution at each position.



[FIG2] Representation of a portion of a multiple sequence alignment of the sequences of 12 homologous proteins. The proteins are in rows, and the columns (positions) represent equivalent residues. Three subfamilies can be defined in this alignment. Fully conserved positions (red) are important for the whole set of proteins. Positions with a subfamily-dependent conservation pattern (blue and green) are related to functional specificity. Conservation is highlighted in grey.

CONSERVED POSITIONS

As mentioned earlier, functional residues tend to be conserved among homologous proteins. Consequently, the first-explored and most obvious pattern related to functionality, extracted from multiple sequence alignments is that related to conserved positions. Although the problem of locating conserved positions might look trivial at first sight, there is not a unique method to solve it [15]. In most cases the conservation is not perfect, e.g., reflected in alignment columns with only one amino acid type. This nonperfect conservation can be due to many reasons, including errors in the alignment of the proteins and conservative substitutions (changes by other amino acids with similar physico-chemical properties). Additionally, the overall similarity between the sequences needs to be taken into account when assessing the conservation of a given position, indeed if the proteins are globally very similar, an observation of conservation is less indicative of functionality.

In most cases, the residues forming the functional/active site of a protein are close in its 3-D structure, to "collaborate" for performing this function and preferentially located in its surface to be accessible to the molecules this protein acts on (Figure 1). Consequently, if the 3-D structure of a member of the homologous family is available, it can be used to assess the spatial clustering and surface exposure of the conserved positions. This additional information can then be used to further filter out those candidate positions that do not satisfy surface and spatial conditions.

One of the best methodologies for detecting conserved positions related to functionality in MSAs is implemented in the Conseq/Consurf server [1], which incorporates a sophisticated method for taking into account the phylogeny of the sequences in the MSA so as to avoid the

artefacts due to the peculiarities or uneven distribution of sequences in the above-mentioned alignment.

FAMILY-DEPENDENT CONSERVED POSITIONS

Some positions in MSAs show a more subtle pattern of conservation. In MSAs that can be partitioned into distinct groups of proteins, these positions are differentially conserved in these different subgroups. They are conserved in a given group but not in another, or the conserved amino acid is different in the different subgroups (Figure 2). These subgroups (also termed *subfamilies*) can be defined based on different criteria: phylogenetic (the groups evolved independently) or functional (the groups have slightly different functions). The fact that these positions are conserved is indicative of their functional importance, whereas the fact that the amino acid type is different for different subfamilies indicates that this importance is subfamily specific. These positions are important for the feature used for defining the subfamilies. If subfamilies are defined according to functional criteria, these positions are related to functional specificity, i.e., they are responsible for the functional differences between the subfamilies. For that reason, they are also termed *specificity-determining positions* (*SDPs*).

There are many methods for detecting this kind of positions in MSAs. One of the first approaches for detecting conservation patterns beyond complete conservation was the evolutionary trace (ET) method [9]. ET explores the hierarchical partitions of the MSA into subfamilies and looks for the conserved positions that show up at each partition. Fully conserved positions become apparent at partition zero (the whole MSA), while partition one will show positions that are differentially conserved in the two main subfamilies within the MSA, and so on. A rank is assigned to each position in the alignment depending on the partition, in which it becomes conserved.

Another widely used family of methodologies for detecting SDPs is based on a vectorial representation of the MSA, in which each protein is represented as a vector in a high-dimensional space based on its amino acid sequence. Then, a dimensionality-reduction method such as principal component analysis (PCA) or multiple component analysis (MCA) is applied to this space to obtain a low-dimensional one, preserving most of the information. In this space, vectors representing similar proteins (subgroups, subfamilies) are clustered together [3]. A similar vectorial treatment for the individual residues produces an equivalent space where those that are informative for explaining the separation of the subfamilies (i.e., SDPs) are located in the same regions of the space where the clusters representing these subfamilies are. A prototypical representative of this family of approaches is the S3Det method [14], which represents proteins and residues as binary vectors and uses MCA for dimensionality reduction (Figure 3).

A CLEVER COMBINATION OF EXPERIMENTAL AND IN SILICO APPROACHES WILL DEFINITIVELY HELP INTERPRET THE GENETIC INFORMATION IN FUNCTIONAL TERMS, WHICH IS THE FINAL GOAL OF THE SO-CALLED POSTGENOMIC ERA.

A third family of approaches is based on the comparison of the mutational behavior of a position (or a set of positions) with the overall mutational behavior of the whole alignment. For example, in Xdet [13], the mutational behavior of a position is represented by a matrix that contains the physico-chemical similarities for all pairs of amino acids at that position. Likewise, the mutational behavior of the whole alignment is represented by an equivalent matrix that contains the overall similarities for the corresponding pairs of proteins. These matrices are compared producing a score for that position of the MSA. Positions with the highest scores are selected as the predicted SDPs. The rationale for this family of approaches is that the mutational pattern of the positions with a family-dependent conservation pattern resembles that of the whole alignment, since this is the expected behavior for positions responsible for subfamily separation.

Finally, another widely used family of methods is based on the evaluation of the mutual information between the subfamily distribution and the distribution of amino acids in a given position, thereby scoring its importance in defining the subfamily specificity [8], [10]. These methods use as additional input the subfamily classification of the MSA. Notice that the methods previously described do not use that information explicitly, but they infer the subfamily composition from the sequence relationships in the same MSA. Nevertheless, there are also versions of these methods that can incorporate the subfamily composition as additional input, instead of inferring it from the MSA. The methods can be classified as supervised and unsupervised [13], depending on whether they take the subfamily composition as additional input or not.

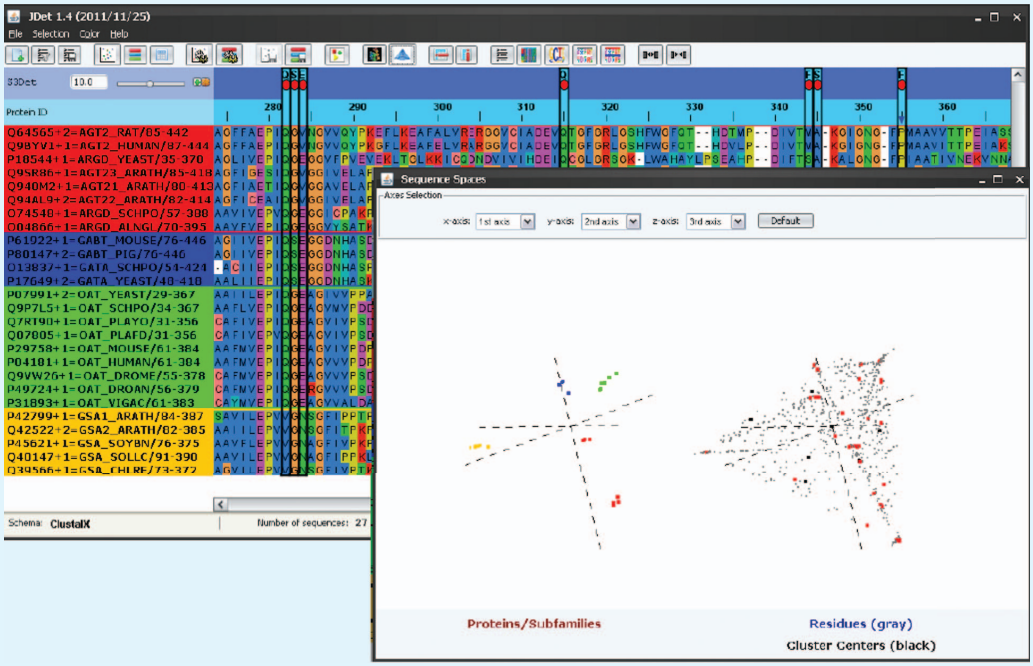
Just as in the case of fully conserved positions, methods for locating SDPs can use 3-D information if available. Generally, that information is used for evaluating the clustering and surface accessibility of the candidate positions.

