Universitat Autònoma de Barcelona

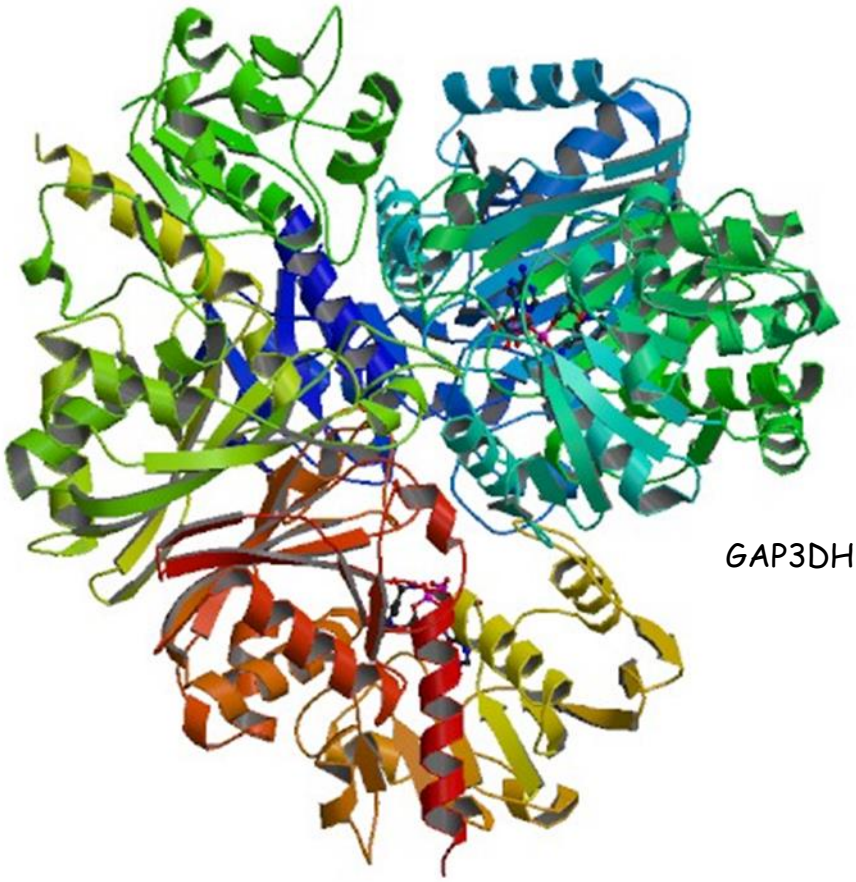
ANÁLISIS BIOINFORMÁTICO DE LAS PROTEÍNAS MULTIFUNCIONALES (MOONLIGHTING)

Sergio Iván Hernández Ranzani

ANÁLISIS BIOINFORMÁTICO DE LAS PROTEÍNAS MULTIFUNCIONALES (MOONLIGHTING)

Sergio Iván Hernández Ranzani

2016



GAP3DH

Universitat Autònoma de Barcelona

ANÁLISIS BIOINFORMÁTICO DE LAS PROTEÍNAS MULTIFUNCIONALES (MOONLIGHTING)

Tesis doctoral presentada por Sergio Iván Hernández Ranzani, Licenciado en Biología y Bioquímica, para optar al grado de Doctor en Biotecnología por la Universitat Autònoma de Barcelona

Este trabajo ha estado realizado en el Institut de Biotecnologia i Biomedicina de la Universitat Autònoma de Barcelona bajo la dirección de los doctores Enrique Querol Murillo y Juan Cedano Rodríguez (CENUR Litoral Norte-Salto, Universidad de la República, Uruguay)

Sergio Iván Hernández Ranzani

Dr. Enrique Querol **Dr. Juan Cedano**

“Annotation is not a routine activity. On the contrary, this is exciting research, somewhat akin to detective work, which has the potential of teasing out deep mysteries of life from genome sequences”.
Koonin E.V. y Galperin M.Y.

“One of the most startling results of the genomic revolution is what we don’t know, rather than what we do”
M. Adams

INDICE

RESUMEN

Moonlighting es la capacidad de algunas proteínas para ejecutar dos o más funciones bioquímicas. En general, las proteínas multifuncionales se identifican experimentalmente por casualidad (serendipia). Por esta razón, sería útil que la Bioinformática pudiera predecir esta multifuncionalidad, sobre todo debido a la gran cantidad de secuencias de proteínas provinientes de los proyectos genoma. En el presente trabajo, analizamos y describimos varios enfoques que utilizan secuencias, estructuras, interactómica y algoritmos y programas bioinformáticos corrientes para tratar de superar este problema. Entre estos enfoques están: a) la búsqueda de homología remota utilizando Psi-Blast, b) la detección de motivos y dominios funcionales, c) utilizar la información contenida en las bases de datos de interactómica (PPIs), d) el análisis de correlación de mutaciones de aminoácidos mediante algoritmos como MISTIC. Los programas diseñados para identificar un motivo o dominio funcional detectan principalmente la función canónica pero generalmente fallan en la detección de la función moonlighting, en todo caso Pfam y ProDom son los mejores métodos. La búsqueda de homología remota por Psi-Blast combinado con los datos de las bases de datos interactómica (PPIs) presentan el mejor rendimiento. La información estructural y análisis de correlación de mutación nos pueden ayudar a mapar los sitios funcionales. El análisis de correlación de mutación sólo se puede utilizar en situaciones muy específicas dado que se requiere la existencia de un multialineamiento de numerosas secuencias de proteínas de una familia funcional, pero esta estrategia puede sugerir cómo tuvo el proceso evolutivo de adquisición de la segunda función. En los análisis del presente trabajo se ha utilizado la base de datos de proteínas multifuncionales MultitaskProtDB (http://wallace.uab.es/multitask/), publicada anteriormente por nuestro grupo. Finalmente, indicar que un gran porcentaje (76%) de las proteínas multifuncionales humanas están implicadas en enfermedades y que el 47% son dianas de fármacos existentes. Esto aumenta el interés por los métodos para la predicción de proteínas moonlighting.

RESUM

Moonlighting és la capacitat d'algunes proteïnes per executar dos o més funcions bioquímiques. En general, les proteïnes multifuncionals s'identifiquen experimentalment per casualitat (serendipia). Per aquesta raó, seria útil que la Bioinformàtica pogués predir aquesta multifuncionalitat, sobretot a causa de la gran quantitat de seqüències de proteïnes provinents dels projectes genoma. En el present treball, analitzem i descrivim diversos enfocaments que utilitzen seqüències, estructures, interactòmica i algoritmes i programes bioinformàtics corrents per intentar superar aquest problema. Entre aquests enfocaments estan: a) la recerca d'homologia remota utilitzant Psi-Blast, b) la detecció de motius i dominis funcionals, c) utilitzar la informació continguda en les bases de dades de interactòmica (PPIs), d) l'anàlisi de correlació de mutacions d'aminoàcids mitjançant algoritmes com MISTIC. Els programes dissenyats per identificar un motiu o domini funcional detecten principalment la funció canònica però generalment fallen en la detecció de la funció moonlighting, en tot cas Pfam i ProDom són els millors mètodes. La recerca de homologia remota per Psi-Blast combinat amb les dades de les bases de dades interactòmica (PPIs) presenten el millor rendiment. La informació estructural i anàlisi de correlació de mutació ens poden ajudar a mapar els llocs funcionals. L'anàlisi de correlació de mutació només es pot utilitzar en situacions molt específiques, doncs es requereix l'existència d'un multialinemaiento de nombroses seqüències de proteïnes de una família funcional, però aquesta estratègia pot suggerir com va el procés evolutiu d'adquisició de la segona funció. En les anàlisis d'aquest treball s'ha utilitzat la base de dades de proteïnes multifuncionals MultitaskProtDB (http://wallace.uab.es/multitask/), publicada anteriorment pel nostre grup. Finalment, indicar que un gran percentatge (76%) de les proteïnes multifuncionals humanes estan implicades en malalties i que el 47% són dianes de fàrmacs existents. Això augmenta l'interès pels mètodes per a la predicció de proteïnes moonlighting.

SUMMARY

Multitasking or moonlighting is the capability of some proteins to execute two or more biochemical functions. Usually, moonlighting proteins are experimentally revealed by serendipity. For this reason, it would be helpful that Bioinformatics could predict this multifunctionality, especially because of the large amounts of sequences from genome projects. In the present work, we analyse and describe several approaches that use sequences, structures, interactomics and current bioinformatics algorithms and programs to try to overcome this problem. Among these approaches are: a) remote homology searches using Psi-Blast, b) detection of functional motifs and domains, c) analysis of data from protein-protein interaction databases (PPIs), d) mutation correlation analysis between amino acids by algorithms as MISTIC. Programs designed to identify functional motif/domains detect mainly the canonical function but usually fail in the detection of the moonlighting one, Pfam and ProDom being the best methods. Remote homology search by Psi-Blast combined with data from interactomics databases (PPIs) have the best performance. Structural information and mutation correlation analysis can help us to map the functional sites. Mutation correlation analysis can only be used in very specific situations –it requires the existence of multialigned family protein sequences - but can suggest how the evolutionary process of second function acquisition took place. The multitasking protein database MultitaskProtDB (http://wallace.uab.es/multitask/), previously published by our group, has been used as a benchmark for the all of the analyses. Finally, a large percentage (76%) of the human moonlighting proteins are involved in human diseases and 47% are targets of current drugs. This augments the interest in methods for predicting moonlighting proteins.

ABREVIATURAS

Blast = Basic Local Alignment Search Tool

Blocks = Base de datos de regiones de secuencia de aminoácidos relacionables con función

DFI = Differential Fluorescence Induction

ESG = Extended Similarity Group

GAPDH = Glyceraldehyde-3-phosphate dehydrogenase

GO = Gene Ontology

GOA = Gene Ontology Anotation

GOC = Gene Ontology Consortium

HMMER de HMM = Hidden Markov Models

HP = Hypothetical Protein

IDP = Intrinsically Disordered Protein

IDR = Intrinsically Disordered Region

InterPro = Servidor de diversos programas de predicción de regiones de secuencia de aminoácidos relacionables con función

IVET = In Vivo Expression Technology,

MISTIC = Mutual Information Server to Infer Coevolution

NMR = Nuclear Magnetic Ressonance

OMIM = Base de datos Human Mendelian Inheritance in Man

PDB = Protein Data Bank

Pfam = Protein Families

PFP = Protein Function Prediction

Pfyre = Protein Homology/Analogy Recognition Engine

PISITE = Protein Interaction Sites

ProDom = Protein Domains

Prosite = Servidor de Protein domain database for functional characterization and annotation

ProtLoc = Protein Localisation

PsiBlast = Position Specific Iterated BLAST

Psort = Prediction of Protein Sorting Signals and Localization Sites in Amino Acid Sequences

PSSM = Position Specific Scoring Matrices

SCOTS = Selective Capture of Transcribed Sequences

STM = Signature-Tagged Mutagenesis

Transmem = Predictor de regiones transmembranas de proteínas

I. INTRODUCCIÓN

I.A. PROTEÍNAS MOONLIGHTING O MULTIFUNCIONALES

Las proteínas moonlighting o multifuncionales son, como su nombre indica, proteínas que presentan más de una función y normalmente la segunda sin relación alguna con la primera. La primera función descrita para una proteína recibe el nombre de función canónica y la segunda(s) función(es) es la moonlighting. El orden funcional es un orden histórico, de la fecha de su identificación, y no implica especial relevancia funcional. La adquisición de la segunda función no implica en ningún caso la pérdida de la función canónica. En muchos casos se observa que la función canónica corresponde a una enzima o proteína del metabolismo primario y la función moonlighting corresponde a una función más compleja, de adquisición posterior (por ejemplo ser proteína del cristalino). La multifuncionalidad no tiene necesariamente que estar asociada a dos dominios independientes de la proteína. Por supuesto que subyacente hay una cuestión fundamental en la que no entraremos ahora: ¿cómo definimos y detectamos una función biológica? Porque los éxitos de la ingeniería genética han llevado a hacer pensar que la función es una característica absoluta y trasladable a cualquier otra célula, organismo o sistema, lo cual no es cierto y la función de una proteína depende del contexto proteico en que se encuentre (Kriston et al., 2010).