OTHER PROTEIN FUNCTIONAL SITES

The functional positions extracted from MSAs are not related to any particular functional feature but to the function of the protein in general. We can say that one of these sites is "important" for something but not whether it is a catalytic site, or a site involved in protein interaction, etc. This is because the evolutionary assumption they rely on (conservation due to functional importance) is common to all functions carried out by the protein. Nevertheless, some methods have been "tuned" and adapted to detect particular types of functional features.

FUNCTIONAL PROFILES AND MOTIFS

If we restrict the proteins in a given MSA to those known to have a particular function, we can use this subalignment for defining a profile or motif associated to that function. These code the amino-acid



[FIG3] Results of S3Det [14] as a prototype of SDP prediction methods based on multidimensional representations of the proteins and residues. Three-dimensional projections of the spaces of proteins and residues after the MCA dimensionality reduction are shown. The colors in the space of proteins represent the subgroups automatically detected by the program. In the multiple sequence alignment in the background, protein names are colored according to the same scheme and the residues detected by the program as associated to the family separation (SDPs) are highlighted. (Figure generated with JDet [11]).

distribution of a particular subset of positions of the MSA (e.g., those known to be part of the active site, or just the conserved ones). These profiles are coded as position specific scoring matrices (PSSMs) or, more recently, hidden Markov models (HMMs). Once defined, these profiles can be used for matching new sequences against them. Positive matches of protein sequences against these motifs can be used not only for predicting the function of uncharacterized proteins, but also for locating their functional sites. The most widely used collection of motifs related to different types of protein functional features is that contained in PROSITE [2].

PROTEIN INTERACTION SITES

Many proteins exert their functions by interacting with other proteins. The residues involved in this interaction are known as protein interaction (or protein binding) sites. As any other functional site, these are subjected to the same evolutionary constraints described before, and hence many of them tend to be conserved (either fully conserved or SDP). However, there are also methods specifically designed for locating interaction sites.

The majority of these methods are based on training. Machine learning (ML) systems, such as neural networks [12], are trained with protein segments known to be involved in protein-protein interactions and later used for predicting this kind of site in new sequences. Some of these ML methods also require information on the protein 3-D structure [7]. In the case of protein binding sites, knowing the 3-D structure of the protein is especially informative since these sites are necessarily located in the surface.

DISORDERED REGIONS

Breaking the paradigm stated at the beginning of this review that a protein has to be folded in a fixed 3-D structure to accomplish its function, there are protein segments (or even full proteins) that need to be unfolded to perform their functions. These "flexible" segments lacking a defined 3-D structure are generally termed *intrinsically disordered regions* (*IDRs*) [6]. There are two main functions exerted by these regions: they can act as flexible connectors or as protein interactions sites. Protein binding mediated by IDRs are especially suitable for certain types of protein interactions, such as transient interactions with multiple partners.

These IDRs show a very particular amino acid composition that is exploited by disorder prediction methods. Early methods were mostly based in ML systems such as neural networks and support vector machines [16]. Such methods are trained with experimentally known IDRs and, afterward, used for predicting these regions in other protein sequences. Other approaches for disorder prediction are based in the amino acid physical properties, such as hydrophobicity and net charge [5].

CONCLUSIONS AND FUTURE DIRECTIONS

The need to obtain functional information for the vast amounts of new

sequences coming from high-throughput sequencing efforts pushed the development of a plethora of methods for detecting functional sites from sequence information. Most of these methods are available to the community as standalone programs or Web servers [11]. Depending on the available information for the protein of interest (e.g., existence of 3-D structure and/or enough sequence homologs) and the type of functional feature one tries to predict (e.g., active sites, protein binding sites) different subsets of these methods can be used.

There are many parallels at different levels between the methodologies for predicting functional sites in proteins and some used in signal processing. Recognizing within a linear protein sequence the few residues responsible for its activity can be seen as a feature extraction or pattern recognition problem. The parallels go deep into the methodologies themselves, e.g., HMMs, now widely used for coding the information of MSAs, were originally developed for pattern recognition in signal processing.

As new methods continue to be developed, it is difficult for the user to assess which one is better. Unfortunately, the original publications are not the best sources of information for comparing the methods since the performance figures reported there are based on different data sets or obtained with different protocols. Additionally, there is the lack of a univocal definition of "functional residue." Moreover, one of the issues that needs to be pursued in the future is the creation of curated data sets of functional sites ("gold standard") that can be used for training, testing, and comparing methods.

In spite of the limitations and problems described above, the discussed methods are clearly in demand, since it is impossible to experimentally obtain functional information for the increasing stream of new sequences. A clever combination of experimental and in silico approaches will definitively help interpret the genetic information in functional terms, which is the final goal of the so-called postgenomic era.

RESOURCES

- Uniprot, the main repository of protein sequences and associated functional features: http://www.uniprot.org/
- Web interface to the BLAST suite of programs for locating homologs for a give protein in sequence databases. The same interface allows to generate a MSA with the found homologs: http://blast.ncbi.nlm.nih.gov/
- JDet software implementing a number of methods for predicting functional positions in MSAs: http://csbg.cnb.csic.es/JDet/.

AUTHORS

Natalia Pietrosemoli (npietrosemoli@cnb.csic.es) is a visiting scholar from Rice University in Houston, Texas, at the National Centre for Biotechnology in Madrid, Spain. Her main research focus is on the relationships between protein sequence, structure, and function with a keen interest for protein intrinsic disorder.

Daniel López (dlopez@cnb.csic.es) is a Ph.D. student at the Computational System Biology Group, National Center for Biotechnology, Madrid, Spain. His research interests include the study of biological networks from a system biology point of view and the study and prediction of protein function and structure.

Aldo Segura-Cabrera (asegurac@ipn.mx) leads the Laboratory of Bioinformatics at Centro de Biotecnología Genómica (IPN), Mexico. His research interests include biophysics, structural bioinformatics and computational systems biology, and the practical applications of these to the drug discovery process.

Florencio Pazos (pazos@cnb.csic.es) leads the Computational Systems Biology Group at the Spanish National Centre for Biotechnology, Madrid. His main research interests are related to the application of bioinformatics techniques to the study of biological networks, the prediction of functional features in proteins, and the prediction of protein-protein interactions.

ACKNOWLEDGMENTS

We thank the members of the Computational Systems Biology Group (CNB-CSIC) and Antonio Rausell, David Juan, and Alfonso Valencia (CNIO) for interesting discussions. This work was partially funded by project BIO2010-22109 from the Spanish Ministry for Economy and Competitiveness.

REFERENCES

[1] A. Armon, D. Graur, and N. Ben-Tal, "ConSurf: An algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information," *J. Mol. Biol.*, vol. 307, pp. 447–463, 2001.

[2] A. Bairoch, "PROSITE: A dictionary of sites and patterns in proteins," *Nucl. Acids Res.*, vol. 20, pp. 2013–2018, 1992.

[3] G. Casari, C. Sander, and A. Valencia, "A method to predict functional residues in proteins," *Nat. Struct. Biol.*, vol. 2, pp. 171–178, 1995.

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

[5] Z. Dosztanyi, V. Csizmok, P. Tompa, and I. Simon, "IUPred: Web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content," *Bioinformatics*, vol. 21, pp. 3433–3434, 2005.

[6] A. K. Dunker, I. Silman, V. N. Uversky, and J. L. Sussman, "Function and structure of inherently disordered proteins," *Curr. Opin. Struct. Biol.*, vol. 18, pp. 756–764, 2008.

[7] P. Fariselli, F. Pazos, A. Valencia, and R. Casadio, "Prediction of protein-protein interaction sites in heterocomplexes with neural networks," *Eur. J. Biochem.*, vol. 269, pp. 1356–1361, 2002.

[8] S. S. Hannenhalli and R. B. Russell, "Analysis and prediction of functional sub-types from protein sequence alignments," *J. Mol. Biol.*, vol. 303, pp. 61–76, 2000.

[9] O. Lichtarge, H. R. Bourne, and F. E. Cohen, "An evolutionary trace method defines binding surfaces common to protein families," *J. Mol. Biol.*, vol. 257, pp. 342–358, 1996.

[10] L. A. Mirny and M. S. Gelfand, "Using orthologous and paralogous proteins to identify specificity-determining residues in bacterial transcription factors," *J. Mol. Biol.*, vol. 321, pp. 7–20, 2002.

[11] T. Muth, J. A. García-Martín, A. Rausell, D. Juan, A. Valencia, and F. Pazos, "JDet: Interactive calculation and visualization of function-related conservation patterns in multiple sequence alignments and structures," *Bioinformatics*, vol. 28, pp. 584–586, 2012.

[12] Y. Ofran and B. Rost, "ISIS: Interaction sites identified from sequence," *Bioinformatics*, vol. 23, pp. e13–16, 2007.

[13] F. Pazos, A. Rausell, and A. Valencia, "Phylogeny-independent detection of functional residues," *Bioinformatics*, vol. 22, pp. 1440–1448, 2006.

[14] A. Rausell, D. Juan, F. Pazos, and A. Valencia, "Protein interactions and ligand binding: From protein subfamilies to functional specificity," *Proc. Natl. Acad. Sci. USA*, vol. 107, pp. 1995–2000, 2010.

[15] W. S. Valdar, "Scoring residue conservation," *Proteins*, vol. 48, pp. 227–241, 2002.

[16] J. J. Ward, J. S. Sodhi, L. J. McGuffin, B. F. Buxton, and D. T. Jones, "Prediction and functional analysis of native disorder in proteins from the three kingdoms of life," *J. Mol. Biol.*, vol. 337, pp. 635–645, 2004.

[SP]

Bioinformatics Advance Access published May 29, 2013

Sequence Analysis

COPRED: Prediction of fold, GO molecular function and functional residues at the domain level

Daniel López and Florencio Pazos*

Computational Systems Biology Group (CNB-CSIC). c/ Darwin, 3. Cantoblanco. 28049 Madrid. Spain.

Associate Editor: Prof. Burkhard Rost

ABSTRACT

Summary: Only recently the first resources devoted to the functional annotation of proteins at the domain level started to appear. The next step is to develop specific methodologies for predicting function at the domain level based on these resources, and to implement them in web servers to be used by the community. In this work we present COPRED, a web server for the concomitant prediction of fold, molecular function and functional sites at the domain level, based on a methodology for domain molecular function prediction and a resource of domain functional annotations previously developed and benchmarked.

Availability and Implementation: COPRED can be freely accessed at: http://csbg.cnb.csic.es/copred . The interface works in all standard web browsers. *WebGL* (natively supported by most browsers) is required for the in-line preview and manipulation of protein three-dimensional structures. The web site includes a detailed help section and usage examples.

Contact: Florencio Pazos (pazos@cnb.csic.es)

1 INTRODUCTION

The computational prediction of functional features for newly sequenced proteins is a very active field of research due to the pace at which these raw sequences are obtained and the difficulties associated with the experimental functional characterization.

The most widely used strategy for function prediction is to transfer to the query sequence the functional features (e.g. global function and functional residues) available for an homologous protein in a given database or annotation resource (Rentzsch and Orengo, 2009; Valencia, 2005). Due to technical and historical reasons, most of these resources associate functional descriptors to whole proteins, instead of individual domains. Nevertheless, domains are the evolutionary, structural and functional units of proteins and, at least in the case of "molecular functions", these can be individually assigned to particular domains or groups of domains within protein chains. The functional independence of protein domains is exemplified by the fact that many domains found in multi-domain proteins also exist in isolation, as independent proteins, in other organisms (Marcotte, et al., 1999). Although associating molecular functions to whole chains without distinguishing the particular domain(s) responsible for them is not an issue for many applications, it can lead to problems in other cases (Lopez and Pazos, 2009). Since molecular functions are generally associated with whole chains in the annotation resources, function prediction methods/servers, based on matching against these annotations, predict molecular function at the whole-chain level.

Only recently the first functional annotations at the domain level started to appear (de Lima Morais, et al., 2011; Lopez and Pazos, 2009), highlighting the importance of associating molecular functions to domains and the consequences of not doing so in particular cases. Accordingly, prediction methods adapted to these resources and intended for predicting functions at the domain level are required. We have developed one of these methods and showed that its performance in assigning function is higher that a traditional sequence-based search (Lopez and Pazos, 2013). The method is based on a resource of Gene Ontology (Ashburner, et al., 2000) molecular function annotations (GO:MF) at the structural domain level (Lopez and Pazos, 2009). In short, profiles are build for the structural domains sharing the same GO:MF term and the same fold (so that they can be structurally aligned). A significant match of a particular region of a query sequence against one of these profiles can be interpreted as a concomitant prediction of fold and GO:MF function for the corresponding domain. Additionally, the conserved positions in these profiles can be interpreted as "functional sites" and consequently transferred to the aligned positions of the query sequence so as to obtain clues on possible functionally important positions. For full details on the methodology and its evaluation see (Lopez and Pazos, 2013).

Here we present COPRED, a web server which allows any user to access this method and generate predictions using, in the simplest case, only the query sequence as input. The predictions can be inspected in a graphical interactive interface, and downloaded in a number of standard formats.

2 FUNCTIONALITY AND INTERFACE

Figure 1 shows some representative screenshots of COPRED's web interface. On the input page, the user can upload the amino acid sequence of the query protein as only input, although an "advanced options" button allows expanding the input form so as to modify some parameters of the method. There is also a "fill with example" button to test the server right away with the example described in the help page.

After a successful run, which normally takes 5-10 seconds, all the results can be accessed from a single page with different panels that can be expanded/collapsed. The first row represents the query sequence itself (blue in Figure 1) and expanding it allows retriev-

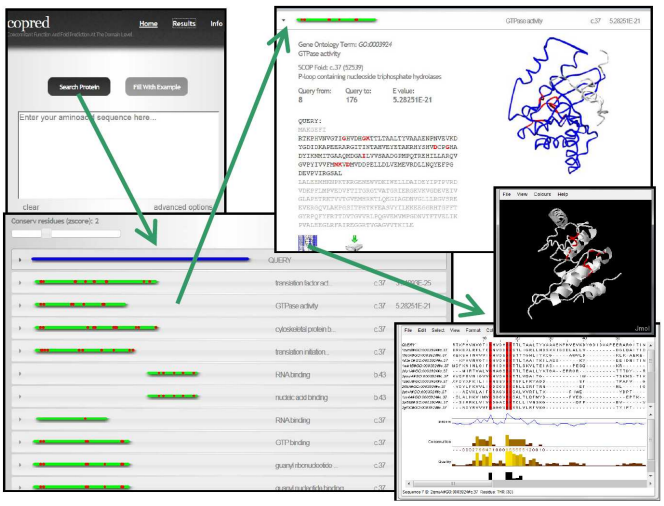


Figure 1. Screenshots of COPRED web interface. Top-left: input page, with links for accessing the help page, inserting an example, and for retrieving previous jobs. Bottom-left: unexpanded list of profiles (green) matching different regions of the query sequence (blue). Top-right: expansion of the "GTPase activity (function) – c.37 (fold)" hit, with details on the region of the query sequence matching that profile, the predicted functional sites, an in-line interactive 3D view of a representative structure of the profile, and links for opening a Javlview applet with the multiple sequence alignment associated to the profile, a Jmol applet with the implicit 3D model and the PDB file with that model (bottom-right).