Las proteínas moonlighting son de reciente descubrimiento. Históricamente las primeras proteínas moonlight son la *aldehyde dehydrogenase* identificada como proteína estructural en el cristalino del ojo, por Wistow & Piatigorsky en 1987, y la neuroleukin, una citoquina identificada por Chaput et al., en 1988 y que resultó ser la *phosphoglucose isomerase*. Siempre se descubre como en esos casos por “serendipity” (hallazgo realizado por casualidad). Al identificar el gen/proteína relacionada con alguna función resulta ser una proteína ya conocida que presenta otra función, generalmente más ancestral y más básica, por ejemplo es muy corriente que corresponda a una función del metabolismo primario. Cuando por primera vez aparecen las proteínas moonlight en un tratado de Bioquímica general, lo son en la última versión del Lehninger (capítulo 16, página 624), en que han sido incorrectamente traducidas como “enzimas del claro de luna”…

El genoma humano aparentemente contiene tan sólo unos 20.000 genes y mecanismos

como el splicing alternativo permite dar lugar a muchas más proteínas, pero el moonlighting añade más capacidad funcional. El moonlighting está extendido por todas las ramas de la Vida, de forma que en procariotas no hay splicing alternativo pero si hay moonlighting. Diversas revisiones sobre el tema de las proteínas moonlighting son: Wool, 1996; Jeffery, 1999, 2003, 2004 and 2009; Piatigorsky 2007; Gancedo and Flores, 2008; Nobeli et al., 2009; Huberts and van der Kiel, 2010; Copley, 2012; Jeffery, 2013, 2014; Henderson and Martin, 2014.

La multifuncionalidad suele estar asociada a las siguientes características (Figura 1):

- **Localización celular**: Por ejemplo, la proteína PutA de *E. coli* es la *pyrroline-5-carboxylate proline dehydrogenase* si está asociada con la membrana plasmática y factor de transcripción cuando está en el citoplasma de la bacteria. En algunos casos no se ha identificado la función moonlighting pero su presencia en una localización anómala sugiere que probablemente la presentará. Por ejemplo la enolasa de *Plasmodium falciparum* se ecuentra también en núcleo o asociada a citoesqueleto pero se desconoce su probable función adicional en esa localización.
- **Intracelular/secretada**: La *phosphoglucose isomerase*, es un enzima de la glucólisis en el citoplasma y también neuroleukin, un factor de crecimiento nervioso, cuando es secretada. La enolasa de microorganismos patógenos, es un enzima de la glucólisis en el citoplasma y un factor de virulencia por unirse al plasminógeno del huésped cuando es secretada.
- **Expresión diferencial**: La neurophilin inducida por el Endothelial Growth Factor estimula la producción de células sanguíneas en las células endoteliales mientras que, inducida por la *semaphorin III*, da lugar al crecimiento correcto del axón en neuronas.
- **Oligomerización**: la *glyceraldehyde-3-phosphate dehydrogenase* (GAPDH), como tetrámero y en citoplasma, es un enzima de la glucólisis y, como monómero y en el núcleo, es la *uracil-DNA glycosylase*. La *pyruvate kinase* es quinasa como tetrámero y factor de unión a hormona tiroidea como monómero.
- **Utilizar distintos sitios de unión**: La proteína ribosomal S10 de *E. coli* además de unirse al rRNA16s interviene en regulación de transcripción vía unión al terminador de transcripción NusB.
- **Modificación postraduccional**: La *phosphoglucose isomerase* fosforilada en la Ser185 no actúa como enzima sino como Autocrine Motility Factor.
- **Presentar "partners" de interacción inesperados:** Por ejemplo la proteína Arg5 (el

enzima *N-acetyl glutamate kinase*) de levadura se une a varias regiones del DNA y ha resultado ser un factor de trascripción.

- En funcion de la **concentración de algún metabolito**: la aconitase a elevada concentración de hierro en vez de enzima del ciclo de Krebs es una proteína Iron-Responsive Element-Binding.

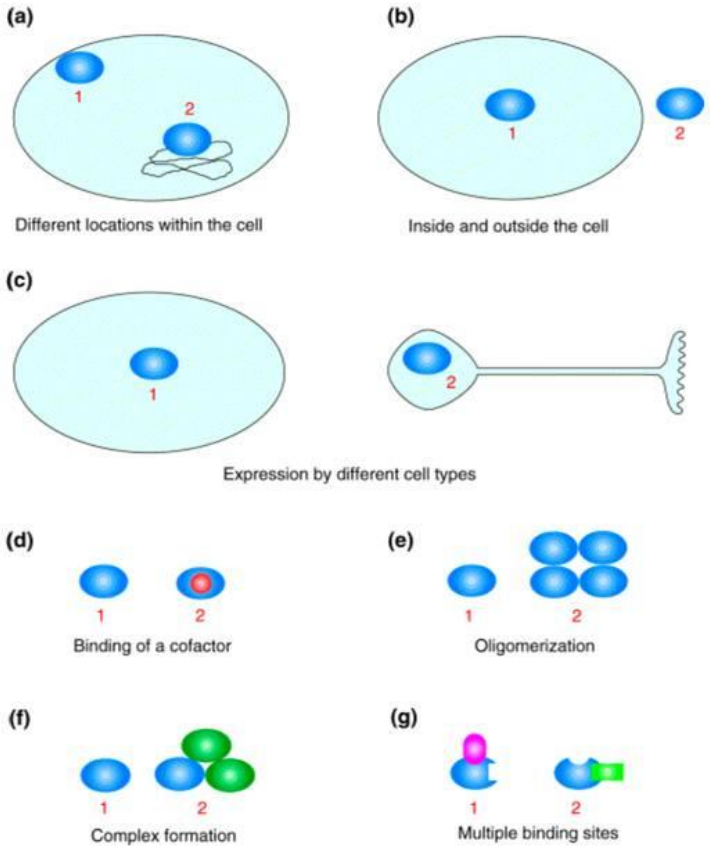


Figura 1. La multifuncionalidad está asociada a diferentes características, procesos, localizaciones celulares, modificaciones post-traduccionales, etc, de las proteínas

En los casos descritos suele estar experimentalmente demostrada la segunda función en una o en unas pocas especies. En principio no se puede universalizar la función moonlighting a más de una especie si no se ha demostrado experimentalmente. Más adelante, en Resultados y Discusión se volverá sobre este punto.

Existen varios términos para describir las proteínas multifuncionales. Uno de sus descubridores, G. Piatigorsky, acuñó el de "*Gene sharing*" pero no se ha impuesto porque el

splicing alternativo también implicaría compartir un gen pero dando lugar a diferentes polipéptidos mientras que en el moonlighting es un polipéptido el que presenta más de una función. Además existen casos en que existen dos genes para la misma proteína pero regulados por diferentes promotores por lo que son expresados de forma diferente dando lugar a un caso de moonlighting sin compartir gen. El término *Moonlighting* se debe a Constance Jeffery (Jeffery, 1999) pero ella lo utiliza de forma más restrictiva que el de *multitasking*, multifuncionalidad. Según Jeffery el término moonlighting estaría restringido a aquellos casos de las proteínas multifuncionales en que la segunda función no proviene de la fusión de genes, ni en la de dos dominios claramente diferenciados, ni presentar dos diferentes actividades catalíticas en el mismo centro activo, aunque sí acepta el presentar dos actividades enzimáticas en diferentes centros activos. De hecho muy pocos investigadores en el tema aceptan la definición tan restrictiva de C. Jeffery, y que además y desde un punto de vista evolutivo puede ser muy difícil de determinar dado que en la mayoría de casos no se puede conocer si una doble funcionalidad proviene evolutivamente de una más o menos lejana fusión de genes de proteínas (o de la región correspondiente a dominios de proteínas) o por mutaciones en un polipéptido. Lo fascinante e importante de estas proteínas es la capacidad de realizar funciones diferentes a partir del mismo polipéptido, no el que provengan de fusiones de genes o dominios.

Otro término también utilizado es el de *proteínas promiscuas* (Noveli et al., 2009). Pero este término que se utiliza bastante más en el caso de enzimas con más de una actividad enzimática, en el caso de la multifuncionalidad estricta crea algunos problemas. Por ejemplo una quinasa sería una proteína promíscua pero no sería una proteína moonlighting ni multifuncional puesto que realiza la misma función, fosforilar una proteína diana, en diferentes dianas que a su vez estarán involucradas en diferentes rutas metabólicas. El ideal de proteína moonlight sería aquella que pertenece a dos clases funcionales muy diferentes, por ejemplo ser enzima y factor de transcripción (que como se verá más adelante, se trata del par funcional más abundante tras el análisis de nuestra base de datos de proteínas moonlighting).

En todo caso **NO serían verdaderos casos de moonlighting** los siguientes:
- Variantes de proteínas por el mecanismo del splicing alternativo o producto de una proteólisis que de lugar a fragmentos con diferentes funciones biológicas.

- Aquellas enzimas que presentan una amplia gama de substratos (p.e., superfamilia aldehído deshidrogenasas, los citocromos P_{450}). Incluso enzimas muy específicos suelen presentar otras funciones secundarias, "adventicias", pero órdenes de magnitud más ineficientes.
- Algunos factores de transcripción que pueden unirse a diferentes promotores.
- Una enzima que interviene como tal en diferentes rutas (p.e., *ribulose-phosphate 3-epimerase* actúa en la ruta de las pentosas fosfato y en el metabolismo del formaldehído).
- Como se ha mencionado anteriormente, para algunos autores (p.e., C. Jeffery) una proteína producto de la fusión de 2 genes (p.e., la HisB de *E.coli* producto de la fusión del *imidazoleglycerol-phosphate dehydratase* y la *histidinolphosphatase*) no sería un verdadero caso de moonlighting dado que ambas funciones, aunque por separado, ya preexistían. Sin embargo para otros muchos autores (p.e., J. Thornton) sí lo sería dado que se trata de una multifuncionalidad. También los autores más restrictivos (p.e., C. Jeffery) consideran que no debería considerarse un verdadero caso de moonlighting una proteína con dos actividades enzimáticas diferentes en el mismo centro activo, mientras otros como J. Thornton sí pues se trata de multifuncionalidad. Thornton las considera ejemplo de proteína "promíscua" (Nobeli et al., 2009). Hay que mencionar que en muchos casos estas segundas actividades enzimáticas promiscuas sólo afectan al fenotipo si se producen en gran cantidad dado que suelen ser bastante ineficientes (Copley, 2003). Jeffery ha propuesto que se utilice el término moonlighting en sentido estricto de acuerdo con la definición suya y multitasking (multitarea o multifuncional…) para el resto, incluyendo las provenientes de fusiones de genes o dominios. Pero muchos autores, como es nuestro caso, las siguen utilizando indistintamente.

SI deberían considerarse verdaderos casos de moonligting aquellos en que las modificaciones post-traduccionales están relacionadas con la multifuncionalidad. p.e., algunas de las diferentes funciones de la GAPDH lo están relacionadas con nitrosilación, fosforilación y acetilación. O la *citrate synthase* que fosforilada es enzima y defosforilada es un componente estructural de los filamentos citoplasmáticos (Figura 2). O la proteína p53 que fosforilada en los resíduos S46+T55 se une al TFβ y fosforilada en los resíduos S15+S20 bloquea la interacción con la proteína MDM2 (Figura 3). También serían verdaderos casos de moonlighting aquellas proteínas que ocasionalmente se anclan en la membrana vía unión a un ácido graso y presentan una función alternativa.