ing some data on the input, including a job ID that can be later used to recover the results of previous runs (via the "Results" option on the input page). The results basically consist of a list of profiles matching different regions of the query sequence, including information on the fold and GO:MF term represented by each of them (Figure 1). For each hit, a graphical representation of the region of the query sequence matched against the profile is shown, using a color scale (from green to black) to represent the hit's significance (E-value). Consequently, this unexpanded list of hits provides a first overview of the domain composition of the query sequence and the possible folds/functions of its domains. In the example shown in Figure 1, the results are clearly pointing to the presence of two domains: an N-terminal one, associated with the c.37 SCOP fold ("P-loop nucleotide phosphate hydrolases") and GTP-related GO:MF terms, and the middle domain associated with the b.43 fold ("Reductase/isomerase/elongation factor common domain") and RNA-binding-related GO:MF terms. These features are in agreement with the domain composition, structural folds and functions of this example query protein (Elongation factor Tu). Additionally, a set of red dots represent the predicted functional sites transferred from the conserved positions of the corresponding profile (according to a conservation threshold controlled by the scroll-bar at the top of the page).

Expanding a given item of the hit list provides additional information on the corresponding profile match (Figure 1), such as detailed information on the GO:MF term and SCOP fold, with links to the corresponding databases. The sequence of the query protein is also shown highlighting the region (domain) matching that particular profile and the transferred functional residues. There is also an interactive representation of the three-dimensional structure of a representative member of that profile. This can be manipulated, e.g. zoomed, rotated, etc. In this 3D view, the region aligned against the query and the conserved residues are highlighted. There is also a link for opening a Jalview (Waterhouse, et al., 2009) applet showing the multiple sequence alignment associated with the profile, which also includes the query sequence (Figure 1). The implicit three-dimensional model for that region (domain) of the query protein is also shown in a Jmol applet (www.jmol.org). Conserved positions are highlighted in both representations. This implicit 3D model, which can be also downloaded as a standard PDB file from another link, contains the backbone atoms of a representative template from the profile with the residues renamed to those of the query sequence.

3 CONCLUSION

Most proteins, especially in eukaryotic organisms comprise multiple domains (Apic et al, 2001). It is important to adapt the function prediction workflows (annotation databases, methods, etc), mainly developed under a "1-chain-1-function" paradigm, to this reality. Although many existing resources are slowly starting to adopt a "domain-centric" view, COPRED is the first server specifically devoted to this task. The main advantages of this server are its ease of use and the fact that it provides a first glimpse of the domain structural and functional features of the problem sequence in a concomitant way. The main limitation of the COPRED server at this point is the relatively small scale of the profile database it uses, compared with other resources. Nevertheless, these profiles can be built from any database of functionally annotated structural domains. Therefore, the server can be in principle expanded with larger sets domain annotations.

ACKNOWLEDGEMENTS

We want to thank David S. León (Plant Molecular Genetics Dep., CNB) for testing the system and providing feedback, and the members of the Computational Systems Biology Group (CNB-CSIC) for comments and suggestions.
Funding: This work was partially funded by the BIO2010-22109 project of the Spanish Ministry of Science and Innovation.

REFERENCES

Apic, G., Gough, J. and Teichmann, S.A. (2001). Domain combinations in archaeal, aubacterial and eukaryotic proteomes. *J Mol Biol.* **310,** 311-325.
Ashburner, M., Ball, C.A., Blake, J.A., *et al* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, *Nat Genet*, **25**, 25-29.
de Lima Morais, D.A., Fang, H., Rackham, O.J., Wilson, D., Pethica, R., Chothia, C. and Gough, J. (2011) SUPERFAMILY 1.75 including a domain-centric gene ontology method, *Nucleic Acids Res*, **39**, D427-434.
Lopez, D. and Pazos, F. (2009) Gene Ontology functional annotations at the structural domain level, *Proteins*, **76**, 598-607.
Lopez, D. and Pazos, F. (2013) Concomitant prediction of function and fold at the domain level with GO-based profiles, *BMC Bioinformatics*, **14**, S12.
Marcotte, E.M., Pellegrini, M., Ho-Leung, N., Rice, D.W., Yeates, T.O. and Eisenberg, D. (1999) Detecting protein function and protein-protein interactions from genome sequences., *Science*, **285**, 751-753.
Rentzsch, R. and Orengo, C. (2009) Protein function prediction - the power of multiplicity, *Trends Biotech*, **27**, 210-219.
Valencia, A. (2005) Automatic annotation of protein function, *Curr Opin Struct Biol*, **15**, 267-274.
Waterhouse, A.M., Procter, J.B., Martin, D.M., Clamp, M. and Barton, G.J. (2009) Jalview Version 2--a multiple sequence alignment editor and analysis workbench., *Bioinformatics*, **25**, 1189-1191.

Bibliografía

Agashe, D., Martinez-Gomez, N. C., Drummond, D. A. y Marx, C. J. Good Codons, Bad Transcript: Large Reductions in Gene Expression and Fitness Arising from Synonymous Mutations in a Key Enzyme. *Molecular biology and evolution*, 2012. ISSN 1537-1719.

Akashi, H. y Eyre-Walker, A. Translational selection and molecular evolution. *Current opinion in genetics & development*, vol. 8(6), páginas 688–93, 1998. ISSN 0959-437X.

Almonacid, D. E., Yera, E. R., Mitchell, J. B. O. y Babbitt, P. C. Quantitative comparison of catalytic mechanisms and overall reactions in convergently evolved enzymes: implications for classification of enzyme function. *PLoS computational biology*, vol. 6(3), página e1000700, 2010. ISSN 1553-7358.

Altschul, S. F., Madden, T. L., Schäffer, a. a., Zhang, J., Zhang, Z., Miller, W. y Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, vol. 25(17), páginas 3389–402, 1997. ISSN 0305-1048.

Andreeva, A., Howorth, D., Brenner, S. E., Hubbard, T. J. P., Chothia, C. y Murzin, A. G. SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic acids research*, vol. 32(Database issue), páginas D226–9, 2004. ISSN 1362-4962.

Armon, A., Graur, D. y Ben-Tal, N. ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *Journal of molecular biology*, vol. 307(1), páginas 447–63, 2001. ISSN 0022-2836.

Attwood, T. K., Flower, D. R., Lewis, A. P., Mabey, J. E., Morgan, S. R., Scordis, P., Selley, J. N. y Wright, W. PRINTS prepares for the new millennium. *Nucleic acids research*, vol. 27(1), páginas 220–5, 1999. ISSN 0305-1048.

BARRETT, T., WILHITE, S. E., LEDOUX, P., EVANGELISTA, C., KIM, I. F., TOMASHEVSKY, M., MARSHALL, K. A., PHILLIPPY, K. H., SHERMAN, P. M., HOLKO, M., YEFANOV, A., LEE, H., ZHANG, N., ROBERTSON, C. L., SEROVA, N., DAVIS, S. y SOBOLEVA, A. NCBI GEO: archive for functional genomics data sets–update. *Nucleic acids research*, vol. 41(Database issue), páginas D991–5, 2013. ISSN 1362-4962.

BARTLETT, G. J., BORKAKOTI, N. y THORNTON, J. M. Catalysing new reactions during evolution: economy of residues and mechanism. *Journal of molecular biology*, vol. 331(4), páginas 829–60, 2003. ISSN 0022-2836.

BASHTON, M. y CHOTHIA, C. The generation of new protein functions by the combination of domains. *Structure (London, England : 1993)*, vol. 15(1), páginas 85–99, 2007. ISSN 0969-2126.

BATEMAN, A., BIRNEY, E., DURBIN, R., EDDY, S. R., FINN, R. D. y SONNHAMMER, E. L. Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. *Nucleic acids research*, vol. 27(1), páginas 260–2, 1999. ISSN 0305-1048.

BATEMAN, O. A., PURKISS, A. G., VAN MONTFORT, R., SLINGSBY, C., GRAHAM, C. y WISTOW, G. Crystal structure of eta-crystallin: adaptation of a class 1 aldehyde dehydrogenase for a new role in the eye lens. *Biochemistry*, vol. 42(15), páginas 4349–56, 2003. ISSN 0006-2960.

BERMAN, H. M., WESTBROOK, J., FENG, Z., GILLILAND, G., BHAT, T. N., WEISSIG, H., SHINDYALOV, I. N. y BOURNE, P. E. The Protein Data Bank. *Nucleic acids research*, vol. 28(1), páginas 235–42, 2000. ISSN 0305-1048.

BOUTSELAKIS, H. E-MSD: the European Bioinformatics Institute Macromolecular Structure Database. *Nucleic Acids Research*, vol. 31(1), páginas 458–462, 2003. ISSN 13624962.

BREST, P., LAPAQUETTE, P., SOUIDI, M., LEBRIGAND, K., CESARO, A., VOURET-CRAVIARI, V., MARI, B., BARBRY, P., MOSNIER, J.-F., HÉBUTERNE, X., HAREL-BELLAN, A., MOGRABI, B., DARFEUILLE-MICHAUD, A. y HOFMAN, P. A synonymous variant in IRGM alters a binding site for miR-196 and causes deregulation of IRGM-dependent xenophagy in Crohn's disease. *Nature genetics*, vol. 43(3), páginas 242–5, 2011. ISSN 1546-1718.

BRYLINSKI, M. y SKOLNICK, J. A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. *Proceedings*

of the National Academy of Sciences of the United States of America, vol. 105(1), páginas 129–34, 2008. ISSN 1091-6490.

CHAGOYEN, M. y PAZOS, F. Quantifying the biological significance of gene ontology biological processes–implications for the analysis of systems-wide data. *Bioinformatics (Oxford, England)*, vol. 26(3), páginas 378–84, 2010. ISSN 1367-4811.

CHANDONIA, J.-M., HON, G., WALKER, N. S., LO CONTE, L., KOEHL, P., LEVITT, M. y BRENNER, S. E. The ASTRAL Compendium in 2004. *Nucleic acids research*, vol. 32(Database issue), páginas D189–92, 2004. ISSN 1362-4962.

CHARTIER, M., GAUDREAULT, F. y NAJMANOVICH, R. Large scale analysis of conserved rare codon clusters suggests an involvement in co-translational molecular recognition events. *Bioinformatics*, 2012. ISSN 1367-4803.

CHURSOV, A., WALTER, M. C., SCHMIDT, T., MIRONOV, A., SHNEIDER, A. y FRISHMAN, D. Sequence-structure relationships in yeast mRNAs. *Nucleic acids research*, vol. 40(3), páginas 956–62, 2012. ISSN 1362-4962.

DANA, A. y TULLER, T. Determinants of Translation Elongation Speed and Ribosomal Profiling Biases in Mouse Embryonic Stem Cells. *PLoS Computational Biology*, vol. 8(11), página e1002755, 2012. ISSN 1553-7358.

DE LIMA MORAIS, D. A., FANG, H., RACKHAM, O. J. L., WILSON, D., PETHICA, R., CHOTHIA, C. y GOUGH, J. SUPERFAMILY 1.75 including a domain-centric gene ontology method. *Nucleic acids research*, vol. 39(Database issue), páginas D427–34, 2011. ISSN 1362-4962.

DESVEAUX, D., SINGER, A. U. y DANGL, J. L. Type III effector proteins: doppelgangers of bacterial virulence. *Current opinion in plant biology*, vol. 9(4), páginas 376–82, 2006. ISSN 1369-5266.

DEVOS, D. y VALENCIA, A. Practical limits of function prediction. *Proteins*, vol. 41(1), páginas 98–107, 2000. ISSN 0887-3585.

DIMMER, E. C., HUNTLEY, R. P., ALAM-FARUQUE, Y., SAWFORD, T., O'DONOVAN, C., MARTIN, M. J., BELY, B., BROWNE, P., MUN CHAN, W., EBERHARDT, R., GARDNER, M., LAIHO, K., LEGGE, D., MAGRANE, M., PICHLER, K., POGGIOLI, D., SEHRA, H., AUCHINCLOSS, A., AXELSEN, K., BLATTER, M.-C., BOUTET, E., BRACONI-QUINTAJE, S., BREUZA, L., BRIDGE, A., COUDERT, E., ESTREICHER, A.,

FAMIGLIETTI, L., FERRO-ROJAS, S., FEUERMANN, M., GOS, A., GRUAZ-GUMOWSKI, N., HINZ, U., HULO, C., JAMES, J., JIMENEZ, S., JUNGO, F., KELLER, G., LEMERCIER, P., LIEBERHERR, D., MASSON, P., MOINAT, M., PEDRUZZI, I., POUX, S., RIVOIRE, C., ROECHERT, B., SCHNEIDER, M., STUTZ, A., SUNDARAM, S., TOGNOLLI, M., BOUGUELERET, L., ARGOUD-PUY, G., CUSIN, I., DUEK-ROGGLI, P., XENARIOS, I. y APWEILER, R. The UniProt-GO Annotation database in 2011. *Nucleic acids research*, vol. 40(Database issue), páginas D565–70, 2012. ISSN 1362-4962.

DONG, H., NILSSON, L. y KURLAND, C. G. Co-variation of tRNA abundance and codon usage in Escherichia coli at different growth rates. *Journal of molecular biology*, vol. 260(5), páginas 649–63, 1996. ISSN 0022-2836.

DOOLITTLE, R. F. y BORK, P. Evolutionarily mobile modules in proteins. *Scientific American*, vol. 269(4), páginas 50–6, 1993. ISSN 0036-8733.

DOSZTÁNYI, Z., CSIZMÓK, V., TOMPA, P. y SIMON, I. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *Journal of molecular biology*, vol. 347(4), páginas 827–39, 2005. ISSN 0022-2836.

ENGELHARDT, B. E., JORDAN, M. I., MURATORE, K. E. y BRENNER, S. E. Protein molecular function prediction by Bayesian phylogenomics. *PLoS computational biology*, vol. 1(5), página e45, 2005. ISSN 1553-7358.

ERDIN, S., LISEWSKI, A. M. y LICHTARGE, O. Protein function prediction: towards integration of similarity metrics. *Current opinion in structural biology*, vol. 21(2), páginas 180–8, 2011. ISSN 1879-033X.

ERNST, J. Codon usage and gene expression. *Trends in Biotechnology*, vol. 6(8), páginas 196–199, 1988. ISSN 01677799.

FALQUET, L., PAGNI, M., BUCHER, P., HULO, N., SIGRIST, C. J. A., HOFMANN, K. y BAIROCH, A. The PROSITE database, its status in 2002. *Nucleic acids research*, vol. 30(1), páginas 235–8, 2002. ISSN 1362-4962.