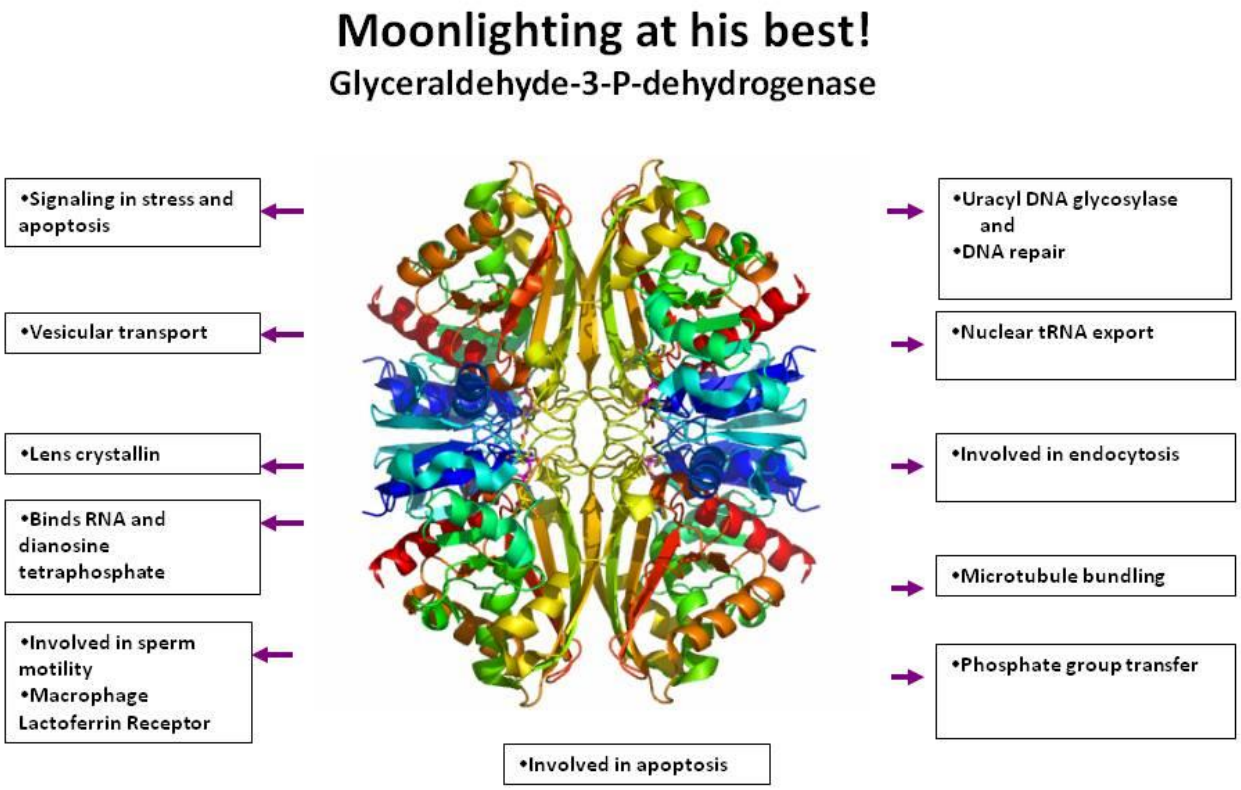


Figura 2. La Glyceraldehyde-3-Phosphate-Dehydrogenase (GAPDH) es la proteína con más funciones, 14, descritas hasta el momento (en diferentes organismos)

El fenómeno del moonlighting está muy ligado a la ancestralidad ("ancient enzymes") de las enzimas del metabolismo primario (glicólisis, ciclo de Krebs, etc) (Sriram et al., 2005). Son enzimas además muy conservados. Las Figuras 4 y 5 muestran que la mayoría de las enzimas de la glicólisis y del ciclo de Krebs están involucradas en dos o más funciones. Hasta el momento la enzima *glyceraldehyde-3-P-dehydrogenase* (GAPDH) es la proteína con más funciones identificadas hasta ahora, 14, (pero esas 14 funciones lo son en diferentes organismos, es obvio que en *E. coli* no puede estar en el cristalino). Son las siguientes 14 funciones:

Funciones de la *Glyceraldehyde-3-P-deshydrogenase* (GAPDH):

- Glyceraldehyde-3-P-DH (función canónica o principal)
- Uracyl DNA glycosylase (reparación del DNA)
- Involved in endocytosis
- Microtubule bundling
- Phosphate group transfer
- Involved in apoptosis
- Involved in sperm motility
- Binds RNA and dianosine tetraphosphate
- Lens crystallin
- Vesicular transport
- Nuclear tRNA export
- Repair of basic sites in DNA
- Signaling molecule between stress factors and cellular apoptotic machinery
- Macrophage Lactoferrin Receptor

Probablemente la proteína reguladora de transcripción **p53** la superará pues no sólo se le conocen ya diversas funciones sino que presenta centenares de interacciones con proteínas reguladoras, etc. Otras proteínas con multiples funciones son las **hsp90α**, **hsp70, HMGB1, aconitase, enolase, Cpn10, dihydrolipolamide dehydrogenase, ubiquitin** y **EFTu**. En el caso de la proteína más abundante en la sangre, la **albumina sérica**, no sólo es un transportador "universal" sino que representa el 13% de la proteína de la córnea y además su "core" hidrofóbico es un sitio activo para diversas catálisis (oxidación NO; formación S-nitrotioles que preservan el NO, *carboxylesterase*, etc). Se ha propuesto una subclasificación de las proteínas moonlighting en *Single Additional Function Moonlighting Proteins* (SAFMPs) y *Multiple Additional Function Moonlighting Proteins* (MAFMPs) según tengan una sola o multiples funciones adicionales (Henderson, 2014).

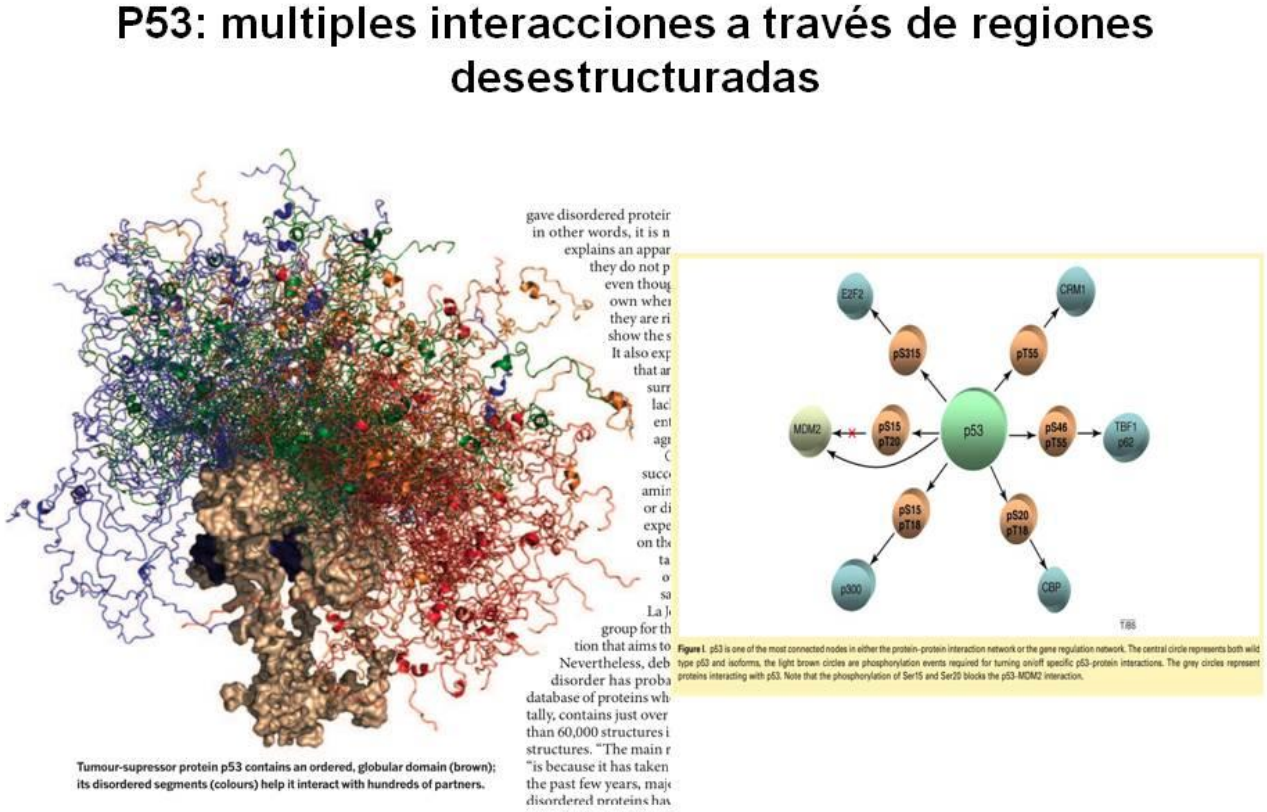


Figura 3. La proteína p53 probablemente superará a la GAPDH en número de funciones

Muchos virus presentan proteínas moonlighting, especialmente aquellos que contienen muy pocos genes/proteínas en su genoma. Se trata de proteínas preferentemente dedicadas a la evasión del sistema de defensa del huésped. Por ejemplo el virus del papiloma humano tan sólo tiene 8 proteínas. La proteína E6 interacciona con la proteína humana p53 y el complejo es degradado en el proteasoma. La proteína E6 también contribuye a la subexpresión de la proteína humana SET7 que metila a la proteína p53. Si la p53 no está metilada disminuye su estabilidad. La proteína E6 interacciona también con proteínas relacionadas con regulación de la transcripción, control de la proliferación celular, apoptosis, adhesión, estabilidad del cromosoma, reconocimiento por el sistema inmune, polarización celular y estructura epitelial (Tungteakkun & Duerksen-Hughes, 2008).

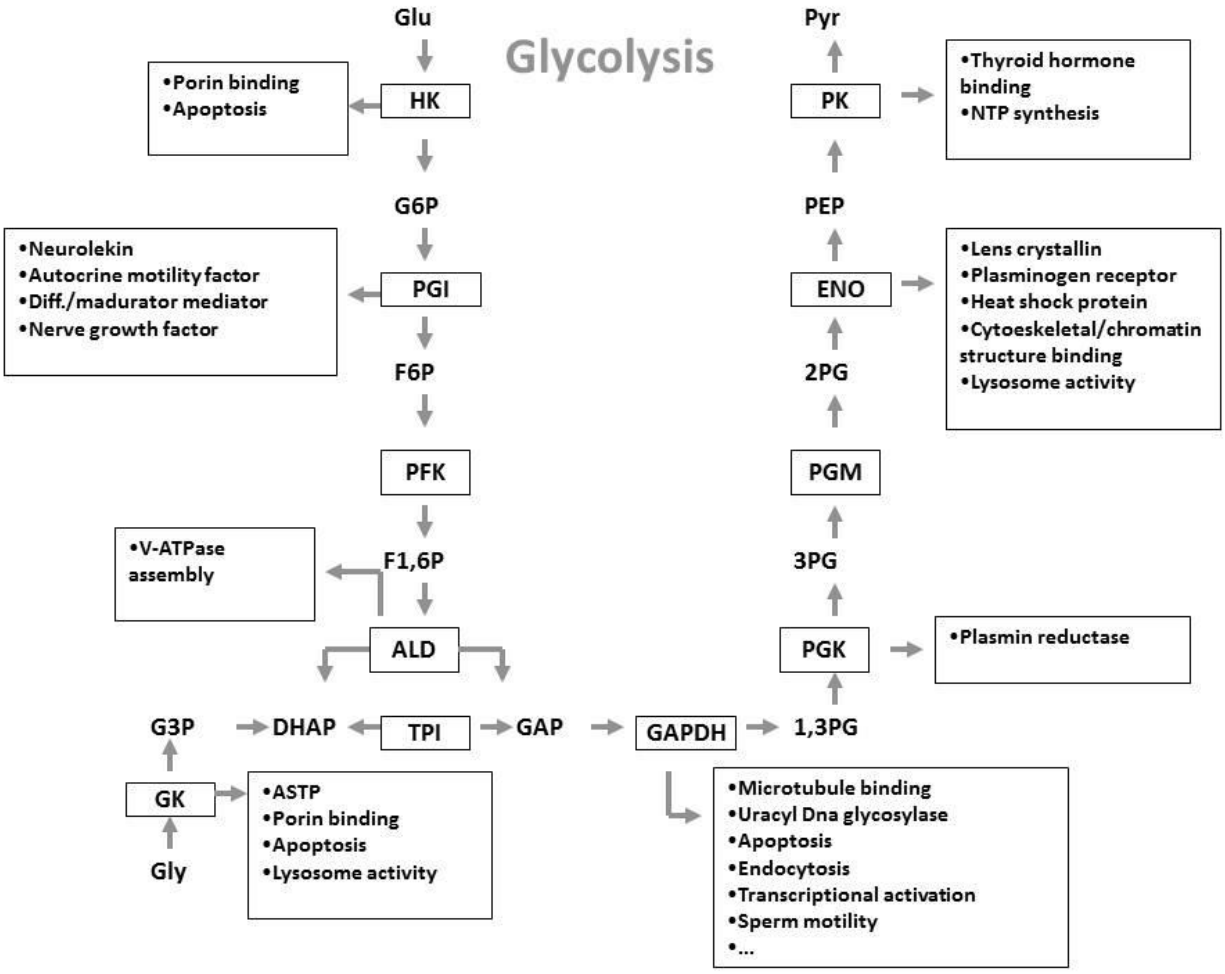


Figura 4. Muchas enzimas del metabolismo primario presentan funciones moonlighting. Por ejemplo la Glicólisis