FITCH, W. M. Homology a personal view on some of the problems. *Trends in genetics : TIG*, vol. 16(5), páginas 227–31, 2000. ISSN 0168-9525.

FRENKEL-MORGENSTERN, M., DANON, T., CHRISTIAN, T., IGARASHI, T., COHEN, L., HOU, Y.-M. y JENSEN, L. J. Genes adopt non-optimal

codon usage to generate cell cycle-dependent oscillations in protein levels. *Molecular systems biology*, vol. 8, página 572, 2012. ISSN 1744-4292.

FRISHMAN, D. Protein annotation at genomic scale: the current status. *Chemical reviews*, vol. 107(8), páginas 3448–66, 2007. ISSN 0009-2665.

GILKS, W. R., AUDIT, B., DE ANGELIS, D., TSOKA, S. y OUZOUNIS, C. A. Modeling the percolation of annotation errors in a database of protein sequences. *Bioinformatics (Oxford, England)*, vol. 18(12), páginas 1641–9, 2002. ISSN 1367-4803.

GLASER, F., ROSENBERG, Y., KESSEL, A., PUPKO, T. y BEN-TAL, N. The ConSurf-HSSP database: the mapping of evolutionary conservation among homologs onto PDB structures. *Proteins*, vol. 58(3), páginas 610–7, 2005. ISSN 1097-0134.

GÖTZ, S., GARCÍA-GÓMEZ, J. M., TEROL, J., WILLIAMS, T. D., NAGARAJ, S. H., NUEDA, M. J., ROBLES, M., TALÓN, M., DOPAZO, J. y CONESA, A. High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic acids research*, vol. 36(10), páginas 3420–35, 2008. ISSN 1362-4962.

HARRIS, M. A., CLARK, J., IRELAND, A., LOMAX, J., ASHBURNER, M., FOULGER, R., EILBECK, K., LEWIS, S., MARSHALL, B., MUNGALL, C., RICHTER, J., RUBIN, G. M., BLAKE, J. A., BULT, C., DOLAN, M., DRABKIN, H., EPPIG, J. T., HILL, D. P., NI, L., RINGWALD, M., BALAKRISHNAN, R., CHERRY, J. M., CHRISTIE, K. R., COSTANZO, M. C., DWIGHT, S. S., ENGEL, S., FISK, D. G., HIRSCHMAN, J. E., HONG, E. L., NASH, R. S., SETHURAMAN, A., THEESFELD, C. L., BOTSTEIN, D., DOLINSKI, K., FEIERBACH, B., BERARDINI, T., MUNDODI, S., RHEE, S. Y., APWEILER, R., BARRELL, D., CAMON, E., DIMMER, E., LEE, V., CHISHOLM, R., GAUDET, P., KIBBE, W., KISHORE, R., SCHWARZ, E. M., STERNBERG, P., GWINN, M., HANNICK, L., WORTMAN, J., BERRIMAN, M., WOOD, V., DE LA CRUZ, N., TONELLATO, P., JAISWAL, P., SEIGFRIED, T. y WHITE, R. The Gene Ontology (GO) database and informatics resource. *Nucleic acids research*, vol. 32(Database issue), páginas D258–61, 2004. ISSN 1362-4962.

HAWKINS, T., LUBAN, S. y KIHARA, D. Enhanced automated function prediction using distantly related sequences and contextual association by PFP. *Protein Science*, páginas 1550–1556, 2009.

HENIKOFF, J. G. y HENIKOFF, S. Blocks database and its applications. *Methods in enzymology*, vol. 266, páginas 88–105, 1996. ISSN 0076-6879.

HOLLIDAY, J. D., HU, C.-Y. y WILLETT, P. Grouping of coefficients for the calculation of inter-molecular similarity and dissimilarity using 2D fragment bit-strings. *Combinatorial chemistry & high throughput screening*, vol. 5(2), páginas 155–66, 2002. ISSN 1386-2073.

HOLM, L. y PARK, J. DaliLite workbench for protein structure comparison. *Bioinformatics (Oxford, England)*, vol. 16(6), páginas 566–7, 2000. ISSN 1367-4803.

HOLM, L. y ROSENSTRÖM, P. Dali server: conservation mapping in 3D. *Nucleic acids research*, vol. 38(Web Server issue), páginas W545–9, 2010. ISSN 1362-4962.

HOLM, L. y SANDER, C. Dali: a network tool for protein structure comparison. *Trends in biochemical sciences*, vol. 20(11), páginas 478–80, 1995. ISSN 0968-0004.

HUNTER, S., JONES, P., MITCHELL, A., APWEILER, R., ATTWOOD, T. K., BATEMAN, A., BERNARD, T., BINNS, D., BORK, P., BURGE, S., DE CASTRO, E., COGGILL, P., CORBETT, M., DAS, U., DAUGHERTY, L., DUQUENNE, L., FINN, R. D., FRASER, M., GOUGH, J., HAFT, D., HULO, N., KAHN, D., KELLY, E., LETUNIC, I., LONSDALE, D., LOPEZ, R., MADERA, M., MASLEN, J., MCANULLA, C., MCDOWALL, J., MCMENAMIN, C., MI, H., MUTOWO-MUELLENET, P., MULDER, N., NATALE, D., ORENGO, C., PESSEAT, S., PUNTA, M., QUINN, A. F., RIVOIRE, C., SANGRADOR-VEGAS, A., SELENGUT, J. D., SIGRIST, C. J. A., SCHEREMETJEW, M., TATE, J., THIMMAJANARTHANAN, M., THOMAS, P. D., WU, C. H., YEATS, C. y YONG, S.-Y. InterPro in 2011: new developments in the family and domain prediction database. *Nucleic acids research*, vol. 40(Database issue), páginas D306–12, 2012. ISSN 1362-4962.

JEFFERY, C. J. Moonlighting proteins. *Trends in biochemical sciences*, vol. 24(1), páginas 8–11, 1999. ISSN 0968-0004.

JEFFERY, C. J. Molecular mechanisms for multitasking: recent crystal structures of moonlighting proteins. *Current opinion in structural biology*, vol. 14(6), páginas 663–8, 2004. ISSN 0959-440X.

JENSEN, L. J., GUPTA, R., BLOM, N., DEVOS, D., TAMAMES, J., KESMIR, C., NIELSEN, H., STAERFELDT, H. H., RAPACKI, K., WORKMAN, C., ANDERSEN, C. A. F., KNUDSEN, S., KROGH, A., VALENCIA, A. y BRUNAK, S. Prediction of human protein function from post-translational

modifications and localization features. *Journal of molecular biology*, vol. 319(5), páginas 1257–65, 2002. ISSN 0022-2836.

JMOL. Visor Java de código abierto para estructuras químicas en tres dimensiones. http://www.jmol.org/. 2013.

JQUERY. Librería javascript. http://jquery.com. 2013.

KERTESZ, M., WAN, Y., MAZOR, E., RINN, J. L., NUTTER, R. C., CHANG, H. Y. y SEGAL, E. Genome-wide measurement of RNA secondary structure in yeast. *Nature*, vol. 467(7311), páginas 103–7, 2010. ISSN 1476-4687.

KIMCHI-SARFATY, C., OH, J. M., KIM, I.-W., SAUNA, Z. E., CALCAGNO, A. M., AMBUDKAR, S. V. y GOTTESMAN, M. M. A "silent"polymorphism in the MDR1 gene changes substrate specificity. *Science (New York, N.Y.)*, vol. 315(5811), páginas 525–8, 2007. ISSN 1095-9203.

KINOSHITA, K., FURUI, J. y NAKAMURA, H. Identification of protein functions from a molecular surface database, eF-site. *Journal of structural and functional genomics*, vol. 2(1), páginas 9–22, 2002. ISSN 1345-711X.

KOMAR, A. A., LESNIK, T. y REISS, C. Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation. *FEBS letters*, vol. 462(3), páginas 387–91, 1999. ISSN 0014-5793.

KUDLA, G., MURRAY, A. W., TOLLERVEY, D. y PLOTKIN, J. B. Coding-sequence determinants of gene expression in Escherichia coli. *Science (New York, N.Y.)*, vol. 324(5924), páginas 255–8, 2009. ISSN 1095-9203.

LASKOWSKI, R. A. SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *Journal of molecular graphics*, vol. 13(5), páginas 323–30, 307–8, 1995. ISSN 0263-7855.

LASKOWSKI, R. A., CHISTYAKOV, V. V. y THORNTON, J. M. PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids. *Nucleic acids research*, vol. 33(Database issue), páginas D266–8, 2005. ISSN 1362-4962.

LEE, D., REDFERN, O. y ORENGO, C. Predicting protein function from sequence and structure. *Nature reviews. Molecular cell biology*, vol. 8(12), páginas 995–1005, 2007. ISSN 1471-0080.

LEE, Y., ZHOU, T., TARTAGLIA, G. G., VENDRUSCOLO, M. y WILKE, C. O. Translationally optimal codons associate with aggregation-prone sites in proteins. *Proteomics*, vol. 10(23), páginas 4163–71, 2010. ISSN 1615-9861.

LEVITT, M. y GERSTEIN, M. A unified statistical framework for sequence comparison and structure comparison. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95(11), páginas 5913–20, 1998. ISSN 0027-8424.

LI, G.-W., OH, E. y WEISSMAN, J. S. The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature*, 2012. ISSN 1476-4687.

LINIAL, M. How incorrect annotations evolve–the case of short ORFs. *Trends in biotechnology*, vol. 21(7), páginas 298–300, 2003. ISSN 0167-7799.

LOPEZ, D. y PAZOS, F. Gene ontology functional annotations at the structural domain level. *Proteins*, vol. 76(3), páginas 598–607, 2009. ISSN 1097-0134.

MAGADUM, S., BANERJEE, U., MURUGAN, P., GANGAPUR, D. y RAVIKESAVAN, R. Gene duplication as a major force in evolution. *Journal of genetics*, vol. 92(1), páginas 155–61, 2013. ISSN 0973-7731.

MARCOTTE, E. M. Detecting Protein Function and Protein-Protein Interactions from Genome Sequences. *Science*, vol. 285(5428), páginas 751–753, 1999. ISSN 00368075.

MARTIN, D. M. A., BERRIMAN, M. y BARTON, G. J. GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes. *BMC bioinformatics*, vol. 5, página 178, 2004. ISSN 1471-2105.

MEDLINEPLUS. Bethesda (MD): National Library of Medicine (US). 2005.

MÉSZÁROS, B., SIMON, I. y DOSZTÁNYI, Z. Prediction of protein binding regions in disordered proteins. *PLoS computational biology*, vol. 5(5), página e1000376, 2009. ISSN 1553-7358.

MICHALSKY, E., DUNKEL, M., GOEDE, A. y PREISSNER, R. SuperLigands - a database of ligand structures derived from the Protein Data Bank. *BMC bioinformatics*, vol. 6, página 122, 2005. ISSN 1471-2105.

NOTREDAME, C. Recent evolutions of multiple sequence alignment algorithms. *PLoS computational biology*, vol. 3(8), página e123, 2007. ISSN 1553-7358.

NOTREDAME, C., HIGGINS, D. G. y HERINGA, J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of molecular biology*, vol. 302(1), páginas 205–17, 2000. ISSN 0022-2836.

PAL, D. y EISENBERG, D. Inference of protein function from protein structure. *Structure (London, England : 1993)*, vol. 13(1), páginas 121–30, 2005. ISSN 0969-2126.

PEARL, F., TODD, A., SILLITOE, I., DIBLEY, M., REDFERN, O., LEWIS, T., BENNETT, C., MARSDEN, R., GRANT, A., LEE, D., AKPOR, A., MAIBAUM, M., HARRISON, A., DALLMAN, T., REEVES, G., DIBOUN, I., ADDOU, S., LISE, S., JOHNSTON, C., SILLERO, A., THORNTON, J. y ORENGO, C. The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. *Nucleic acids research*, vol. 33(Database issue), páginas D247–51, 2005. ISSN 1362-4962.

PIETROSEMOLI, N., LOPEZ, D., SEGURA-CABRERA, A. y PAZOS, F. Computational Prediction of Important Regions in Protein Sequences [Life Sciences]. *IEEE Signal Processing Magazine*, vol. 29(6), páginas 143–147, 2012. ISSN 1053-5888.

POLACCO, B. J. y BABBITT, P. C. Automated discovery of 3D motifs for protein function annotation. *Bioinformatics (Oxford, England)*, vol. 22(6), páginas 723–30, 2006. ISSN 1367-4803.

POWER, P. M., JONES, R. A., BEACHAM, I. R., BUCHOLTZ, C. y JENNINGS, M. P. Whole genome analysis reveals a high incidence of non-optimal codons in secretory signal sequences of Escherichia coli. *Biochemical and biophysical research communications*, vol. 322(3), páginas 1038–44, 2004. ISSN 0006-291X.

PUNTA, M., COGGILL, P. C., EBERHARDT, R. Y., MISTRY, J., TATE, J., BOURSNELL, C., PANG, N., FORSLUND, K., CERIC, G., CLEMENTS, J., HEGER, A., HOLM, L., SONNHAMMER, E. L. L., EDDY, S. R., BATEMAN, A. y FINN, R. D. The Pfam protein families database. *Nucleic acids research*, vol. 40(Database issue), páginas D290–301, 2012. ISSN 1362-4962.

PUNTA, M. y OFRAN, Y. The rough guide to in silico function prediction, or how to use sequence and structure information to predict protein function. *PLoS computational biology*, vol. 4(10), página e1000160, 2008. ISSN 1553-7358.

PUNTERVOLL, P. L., LINDING, R., GEMÜND, C., CHABANIS-DAVIDSON, S., MATTINGSDAL, M., CAMERON, S., MARTIN, D. M. A., AUSIELLO, G., BRANNETTI, B., COSTANTINI, A., FERRÈ, F., MASELLI, V., VIA, A., CESARENI, G., DIELLA, F., SUPERTI-FURGA, G., WYRWICZ, L., RAMU, C., MCGUIGAN, C., GUDAVALLI, R., LETUNIC, I., BORK, P., RYCHLEWSKI, L., KÜSTER, B., HELMER-CITTERICH, M., HUNTER, W. N., AASLAND, R. y GIBSON, T. J. ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic acids research*, vol. 31(13), páginas 3625–30, 2003. ISSN 1362-4962.

QUEVILLON, E., SILVENTOINEN, V., PILLAI, S., HARTE, N., MULDER, N., APWEILER, R. y LOPEZ, R. InterProScan: protein domains identifier. *Nucleic acids research*, vol. 33(Web Server issue), páginas W116–20, 2005. ISSN 1362-4962.

RAES, J., HARRINGTON, E. D., SINGH, A. H. y BORK, P. Protein function space: viewing the limits or limited by our view? *Current opinion in structural biology*, vol. 17(3), páginas 362–9, 2007. ISSN 0959-440X.

RAUSELL, A., JUAN, D., PAZOS, F. y VALENCIA, A. Protein interactions and ligand binding: from protein subfamilies to functional specificity. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107(5), páginas 1995–2000, 2010. ISSN 1091-6490.