Por otra parte la misma función moonlight puede ser producida por diferentes proteínas en diferentes especies, incluso cercanas. En el caso de las proteínas del cristalino existen diversos ejemplos (Piatigorsky, 2007) y también en enzimas citosólicos de patógenos (por ejemplo GAPDH, *enolase*, *phosphoglucomutase*, *phosphoglycerate kinase*, DnaK, *peroxiredoxin* y *elongation factor Tuf*, en que todas ellas son proteínas de unión al plasminógeno del huésped (Henderson and Martin, 2011; 2013).

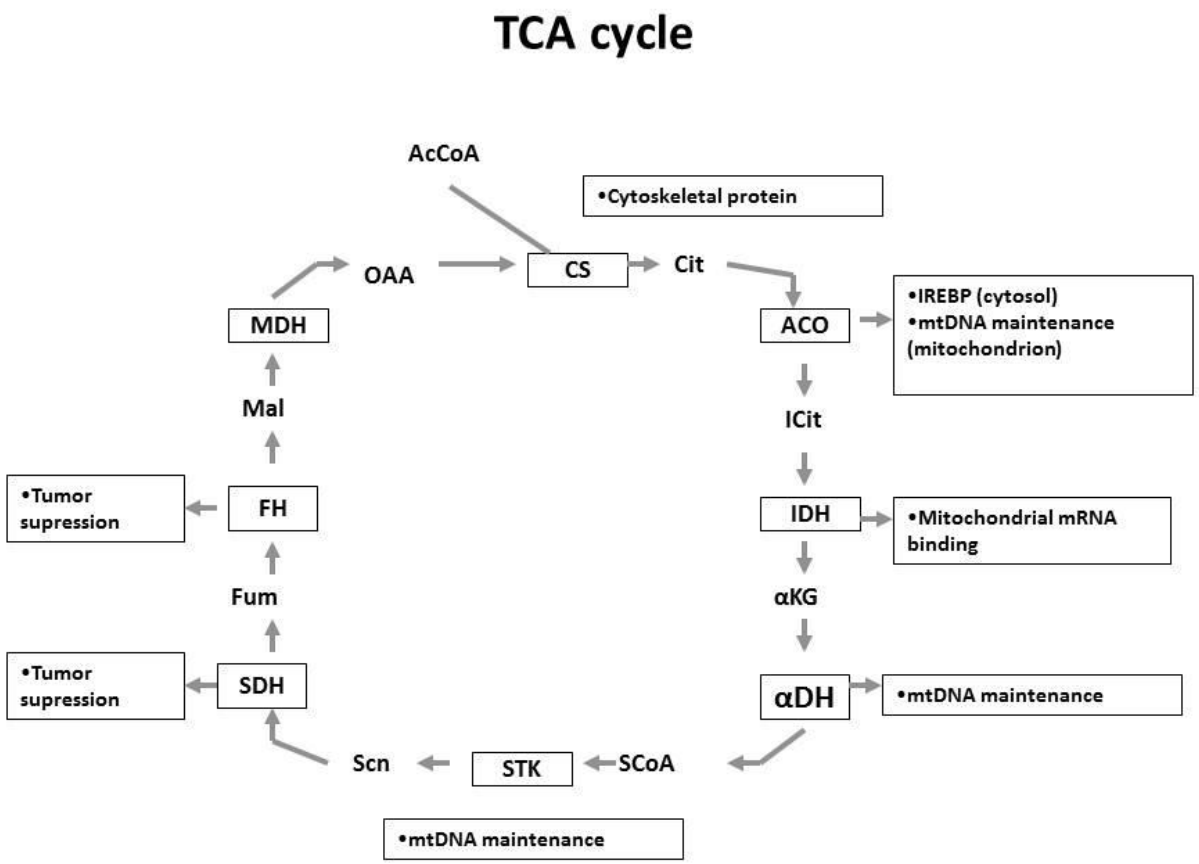


Figura 5. Muchas enzimas del metabolismo primario presentan funciones moonlighting. Por ejemplo el ciclo de Krebs

I.B. REPERCUSIÓN DE LA MULTIFUNCIONALIDAD EN DIFERENTES ÁMBITOS DE LA BIOQUÍMICA DE PROTEÍNAS

El fenómeno del moonlighting perturba la interpretación de los experimentos de:

- Anotación funcional genómica: Inexactitudes o descripción insuficiente en la misma.
- Experimentos de Knock-outs y Knock-downs para el análisis de la función génica.
- El mecanismo de "Enzyme recruitment" y el "non orthologous gene displacement".
- Análisis metabólico y predicción de rutas metabólicas, etc.
- Análisis de redes de interactómica. Por ejemplo interacciones con partners que pueden ser considerados Falsos Positivos, análisis de redes de genes y por todo ello en Biología de Sistemas. Y la existencia de la multifuncionalidad incrementará la complejidad de las redes

de interactómica.

- La acción de fármacos ("drug targeting", efectos secundarios y toxicidad, polifarmacología…).
- Existen enfermedades por adquisición de una segunda función por parte de una proteína ("neomorphic moonlighting"). También ha sido involucrado en la aparición de resistencia bacteriana a los antibióticos (p.e., la *glutamate racemase* de *M. tuberculosis* presenta como función moonlighting la resistencia al antibiótico *ciprofloxacin*). O por exceso de actividad moonlighting de una proteína, p.e. GAPDH que presenta también actividad apoptótica estaría involucrada en Alzheimer, Parkinson e isquemia cerebral.
- En microorganismos patógenos hay muchos ejemplos de proteínas o enzimas del metabolismo primario (p.e., la enolasa de la glicólisis…) que, localizados en la membrana, presentan una segunda función como factor de virulencia para facilitar su unión a las células del huésped.
- Además, el moonlighting daría lugar al denominado Conflicto Adaptativo (las mutaciones beneficiosas para una función pueden ser desfavorables o deletéreas para la otra). Los conflictos adaptativos se resuelven por duplicación génica y evolución independiente de los parálogos. Esto ha dado lugar a superfamilias de proteínas (transportadores, reguladores de transcripción…).
- Finalmente y como se comentará en el Apartado IV.G. a lo largo del presente trabajo se ha encontrado que el 76% de las enfermedades humanas con base genética corresponden a proteínas moonlighting. Y que el 47% de las dianas farmacéuticas conocidas también son proteínas moonlighting.

Todo ello hace que la predicción bioinformática de las proteínas moonlghting sea un objetivo importante.

I.C. CLASES FUNCIONALES QUE PRESENTAN LAS PROTEÍNAS MOONLIGHTING

Las proteínas moonlight presentan una amplia gama de pares de funciones. A partir de la base de datos que hemos creado (Hernández et al., 2014) y como se describirá en el capítulo IV.A. la combinación más abundante es la de "enzyme" y "nucleic acid binding protein" (incluye aquí los factores de transcripción) siendo la segunda "enzyme" y "cell adhesion". En la Tabla 2 del apartado IV.A. de Resultados se detallan los pares de funciones.

Pero pueden encontrarse prácticamente todas las grandes clases funcionales, receptores, proteínas del citoesqueleto, chaperonas, etc. Y suelen corresponder a dos funciones realizadas en diferente compartimento celular, al menos en las células eucariotas. Pueden presentar las dos actividades simultáneamente o en diferente momento del ciclo celular. En todo caso las tres grandes clases funcionales más representadas son: (a) enzimas, (b) proteínas que se unen al DNA o RNA y (c) proteínas estructurales, generalmente citosólicas. Desde un punto de vista estructural hay pocas proteínas de membrana, lo cual es lógico debido a la complicación en el plegamiento y conformación que tendría lugar por presentar una o más estructuras transmembrana y una segunda función en el citoplasma o en el núcleo. Sin embargo, como ya se ha mencionado antes existen proteínas solubles citoplasmáticas (e incluso nucleares) carentes de secuencia o péptido señal ni de hélices transmembrana, son exportadas a la membrana plasmática de la célula (o a la pared celular de microorganismos) y utilizadas como receptor de la célula del huésped. Por ejemplo la histona H1 es utilizada como receptor para la proteína *thyroglobulin* (Brix et al., 1998). Y muchos patógenos utilizan enzimas de la glicólisis para unirse a la célula huésped o a la matriz extracelular, p.e., la *enolase*, etc, que es utilizada por distintos patógenos (Henderson and Martin, 2011; 2013). Es obvio que no utilizan el sistema ER-Golgi sino algún mecanismo alternativo no convencional de secreción. En el caso de la célula eucariota las proteínas moonlighting pueden hacer uso de las interacciones selectivas con partners específicos o vía modificaciones post-traduccionales que faciliten su localización en diferentes compartimentos celulares. Es por ejemplo el caso de la GAPDH (Tristan et al., 2011).

Por otra parte y además de presentar una segunda función, las proteínas multifuncionales pueden contribuir a coordinar diferentes actividades celulares, facilitar el “switching” o las conexiones entre diferentes rutas metabólicas, proporcionar mecanismos regulatorios tipo “feedback”, etc. Pueden corresponder a hubs en las redes metabólicas y de interactómica. Ya se ha mencionado anteriormente que muchas enzimas del metabolismo primario están involucradas en actividad moonlighting y por otra parte a partir de los datos de la interactómica se conoce que las proteínas con mayor número de conexiones son las del metabolismo energético y del mecanismo de la traducción.

En general las proteínas moonlighting corresponden a genes de regulación constitutiva lo cual es lógico dado que muchas son, al menos en su función canónica, proteínas del

metabolismo primario (http://www.proteinatlas.org/humanproteome/housekeeping).

Se suele considerar, sin un soporte experimental y conceptual sólido, que la función canónica es la "verdadera" o más "importante" y la moonlighting "secundaria". Esto lo corroboraría el que muchas actividades canónicas correspondan al metabolismo primario, por ejemplo glicólisis o ciclo de Krebs y la deleción del gen sería letal por lo que es difícil de determinar la relevancia biológica de la función moonlighting.