REEVES, G. A., EILBECK, K., MAGRANE, M., O'DONOVAN, C., MONTECCHI-PALAZZI, L., HARRIS, M. A., ORCHARD, S., JIMENEZ, R. C., PRLIC, A., HUBBARD, T. J. P., HERMJAKOB, H. y THORNTON, J. M. The Protein Feature Ontology: a tool for the unification of protein feature annotations. *Bioinformatics (Oxford, England)*, vol. 24(23), páginas 2767–72, 2008. ISSN 1367-4811.

RILEY, M. Searchlight on domains. *Structure (London, England : 1993)*, vol. 15(1), páginas 1–2, 2007. ISSN 0969-2126.

SAUNDERS, R. y DEANE, C. M. Synonymous codon usage influences the local protein structure observed. *Nucleic acids research*, vol. 38(19), páginas 6719–28, 2010. ISSN 1362-4962.

Schrödinger, L. The PyMOL Molecular Graphics System, Version 1.3r1, 2010.

Schuster, S. Next-generation sequencing transforms today's biology. *Nature*, vol. 5(1), páginas 16–18, 2007.

Shabalina, S. a., Spiridonov, N. a. y Kashina, A. Sounds of silence: synonymous nucleotides as a key to biological regulation and complexity. *Nucleic acids research*, vol. 41(4), páginas 2073–2094, 2013. ISSN 1362-4962.

Shulman-Peleg, A., Shatsky, M., Nussinov, R. y Wolfson, H. J. MultiBind and MAPPIS: webservers for multiple alignment of protein 3D-binding sites and their interactions. *Nucleic acids research*, vol. 36(Web Server issue), páginas W260–4, 2008. ISSN 1362-4962.

Sorensen, M. A. Charging levels of four tRNA species in Escherichia coli Rel(+) and Rel(-) strains during amino acid starvation: a simple model for the effect of ppGpp on translational accuracy. *Journal of molecular biology*, vol. 307(3), páginas 785–98, 2001. ISSN 0022-2836.

Stebbins, C. E. y Galán, J. E. Structural mimicry in bacterial virulence. *Nature*, vol. 412(6848), páginas 701–5, 2001. ISSN 0028-0836.

Tatusov, R. L., Koonin, E. V. y Lipman, D. J. A genomic perspective on protein families. *Science (New York, N.Y.)*, vol. 278(5338), páginas 631–7, 1997. ISSN 0036-8075.

Tavares, G. WebGL Fundamentals. *HTML5 Rocks*, 2012.

Thanaraj, T. a. y Argos, P. Ribosome-mediated translational pause and protein domain organization. *Protein science : a publication of the Protein Society*, vol. 5(8), páginas 1594–612, 1996. ISSN 0961-8368.

Theissen, G. Secret life of genes. *Nature*, vol. 415(6873), página 741, 2002. ISSN 0028-0836.

Threejs. Librería javascript 3D. http://threejs.org/. 2013.

Tottey, S., Waldron, K. J., Firbank, S. J., Reale, B., Bessant, C., Sato, K., Cheek, T. R., Gray, J., Banfield, M. J., Dennison, C. y Robinson, N. J. Protein-folding location can regulate manganese-binding versus copper- or zinc-binding. *Nature*, vol. 455(7216), páginas 1138–42, 2008. ISSN 1476-4687.

TSAI, C.-J., SAUNA, Z. E., KIMCHI-SARFATY, C., AMBUDKAR, S. V., GOTTESMAN, M. M. y NUSSINOV, R. Synonymous mutations and ribosome stalling can lead to altered folding pathways and distinct minima. *Journal of molecular biology*, vol. 383(2), páginas 281–91, 2008. ISSN 1089-8638.

TSENG, Y. Y., DUNDAS, J. y LIANG, J. Predicting protein function and binding profile via matching of local evolutionary and geometric surface patterns. *Journal of molecular biology*, vol. 387(2), páginas 451–64, 2009. ISSN 1089-8638.

TULLER, T., CARMI, A., VESTSIGIAN, K., NAVON, S., DORFAN, Y., ZABORSKE, J., PAN, T., DAHAN, O., FURMAN, I. y PILPEL, Y. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell*, vol. 141(2), páginas 344–54, 2010. ISSN 1097-4172.

UNIPROT CONSORTIUM. The Universal Protein Resource (UniProt) in 2010. *Nucleic acids research*, vol. 38(Database issue), páginas D142–8, 2010. ISSN 1362-4962.

VALDAR, W. S. J. Scoring residue conservation. *Proteins*, vol. 48(2), páginas 227–41, 2002. ISSN 1097-0134.

VALENCIA, A. Automatic annotation of protein function. *Current opinion in structural biology*, vol. 15(3), páginas 267–74, 2005. ISSN 0959-440X.

VARENNE, S., BUC, J., LLOUBES, R. y LAZDUNSKI, C. Translation is a non-uniform process. *Journal of Molecular Biology*, vol. 180(3), páginas 549–576, 1984. ISSN 00222836.

WALLACE, I. M., BLACKSHIELDS, G. y HIGGINS, D. G. Multiple sequence alignments. *Current opinion in structural biology*, vol. 15(3), páginas 261–6, 2005. ISSN 0959-440X.

WATERHOUSE, A. M., PROCTER, J. B., MARTIN, D. M. A., CLAMP, M. y BARTON, G. J. Jalview Version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics (Oxford, England)*, vol. 25(9), páginas 1189–91, 2009. ISSN 1367-4811.

WATSON, J. D., SANDERSON, S., EZERSKY, A., SAVCHENKO, A., EDWARDS, A., ORENGO, C., JOACHIMIAK, A., LASKOWSKI, R. A. y THORNTON, J. M. Towards fully automated structure-based function prediction in structural genomics: a case study. *Journal of molecular biology*, vol. 367(5), páginas 1511–22, 2007. ISSN 0022-2836.

WETLAUFER, D. B. Nucleation, rapid folding, and globular intrachain regions in proteins. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 70(3), páginas 697–701, 1973. ISSN 0027-8424.

WHISSTOCK, J. C. y LESK, A. M. Prediction of protein function from protein sequence and structure. *Quarterly reviews of biophysics*, vol. 36(3), páginas 307–40, 2003. ISSN 0033-5835.

WILCOX, C., HU, J. S. y OLSON, E. N. Acylation of proteins with myristic acid occurs cotranslationally. *Science (New York, N.Y.)*, vol. 238(4831), páginas 1275–8, 1987. ISSN 0036-8075.

WILSON, C. A., KREYCHMAN, J. y GERSTEIN, M. Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *Journal of molecular biology*, vol. 297(1), páginas 233–49, 2000. ISSN 0022-2836.

WILSON, D., MADERA, M., VOGEL, C., CHOTHIA, C. y GOUGH, J. The SUPERFAMILY database in 2007: families and functions. *Nucleic acids research*, vol. 35(Database issue), páginas D308–13, 2007. ISSN 1362-4962.

ZHANG, G., HUBALEWSKA, M. y IGNATOVA, Z. Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. *Nature structural & molecular biology*, vol. 16(3), páginas 274–80, 2009. ISSN 1545-9985.

–¿Qué te parece desto, Sancho? – Dijo Don Quijote –
Bien podrán los encantadores quitarme la ventura,
pero el esfuerzo y el ánimo, será imposible.

Segunda parte del Ingenioso Caballero
Don Quijote de la Mancha
Miguel de Cervantes

–Buena está – dijo Sancho –; fírmela vuestra merced.
–No es menester firmarla – dijo Don Quijote–,
sino solamente poner mi rúbrica.

Primera parte del Ingenioso Caballero
Don Quijote de la Mancha
Miguel de Cervantes