I.D. IDENTIFICACIÓN DE LAS PROTEÍNAS MULTIFUNCIONALES

Si ya es difícil identificar y demostrar la función biológica de una proteína mucho más lo es demostrar una doble función! Ya se ha mencionado que se suelen descubrir por serendipia. Sin embargo hay algunos indicadores que sugieren un posible caso de proteína moonlight y son:

1) Encontrar un enzima en otro **compartimento** que "no toca". P.e., *lactate dehydrogenase*, *phosphoglycerate synthase*, *aldolase* y GAPDH han sido encontradas en núcleo, o sea cabe la posibilidad de actuar como un factor de transcripción; la *lactate dehydrogenase* en el cristalino; o ser exportada y carecer de los motifs apropiados para secreción (es el caso de la enolasa en algunos microorganismos patógenos).

2) Que esté presente en mayor **cantidad** respecto a la necesaria para su papel catalítico en la ruta en que participa como "función canónica" (p.e., *lactate dehydrogenase* LDHB4 es el 5% de la proteína de oocito de ratón; la albúmina sérica es el 13% de las proteínas de la córnea…).

3) Encontrar un **fenotipo inesperado** al noquear un gen.

4) Por **interactómica**: encontrar que la proteína interacciona con partners inesperados y no tratarse de una proteína promiscua ni pegajosa (sticky, en la terminología de interactómica son aquellas proteínas que presentan muchísimas interacciones pero biológicamente irrelevantes).

5) **Bioinformaticamente**: p.e., combinando programas de predicción de dominios/motifs + de homología, especialmente remota, PsiBlast/ByPass + A partir de la información contenida en las bases de datos de Interactómica (PPIs), etc. Forma parte de los OBJETIVOS del presente trabajo.

I.E. BASE ESTRUCTURAL Y EVOLUTIVA DE LA MULTIFUNCIONALIDAD

Incluso en las proteínas multifuncionales con estructura 3D conocida (101 proteínas a partir de nuestra base de datos) hay muy pocos trabajos que describan, de forma precisa, la localización (motifs funcionales, dominios) de la función moonlighting en la estructura de la proteína. De hecho tras crear la base de datos contactamos por mail a los autores de los trabajos para conocer si habían mapado las dos funciones y en general desconocían los sitios funcionales, especialmente para la función moonlighting. En el apartado IV.E. en que se comenta la posible conservación evolutiva se describirá algún caso concreto. Ya se ha mencionado antes que la multifuncionalidad no tiene necesariamente que estar asociada a dos dominios independientes de la proteína. En algunos casos en que se han mapado las diferentes funciones, la multifuncionalidad puede estar relacionada con una mínima parte de la estructura y secuencia de la proteína como por ejemplo en los casos de las proteínas *chaperone GroEL* (Yoshida et al., 2001), la Cpn60 (Henderson et al., 2013) o la GAPDH (Sirover, 2014). Aunque en el caso de la GAPDH, que tan sólo tiene 37kDa, las diferentes funciones estan relacionadas con ser monómero o tetrámero y con diferentres modificaciones post-traduccionales. De hecho, aunque no se hayan mapado las funciones en la estructura de una proteína, en diversos casos puede asumirse que la multifuncionalidad dependerá de muy pocos aminoácidos. Por ejemplo, la EFTu de *Mycoplasma pneumoniae* y *Mycoplasma genitalium* presentan un 96% de identidad de secuencia pero se diferencian en su interacción con la fibronectina, lo que implica que esa interacción depende de unos pocos aminoácidos (Balasubramanian et al., 2009).

Aunque las funciones moonlight suelen estar en diferentes regiones/dominios de una proteína también pueden solapar o estar contíguas con la región funcional de la actividad canónica. Es el caso de la proteína ribosomal S10 cuya función moonlight como proteína de regulación de transcripción practicamente solapa con la de unión al rRNA (Figura 6) (Luo et al., 2008). También se ha mencionado que las dos funciones pueden estar relacionadas con cambios conformacionales en la proteína o en alguna de sus regiones. Por ejemplo la proteína reguladora de transcripción p53 (Figura 3) que presenta varios lazos involucrados en diferentes interacciones con sirtuin, cyclin A, CBP y S100bb, y así presentar diversas funciones (Olfield et al., 2005).

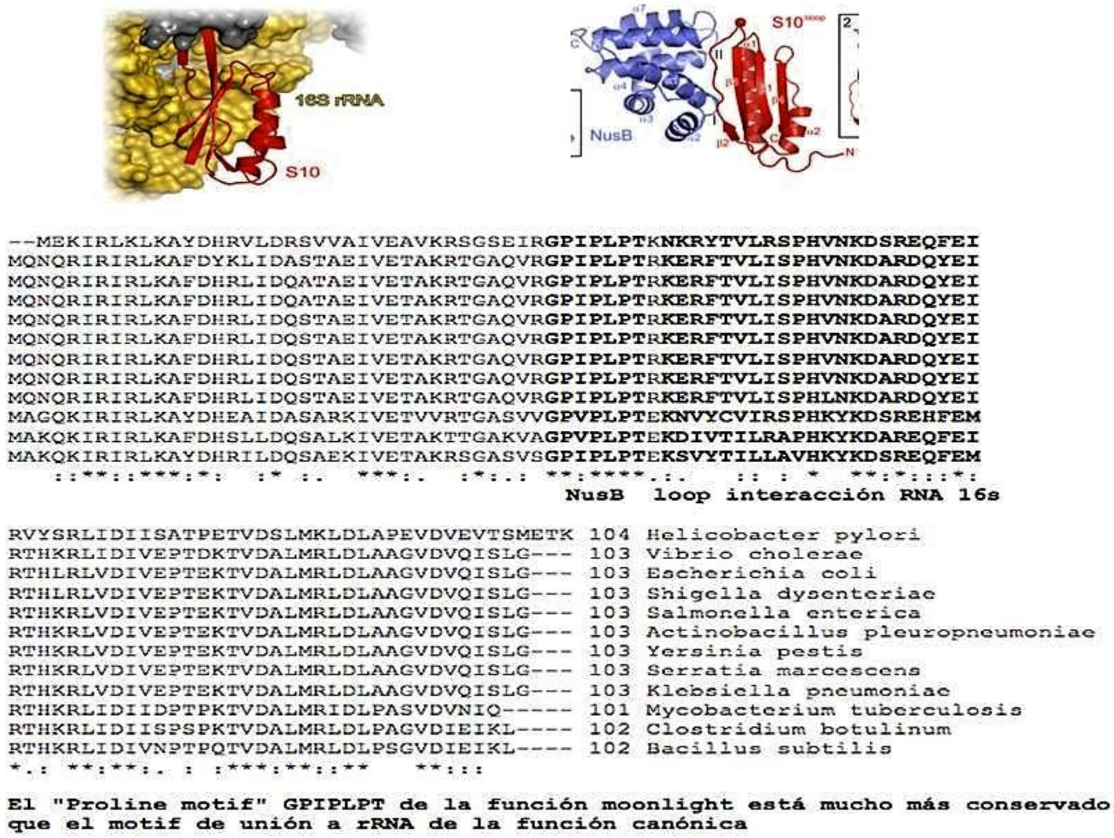


Figura 6. Estructura 3D de la proteína ribosomal S10 de *E.coli* que muestra los dos motivos responsables de sus dos funciones (unión a RNAR 16s y al factor de transcripción NusB) y su alineamiento con las secuencias ortólogas de diversas especies. Puede verse que el motif de la función moonlighting (unión al factor terminador de transcripción NusB) está más conservado que el de unión al rRNA 16s

Diversos autores (Tompa et al., 2005) han propuesto que las proteínas moonlight pertenecerían a la clase de las IDPs (Intrisically Disordered Proteins) y esto facilitaría la interacción con diversos partners, dando lugar a diferentes funciones. Sin embargo, como describimos en el apartado IV.D., en la mayoría de casos las proteínas multifuncionales no mostrarían un nivel de desestructuración total o parcial (IDP o IDR) superior al de las demás proteínas (Hernández et al., 2012). Hay casos en que las dos funciones están relacionadas con un profundo cambio estructural de la proteína. Por ejemplo la *human chemokine lymphotactin* presenta 2 formas: la Ltn10 como monomero y con un fold tipo *chemokine* canónico y como dimérica y adoptando un fold distinto, Ltn40, que se une a glucosaminoglicanos (Tuinstra et al., 2008).

Las proteínas moonlight son un ejemplo de que la Naturaleza reutiliza ampliamente secuencias y estructuras de proteínas, hace bricolage molecular en vez de diseñar continuamente “de novo”. Y también de que contrariamente a lo que opinan otros autores (Jeffery, 2013) no se requieren largos periodos de tiempo, incluso miles o millones de años de evolución, para adquirir una segunda función. En muchos casos, por ejemplo en algunas proteínas del cristalino de los mamíferos, la adquisición de la segunda función no conlleva ningún cambio secuencial sino meramente de patrón de expresión (Cvekl and Piatigorsky, 1996). Aunque la evolución, adquisición, de la multifuncionalidad es un fenómeno bastante desconocido, en muchos casos se ha visto que requiere muy pocos cambios en la proteína. Por ejemplo, la chaperona GroEL de *Enterobacter aerogenes* presenta una segunda actividad toxina para insectos. Esta GRoEL difiere en tan sólo 11 aminoácidos de la de *E. coli*, que no es toxina. Y mediante análisis mutacional se ha visto que tan sólo 4 aminoácidos de los 11 son los responsables de la función toxina (Yoshida et al., 2001). Y también, nosotros sugerimos que el Non Orthologous Gene Displacement o de reclutamiento enzimático sería un mecanismo para obtener una función moonlighting a partir de una proteína previa.

Una cuestión importante es ¿se conserva la multifuncionalidad de una proteína entre especies? Este es un objetivo muy interesante. En general se considera que identificar una proteína como moonlighting en un organismo no implica que lo haya de ser ni en especies cercanas. Por ejemplo una *pyruvate carboxylase* de *Saccharomyces cerevisiae* presenta un 80% de identidad de secuencia con la de *H. polymorpha*. La primera carece de la función moonlight (translocación de peroxisoma) de la segunda. Pero a su vez la *pyruvate carboxylase* de *P. pastoris* presenta un 80% de identidad de secuencia con *H. polymorpha*, y en éste caso sí conserva la función moonlight (Ozimek et al., 2006).

Una proteína puede ser multifuncional en dos especies presentando la misma función canónica pero sin mantener la misma segunda función en otra especie. Por ejemplo, la *aconitase* presenta como función canónica el catalizar la conversión de citrato a isocitrato en los diferentes organismos pero la función moonlighting puede ser como “Iron-Responsive Element”, “DNA maintenance”, etc, dependiendo de qué especie se trate. Las proteínas de la lente del cristalino o las Cpn60 presentan numerosos ejemplos similares. La demostración de

que presenten la misma función moonlighting requiere análisis experimental, que puede ser complicado. Sin embargo y como describiremos en el apartado IV.E. creemos que la conservación de secuencia, especialmente de “motifs” funcionales sugiere fuertemente que habrá conservación de la multifuncionalidad.

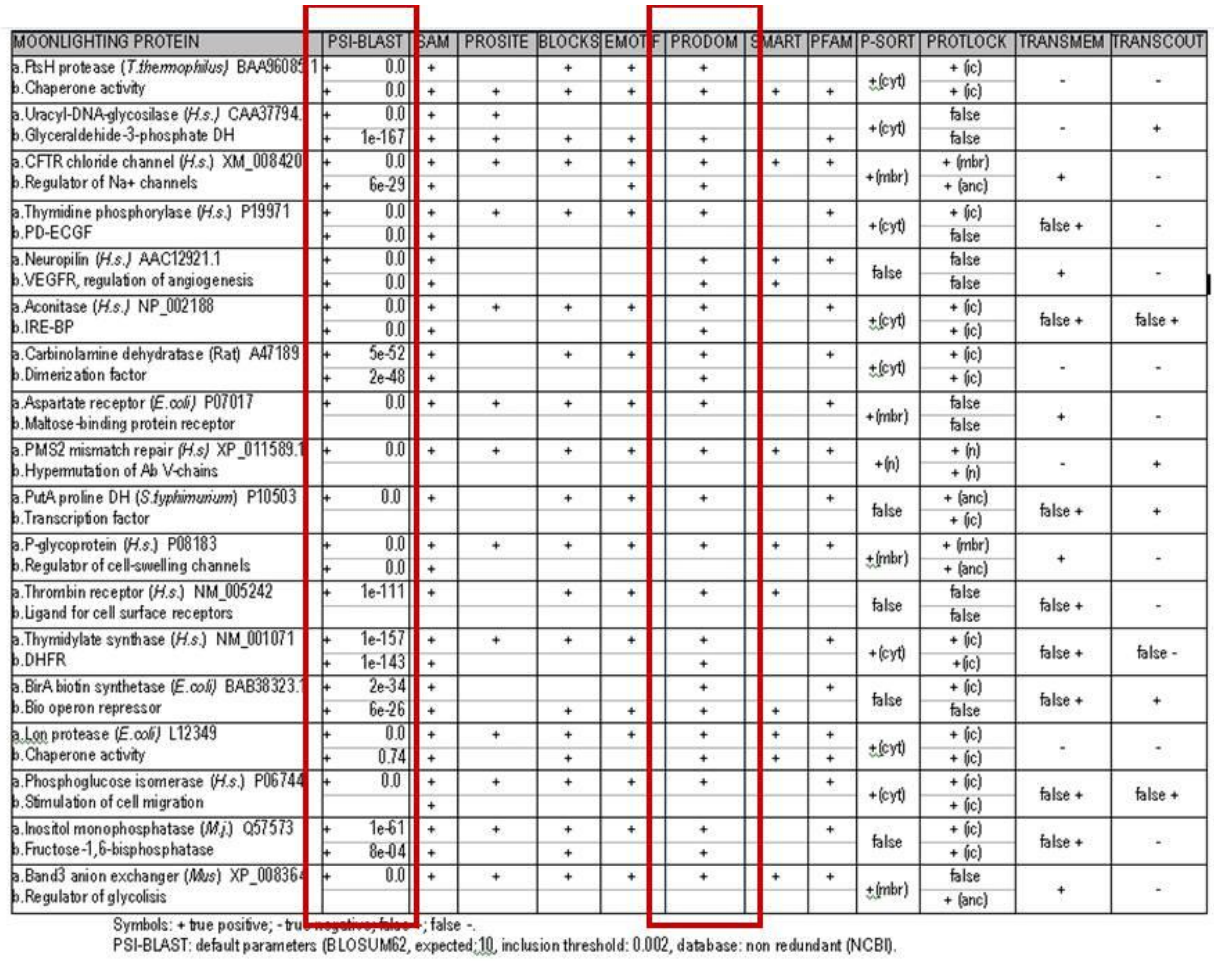
La existencia de la multifuncionalidad incrementa la complejidad del interactoma de una proteína moonlighting. Esa complejidad dependerá de cuantas proteínas moonlighting haya, pero se ha hecho muy pocos estudios de esto. En un trabajo se analizaron cuantas proteínas humanas podian ser de union a DNA. Se expresaron 4000 proteínas humanas no redundantes y el 22.4% unían DNA (Hu et al., 2009).

Como ya se ha mencionado anteriormente la multifuncionalidad afecta a la anotación funcional de las proteínas, y genes, lo cual tiene una gran importancia dada la enorme cantidad de información que entra en las bases de datos procedentes de la genómica. De hecho, incluso en los casos en que se conoce que una proteína de una especie es moonlighting, no está anotada como tal en las principales bases de datos de secuencias (ncbi, ebi…). Tan sólo en alguna base de datos, por ejemplo en UniProt (www.uniprot.org), se describe esta información para unos pocos casos. Por ello era necesaria la creación de un base de datos de las proteínas multifucionales. Y el análisis de la base de datos que hemos construido (ver Apartado IV.A.) muestra que la multifuncionalidad es un fenómeno presente en todos los reinos de la Vida.

I.F. INTENTOS PREVIOS DE PREDICCIÓN BIOINFORMÁTICA DE MULTIFUNCIONALIDAD A PARTIR DE LA SECUENCIA/ESTRUCTURA DE LAS PROTEÍNAS

La mayor parte de este enfoque bioinformático, a nivel internacional, ha venido siendo realizado por el grupo, y una importante parte del mismo corresponde a los objetivos de la presente tesis. Una primera aproximación consistió en determinar si los programas de análisis de “motifs” y dominios funcionales, concretamente Prosite, Blocks, ProDom, Pfam y E-Motif eran capaces de identificar los 2 dominios relacionados con cada función, canónica y moonlight. Asimismo otro objetivo era averiguar si programas de predicción de localización celular que dan un listado de localizaciones y probabilidades (ProtLoc y psort) contribuían a

la predicción a partir de considerar las 2 localizaciones más probables. Finalmente si programas de homología remota como PsiBlast y Sam presentaban en el listado de su output dianas correspondientes a las dos o más funciones. Esto se realizó con el pequeño número de proteínas moonlight conocidas en ese momento, unas 30 y fue publicado (Gómez et al., 2003) (Figura 7). Estos análisis se han realizado también en la presente tesis para las 300 proteínas de nuestra base de datos y serán más ampliamente comentados en el apartado de Resultados.



MOONLIGHTING PROTEIN		PSI-BLAST	SAM	PROSITE	BLOCKS	EMOTIF	PRODOM	SMART	PFAM	P-SORT	PROTLOCK	TRANSMEM	TRANSCOUT
a.FtsH protease (*T.thermophilus*) BAA96085.1	+	0.0	+		+	+	+			+(cyt)	+ (ic)	-	-
b.Chaperone activity	+	0.0	+	+	+	+	+	+	+		+ (ic)		
a.Uracyl-DNA-glycosilase (*H.s.*) CAA37794.	+	0.0	+	+						+(cyt)	false	-	+
b.Glyceraldehide-3-phosphate DH	+	1e-167	+	+	+	+	+		+		false		
a.CFTR chloride channel (*H.s.*) XM_008420	+	0.0	+	+	+	+	+	+	+	+(mbr)	+ (mbr)	+	-
b.Regulator of Na+ channels	+	6e-29	+			+	+				+ (anc)		
a.Thymidine phosphorylase (*H.s.*) P19971	+	0.0	+	+	+	+	+		+	+(cyt)	+ (ic)	false +	-
b.PD-ECGF	+	0.0	+								false		
a.Neuropilin (*H.s.*) AAC12921.1	+	0.0	+				+	+	+	false	false	+	-
b.VEGFR, regulation of angiogenesis	+	0.0	+				+	+			false		
a.Aconitase (*H.s.*) NP_002188	+	0.0	+	+	+	+	+		+	+(cyt)	+ (ic)	false +	false +
b.IRE-BP	+	0.0	+				+				+ (ic)		
a.Carbinolamine dehydratase (Rat) A47189	+	5e-52	+		+	+	+		+	+(cyt)	+ (ic)	-	-
b.Dimerization factor	+	2e-48	+				+				+ (ic)		
a.Aspartate receptor (*E.coli*) P07017	+	0.0	+	+	+	+	+		+	+(mbr)	false	+	-
b.Maltose-binding protein receptor											false		
a.PMS2 mismatch repair (*H.s*) XP_011589.1	+	0.0	+	+	+	+	+	+	+	+(n)	+ (n)	-	+
b.Hypermutation of Ab V-chains											+ (n)		
a.PutA proline DH (*S.typhimurium*) P10503	+	0.0	+		+	+	+		+	false	+ (anc)	false +	+
b.Transcription factor											+ (ic)		
a.P-glycoprotein (*H.s.*) P08183	+	0.0	+	+	+	+	+	+	+	+(mbr)	+ (mbr)	+	-
b.Regulator of cell-swelling channels	+	0.0	+								+ (anc)		
a.Thrombin receptor (*H.s.*) NM_005242	+	1e-111	+		+	+	+	+		false	false	false +	-
b.Ligand for cell surface receptors											false		
a.Thymidylate synthase (*H.s.*) NM_001071	+	1e-157	+	+	+	+	+		+	+(cyt)	+ (ic)	false +	false -
b.DHFR	+	1e-143	+				+				+(ic)		
a.BirA biotin synthetase (*E.coli*) BAB38323.1	+	2e-34	+				+		+	false	+ (ic)	false +	+
b.Bio operon repressor	+	6e-26	+		+	+	+	+			false		
a.Lon protease (*E.coli*) L12349	+	0.0	+	+	+	+	+	+	+	+(cyt)	+ (ic)	-	-
b.Chaperone activity	+	0.74	+		+		+	+	+		+ (ic)		
a.Phosphoglucose isomerase (*H.s.*) P06744	+	0.0	+	+	+	+	+		+	+(cyt)	+ (ic)	false +	false +
b.Stimulation of cell migration			+								+ (ic)		
a.Inositol monophosphatase (*M.j.*) Q57573	+	1e-61	+	+	+	+	+		+	false	+ (ic)	false +	-
b.Fructose-1,6-bisphosphatase	+	8e-04	+		+		+				+ (ic)		
a.Band3 anion exchanger (*Mus*) XP_008364	+	0.0	+	+	+	+	+	+	+	+(mbr)	false	+	-
b.Regulator of glycolisis											+ (anc)		

Symbols: + true positive; - true negative/false +; false -.
PSI-BLAST: default parameters (BLOSUM62, expected:10, inclusion threshold: 0.002, database: non redundant (NCBI).

Figura 7. Tabla mostrando los primeros análisis de homología remota (PsiBlast y SAM) y de buscadores de motivos y dominios funcionales realizados en el grupo a partir del escaso número de proteínas moonlighting conocidas en aquel momento

Un grupo de investigadores (el grupo de Kihara) ha utilizado también el algoritmo PSIBLAST para tratar de predecir bioinformaticamente la multifuncionalidad. Sugieren que es mejor utilizar como matriz de alineamiento la Blosum 45 (Khan et al., 2012; 2014a and b) y la anotación restringida a las existentes en el Gene Ontology (GO) www.geneontology.org. Por

su parte el grupo de investigación de Brun describen una aproximación basada en utilizar los datos de interactómica y expresión génica y la teoría de grafos, pero también restringiendo el análisis según la anotación funcional GO (Becker et al., 2012; Chapple et al., 2015a y 2015b). Sin embargo, utilizar tan sólo ejemplos en que la anotación funcional sea la GO restringe mucho las posibles anotaciones comparadas puesto que la base de datos GO tan sólo tiene incluidas en este momento 9500 anotaciones funcionales y existen muchísimas más. Lamentablemente las bases de datos de secuencias de proteínas (ncbi, ebi…) están llenas de anotaciones anárquicas, que no seguían criterios semánticos como GO. Existen anotaciones tan imprecisas como “17 kilodalton protein”. Esto dificulta la comparación automática de los resultados de programas de homología remota y de interactómica.

Como se describirá en el apartado IV.B.2. otra aproximación que hemos realizado es utilizar la información de las bases de datos de interactómica de proteínas (PPIs) suponiendo que a partir de los partners de interactómica que presenta una proteína pueden estar los relacionados con la función moonlighting, y que ahora se desechan como falsos positivos, Figura 8 (Gómez et al., 2011). Cotejando los partners de interactómica con los homólogos remotos hemos visto que aumenta la predicción bioinformática de la multifuncionalidad de una proteína.

Protein	Known moonlighting functions	Database interacting partners	GO related functions	GO enrichment *P*-value
Aconitase	mtDNA maintenance	ATP-dependent DNA helicase MER3	GO:0017111: nucleoside-triphosphatase activity	0.00461
			GO:0030554: adenyl nucleotide binding	0.00648
			GO:0001883: purine nucleoside binding	0.00664
			GO:0001882: nucleoside binding	0.00685
			GO:0008135: translation factor activity, nucleic acid binding	0.00017
Aldolase	Vacuolar H+-ATPase assembly	V-type proton ATPase subunit E 1	GO:0008553: hydrogen-exporting ATPase activity	0.00361
			GO:0042623: ATPase activity, coupled	0.00615
			GO:0051117: ATPase binding	0.00677
			GO:0046961: proton-transporting ATPase activity, rotational	0.00857
			GO:0016887: ATPase activity	0.00857
Enolase	Bind to cytoskeletal structures	Actin	GO:0034621: cellular macromolecular complex organization	7.54×10^{-5}
			GO:0032506: cytokinetic process	0.0053
			GO:0007109: cytokinesis, completion of separation	0.0021
		Microtubule-associated protein 4	GO:0007017: microtubule-based process	0.00286
			GO:0051488: activation of anaphase-promoting complex	0.00314
			GO:0000920: cytokinetic cell separation	0.00418
Glyceraldehyde-3-phosphate dehydrogenase	Microtubule bundling	Tubulin polymerization-promoting protein	GO:0051015: actin filament binding	0.0071
			GO:0001948: beta-catenin binding	0.00594
			GO:0008017: microtubule binding	0.00251
	Phosphate group transfer	Phosphoglycerate kinase 1	GO:0017111: nucleoside-triphosphatase activity	0.00222
			GO:0016462: pyrophosphatase activity	0.00316
			GO:0016772: transferase activity, transferring phosphorus-containing groups	0.00104

Figura 8. Tabla mostrando los primeros análisis de bases de datos de interactómica con objeto de determinar si muchos de los partners considerados falsos positivos eran identificables las funciones moonlighting de proteínas multifuncionales conocidas

En todo caso la predicción bioinformática de la función de una proteína es difícil y como dice Koonin "Annotation is not a routine activity. On the contrary, this is exciting research, somewhat akin to detective work, which has the potential of teasing out deep mysteries of life from genome sequences" (Koonin & Galperin, 2003). Predecir una segunda función es todavía más complicado.

I.G. CREACIÓN DE UNA BASE DE DATOS DE PROTEÍNAS MOONLIGHT

Durante la mayor parte del desarrollo del presente trabajo y para varias publicaciones previas o del mismo, se han utilizado como ejemplos de proteínas moonlighting aquellos que se describían en los reviews del tema (Wool, 1996; Jeffery, 1999, 2003, 2004 and 2009;

Piatigorsky 2007; Gancedo and Flores, 2008; Nobeli et al., 2009; Huberts and van der Kiel, 2010; Copley, 2012). En estos reviews se mencionaba que se conocían unas 80-100 proteínas moonlight, pero tan sólo se presentaban tablas con 20 o 30, y eso la que más. Por ello, ya bastante hacia el final de este trabajo decidimos recopilar a partir de la bibliografía las proteínas moonlight conocidas, algo más de 300 en ese momento (actualmente ya hemos recopilado 650), y crear una base de datos de las mismas (http://wallace.uab.es/multitask/), (Hernández et al., 2014), que corresponde al Objetivo 1 de la presente tesis. A partir de disponer de esta base de datos se han realizado mejores aproximaciones a la predicción, evolución, etc, de las proteínas moonlight. Posteriormente a la publicación de nuestra base de datos el grupo de Jeffery ha publicado la suya (Mani et al., 2015), que es más incompleta y más complicada de utilización (por ejemplo hay menos links o no los hay a las secuencias, a UniProt, etc).

I.H. RELACIÓN DEL MOONLIGHTING CON LA INFECCIÓN POR MICROORGANISMOS PATÓGENOS Y CON PATOLOGÍAS HUMANAS

Ya se ha descrito anteriormente que muchas proteínas del metabolismo primario (enolasa, GAPHD…) de diversos microorganismos patógenos son utilizadas como factores de virulencia, por ejemplo para la adhesión al huésped (Henderson and Martin, 2011; 2013). Pero además existe el mecanismo del moonlighting forzado ("enforced moonlighting") o secuestro de proteínas del huésped para forzarlas a una segunda función que facilite la actividad del patógeno. Por ejemplo *E. coli* recluta la actina del huésped para facilitar la unión al epitelio intestinal del huésped (Backert et al. 2008). Y los virus son grandes especialistas en secuestrar proteínas del huésped para nuevas funciones en contra del mismo. Es obvio que poder predecir bioinformáticamente si una proteína es multifunctional sería de gran ayuda para identificar dianas vacunales y farmacéuticas, lo cual tiene gran interés para el grupo por cuanto una importante parte de su investigación está relacionada con la obtención de vacunas.

Ya se ha descrito que las proteínas del metabolismo primario son muy propensas a presentar función moonlighting, probablemente por su ancestralidad. Pero otro hecho relacionado con la patogenicidad y virulencia de microorganismos es que diferentes proteínas del metabolismo primario tengan igual función moonlighting en relación con la virulencia. Por

ejemplo la *enolase* de cerca de 27 diferentes especies (Tabla 1) se une al plasminógeno del huésped, pero en este caso todas ellas presentan una muy alta homología de secuencia. Pero a su vez la *glyceraldehyde phosphate dehydrogenase* de 18 especies también une plasminógeno. O la *phosphoglycerate mutase* de 4 microorganismos. Y la *triosephosphate isomerase* de 3 microorganismos, etc. Hay más ejemplos en nuestra base de datos. Esto implica que estas enzimas que no presentan similitud de secuencia comparten alguna característica conformacional, o algún motif no identificado. Entre los objetivos de otra tesis del grupo está el identificar qué motifs o dominios estarían involucrados, y conservados, en esta interacción con una proteína del huésped.

Las proteínas moonlighting también pueden estar relacionadas con las patologías humanas, normalmente por alguna mutación. Por ejemplo la GAPDH está involucrada en neurodegeneración y Alzheimer, la PGI en anemia hemolítica, etc. (Sriram et al., 2005). Por otra parte hay numerosos ejemplos de proteínas en que la función moonlighting no es una función "normal" y está relacionada con diferentes patologías por una ganancia de función tóxica (*gain-of-toxic-function*). Jeffery ha acuñado el término de Función Neomórfica (*Neomorphic Moonlighting Function*) (Jeffery, 2011). Aunque no es un objetivo de la presente tesis sino de otra del grupo, cabe mencionar que partir del análisis de las proteínas de nuestra base de datos hemos encontrado que el 76% de las proteínas moonlighting humanas de nuestra base de datos están involucradas en patologías actualmente conocidas de acuerdo con la base de datos OMIM (Hamosh et al., 2005) http://www.omim.org, y la Human Gene Mutation Database, http://www.hgmd.cf.ac.uk/ac.index.php (Cooper and Krawczak,1998). Y que el 47% de las proteínas moonlighting humanas de nuestra base de datos corresponden a dianas farmacéuticas de fármacos existentes, de acuerdo con las bases de datos de dianas farmacéuticas The Therapeutic Target Database (Qin et al., 2014), http://xin.cz3.nus.edu.sg/group/ttd/ttd/asp, y la DrugBank database (Wishart et al., 2008), http://www.drugbank.ca. Todo esto complica el análisis de dianas farmacológicas, la toxicidad de los fármacos, etc. Pero a su vez en algunos casos pueden representar una ventaja farmacocinética si el fármaco (por ejemplo un anticuerpo monoclonal) no penetra en la célula y sólo afecta a la actividad moonlighting, si esta es extracelular, o si el fármaco es específico para bloquear la interacción con algún partner concreto presentaría menos efectos secundarios (Butler and Overall, 2009). Por ejemplo, THC346, un derivado del deprenyl es neuroprotector por impedir la S-nitrosilación y por ello la interacción de la GAPHD con la

Siah1 (Hara et al., 2005).

I.I. ALGUNAS PREGUNTAS RELEVANTES ACERCA DE LAS PROTEÍNAS MULTIFUNCIONALES

El fenómeno de la multifuncionalidad conduce a un cierto número de preguntas importantes tanto para la función como para la evolución de proteínas que todavía no tienen respuesta. En la presente tesis se tratará de obtener algunas respuestas o de estimar o especular sobre otras. Algunas preguntas son por ejemplo:

(a) ¿Qué ventaja evolutiva representa el que una proteína tenga más de una función, a veces ambas funciones indispensables, en vez de utilizar la duplicación de un gen y evolución independiente del parálogo, lo cual evitaría el denominado "conflicto adaptativo"?. Como Conflicto adaptativo se entiende que una mutación que pueda mejorar una de las funciones de una proteína moonlighting podría afectar negativamente a la otra función.

(b) ¿Cuál es la base estructural de la multifuncionalidad?

(c) ¿Cuál es el mecanismo que conduce a la aparición de la segunda función?

(d) ¿Hay conservación filogenética de la multifuncionalidad?

(e) ¿Cuántas proteínas moonlight hay?

(f) ¿Son "hubs" de redes de interactómica?

(g) ¿Son ejemplos de "non orthologous gene displacement"?

(h) ¿Por qué tantos casos de moonlighting lo son de proteínas de adhesión y virulencia de microorganismos patógenos y muchas de ellas enzimas de la glicólisis?

(i) ¿Qué papel juegan las proteínas moonlighting en las enfermedades humanas?

(j) Cuántas proteínas moonlighting son posibles dianas terapéuticas y hasta que punto son responsables de la toxicidad y efectos secundarios de muchos fármacos?

(k) Finalmente y es el objetivo fundamental de la presente tesis ¿es predecible bioinformaticamente la multifuncionalidad?

Respecto a la primera pregunta la respuesta sería doble: (a) que con menos genes el organismo puede llevar a cabo más funciones y (b) la Vida más que diseñar polipéptidos enteramente nuevos suele reutilizar lo previamente existente, pues ya ha sido optimizado para dar lugar a un plegamiento y conformación estables y es más fácil añadir una función a una estructura previa que en muchos casos, como ya se ha mencionado en el apartado I.E.,

tan sólo requieren unos pocos cambios en aminoácidos. La Vida precisa de numerosas funciones biólogicas. Una manera de obtenerlas es incrementando el número de genes, pero implica una mayor carga mutacional. Una segunda manera es el splicing alternativo. Es la más utilizada por los organismos eucariotas, pero conlleva una mayor complejidad del mecanismo, existencia de proteínas estimuladoras y silenciadoras del splicing, etc. Y es un mecanismo no presente en procariotas. Finalmente, sin aumentar el número de genes ni una complejidad excesiva estaría la multifuncionalidad a partir de cambios mínimos en la secuencia de la proteína. Cabe mencionar también que a partir de los resultados del primer borrador y del Atlas del proteoma humano no parecen haber tantas proteínas como se pensaba (se estimaba en por lo menos un millón) (Kim et al., 2014; Wilhelm et al., 2014; Uhlen et al.,2015). Lo cual abundaría en optimizar al máximo las proteínas para conseguir el número de funciones biológicas necesarias.

Respecto al mecanismo que conduce a la aparición de la segunda función no se sabe pero ya se ha mencionado anteriormente el caso de la chaperona GroEL en la que mutando 4 aminoácidos dan lugar a una segunda función toxina de insectos. Probablemente el primer paso para la obtención de una nueva función sería la aparición de nuevas interacciones con otras proteínas. Finalmente tendrían lugar casos de Non Orthologous Gene Displacement o de reclutamiento enzimático. Y en otros casos aparecen por la fusión de genes o dominios de proteínas dando lugar a una proteína multifuncional (Gancedo and Flores, 2008).

Respecto a la conservación filogenética de la multifuncionalidad no se conoce apenas dado que habría que realizar experimentación en cada organismo del que se sospecha que presente multifuncionalidad. A partir de multialineamientos vemos que hay una alta conservación de secuencia, en algunos casos altísima (más del 90% de identidad). Y de hecho en los casos de proteínas de la glicólisis o del ciclo de Krebs relacionadas con virulencia hay numerosos casos conocidos de presentar las 2 funciones en diferentes especies de microorganismos patógenos. Por ejemplo, la enzima de la glicólisis enolasa es proteína moonlight como unión a plasminógeno en 27 especies de microorganismos, con otras funciones en 17 organismos más (Tabla 1**)**. Y la GADPH como unión a plasminógeno en 18 especies y con otras funciones en 15 organismos más En Material suplementario, las Tablas S1 y S2 muestran todos los casos conocidos de proteínas que se repiten como multifuncionales en diferentes especies (Tabla S1) y funciones moonlighting que se repiten

en diferentes especies (Tabla S2). Todo esto indica que hay proteínas con una gran facilidad para presentar más de una función así como funciones propensas a presentarse como moonlighting en diferentes especies y en diferentes secuencias/estructuras de proteínas, siendo una cuestión abierta la base estructural de esta facilidad y propensión.

Tabla 1. Dos ejemplos de proteínas que son moonlighting en diferentes organismos

Protein name	Organism	Moonlighting function
Aconitase	Homo sapiens Paracoccidioides brasiliensis Bos taurus	Iron responsive element
Enolase	S. pneumoniae Aeromonas hydrophila Bacillus anthracis Bifidobacterium animalis Lactobacillus johnsonii Mycoplasma fermentans Neisseria meningitidis Trichomonas vaginalis Lactobacillus crispatus Streptococcus pyogenes Candida albicans Onchocerca volvulus Streptococcus oralis Streptococcus mutans Leishmania mexicana Paenibacillus larvae Bifidobacterium bifidum Bifidobacterium breve Bifidobacterium longum Borrelia burgdorferi Lactobacillus plantarum Listeria monocytogenes Mycoplasma pneumoniae Staphylococcus aureus Streptococcus anginosus Streptococcus oralis Taenia pisiformis	Plasminogen binding
	Lactobacillus plantarum Streptococcus suis Paracoccidioides brasiliensis	Fibronectin binding
	Lactobacillus johnsonii Staphylococcus aureus	Laminin binding
	Streptococcus mutans Streptococcus gordonii	Binding to salivary mucin
	Plasmodium falciparum Echinococcus granulosus Streptococcus suis Plasmodium berghei Enterococcus faecalis Toxoplasma gondii	Adhesion to host (except the previous functions)
	Homo sapiens Mus musculus	Promotes cell survival
	Trachemys scripta Petromyzon marinus	Lens crystallin

Las preguntas sobre cuántas proteínas moonlighting hay y si son hubs de redes de interactómica no tienen respuesta fácil. Nosotros creemos que como dice C. Jeffrey (2004) “Current moonlighting appear to be ony the tip of the iceberg”. Y respecto a si pueden corresponder a hubs en las redes metabólicas y de interactómica ya se ha mencionado anteriormente que muchas enzimas del metabolismo primario están involucradas en actividad moonlighting y por otra parte a partir de los datos de la interactómica se conoce que las proteínas con mayor número de conexiones son las del metabolismo energético y del mecanismo de la traducción. O sea, sí habría muchos ejemplos de proteínas moonlighting que correspondan a un hub. Es obvio que una segunda función podría hacer que una proteína conectase dos rutas metabólicas de una manera inesperada y relevante para las aproximaciones de Biología de Sistemas.

Respecto a si las proteínas moonlighting son ejemplos de “non orthologous gene displacement” es de esperar que lo sean y represente una vía evolutiva de llegar a ser multifuncional, pero tan sólo hemos encontrado 5 ejemplos de proteína moonlight y que a la vez sea un caso de desplazamiento no ortólogo (de acuerdo con los contenidos en la base de datos de enzimas análogos (Omelchenko et al., 2010; Galperin et al., 2012). Sin embargo, este pequeño número de casos probablemente responde al insuficiente número de ejemplos actualmente conocidos de ambas clases, proteínas ortólogas y moonlighting.

Respecto a por qué tantos casos de moonlighting lo son de proteínas de adhesión y virulencia de microorganismos patógenos y muchas de ellas enzimas de la glicólisis, ésta representa uno de las rutas metabólicas más ancestrales y el moonlighting representa la reutilización de las moléculas y los procesos ya existentes. Y las proteínas relacionadas con la adhesión del patógeno al huésped ha sido un campo muy estudiado por motivos de diseño de vacunas y fármacos antimicrobianos, aparte de ser comparativamente más fácil su análisis desde el punto de vista experimental.

II. OBJETIVOS

El **Objetivo General** del presente trabajo es determinar si los algoritmos y programas bionformáticos existentes pueden contribuir a predecir las proteínas multifuncionales a partir de su secuencia.

Los **Objetivos Específicos** son:

1.- Diseñar la primera base de datos de las proteínas multifuncionales conocidas.

2.- Determinar si los programas de análisis de homología remota, especialmente PsiBlast, permiten identificar posibles casos de multifuncionalidad.

3.- Determinar si las bases de datos de interacción proteína-proteína (PPIs) contienen información útil para identificar posibles casos de multifuncionalidad.

4.- Determinar si el solapamiento de los *outputs* de los programas de alineamiento de secuencia con los de las bases de datos de interactómica mejoran la identificación de posibles casos de multifuncionalidad.

5.- Determinar si los programas de identificación de motifs/dominios funcionales a partir de la secuencia de una proteína revelan posibles casos de multifuncionalidad.

6.- Determinar si la multifuncionalidad está relacionada con las regiones desestructuradas de las proteínas y si éstas corresponden a IDPs (Intrinsically Disordered Proteins).

7.- Determinar si los programas de predicción de localización subcelular a partir de la secuencia de una proteína representan una ayuda para identificar posibles casos de multifuncionalidad.

8.- Determinar si los programas de correlación mutacional en proteínas pueden ayudar a mapar las funciones canonica y moonlighting de las mismas en los casos en que existe una extensa familia de secuencias de proteínas a comparar mediante multialineamiento.

III. METODOS

III.A. BASES DE DATOS Y SERVIDORES

III.A.1 BASES DE DATOS Y SERVIDORES UTILIZADOS

Además de la base de datos bibliográficos, de secuencias, etc, del Genbank ncbi (www.ncbi.nlm.nih.gov) se han utilizado las siguientes bases de datos o en ausencia de ellas, como en el caso de las proteínas moonlighting (hasta la creación de la base de datos MultitaskProtDB del presente trabajo) las tablas descritas en diferentes artículos.

Proteínas moonlight

Gran parte del trabajo se ha realizado utilizando las tablas de proteínas moonlight descritas en las siguientes revisiones de proteínas moonlighting publicadas (Wool, 1996; Jeffery, 1999, 2003, 2004, 2009 and 2014; Piatigorsky 2007; Gancedo and Flores, 2008; Nobeli et al., 2009; Huberts and van der Kiel, 2010; Copley, 2012). En estos reviews se mencionaba que se conocían unas 80-100 proteínas moonlight, pero tan sólo se presentaban tablas con 20 o 30, y eso la que más. Entre todos estos trabajos habían sido descritas unas 60 proteínas moonlight y en algunos casos sin especificar especie o referencias blibliográficas. Por ello, y en una fase avanzada de la tesis, decidimos diseñar nuestra base de datos (ver apartado III.A.2), **http://wallace.uab.es/multitask/** la primera existente y a partir de la cual se han realizado muchos de los análisis que se describen en el apartado IV.

Bases de datos de Interactómica (PPIs)

Las proteínas asociadas por interacción (en adelante las denominaremos por su término inglés “partner”) a las proteínas multifuncionales de la base de datos MultitaskProtDB se identificaron en el servidor APID (Prieto et al., 2006) (**http://bioinfow.dep.usal.es/apid/index.htm**). APID contiene la mayor parte de los datos de las bases de datos de proteómica MINT, DIP, BIOGRID, IntAct, HPRD y BIND. Además presenta las proteínas de acuerdo con la anotación GO (Gene Ontology) (**www.geneontology.org**) (Ashburner et al., 2000). Hemos considerado que los datos de interactómica revelan la segunda función de una proteína moonlight si la base de datos PPI identifica como partner una función molecular, o en algunos casos un proceso biológico (de

acuerdo con la anotación GO), si esta función coincide con la función moonlighting esperada. Para filtrar los aciertos y mejorar la precisión, es aconsejable realizar un análisis de enriquecimiento de ontología de genes utilizando el paquete GOStat R (Beissbarth y Speed, 2004) como se ha descrito anteriormente (Gómez et al., 2011). Cabe indicar que en el caso de los datos de interactómica es conveniente utilizar bases de datos "no curadas" (p.e., DIP, MINT y el servidor APID) porque en las curadas se ha eliminado muchas interacciones consideradas falsos positivos y que en realidad pueden ocultar una segunda función identificable por los partners. Las direcciones web de las principales bases de datos y servidores de interactómica son las siguientes (Figura 9):

APID: http://bioinfow.dep.usal.es/apid/index.htm

MINT: http://mint.bio.uniroma2.it/mint

DIP**: http://dip.doe-mbi.ucla.edu/Main.cgi**

BOND: **http://bond.unleashedinformatics.com/Action**

HPRD: **http://www.hprd.org**

BioGrid**: http://thebiogrid.org**

IntAct: **http://www.ebi.ac.uk/intact/main.xhtml**

BIND: **http://www.bind.ca**

KEGG: **http://www.genome.jp/kegg/**

STRING: **http://string-db.org**

Expasy/Swiss-Prot (**www.expasy.org/swissprot**)

Se trata de una base de datos y servidor para el análisis de proteínas establecida en 1986 y mantenida en colaboración entre el Instituto Suizo de Bioformatica (SIB) y el Instituto Europeo de Bioinformática (EBI). Proporciona un alto nivel de anotación, un nivel mínimo de redundancia de secuencias, un alto nivel de integración con otras bases de datos biomoleculares, y una extensa documentación externa. También numerosos programas para el análisis de estructura, función y evolución de proteínas.

UniProt (www.uniprot.org)

Se trata de una excelente base de datos de información de estructura, función y bibliografía sobre proteínas en **www.uniprot.org** (The Uniprot Consortium, 2103). En algunos casos, muy pocos, UNiProt describe la segunda función de la proteína.

InterPro (www.ebi.ac.uk/interpro/)

InterPro es un servidor para la identificación de familias de proteínas, dominios, motivos ("motifs") y sitios funcionales (Mitchell et al., 2015). El análisis comparativo de las familias de secuencias de proteínas muestra que algunas regiones han sido mejor conservadas que otras durante la evolución. Estas regiones son generalmente importantes para la función de una proteína y/o para el mantenimiento de su estructura tridimensional. Permiten establecer una "firma" para una familia de proteínas, o de dominios, que distingue a sus miembros de todas las otras proteínas no relacionadas. Una firma de proteína se puede utilizar para asignar una nueva proteína secuenciada a una familia específica de proteínas y por lo tanto para formular hipótesis acerca de su función. InterPro integra diversos subprogramas: Prosite, Pfam, Prints, ProDom, Smart, TIGRfams, HAMap, PIRsf, Superfamily, CathGene 3D y Panther. La última versión de InterPro (Julio 2015) contiene 18.816 familias; 7.571 dominios, 851 sitios funcionales, 281 repeticiones y 16 modificaciones postraduccionales.

Blocks (http://blocks.fhcrc.org)

Blocks es una base de datos de alineamientos de "motifs" secuenciales relacionados con dominios y sitios funcionales de proteínas (Henikoff et al., 1999). En principio es más aconsejable utilizar InterPro dado que Blocks no ha sido actualizada desde el año 2006. Pero precisamente por no haber sido actualizado, "curado", para identificar preferentemente y presentar unicamente el motif de mejor puntuación permite que pueda indentificar otros motifs relacionados con las funciones moonlighting.

Pfam (http://pfam.xfam.org/)

Pfam (Finn et al., 2011), es una base de datos y servidor (InterPro también lo incluye en su servidor) que permite identificar familias y dominios de proteínas de forma más o menos restrictiva (PfamA es más "curado" y restrictivo y PfamB menos). En nuestro caso hemos comprobado que para la identificación de proteínas moonlighting es mejor utilizar PfamB

dado que al identificar dianas potenciales de forma menos restrictiva puede desenmascarar las segundas funciones de las proteínas. Por defecto el servidor InterPro tan solo muestra el "output" PfamA. PfamB debe ser activado por el usuario en la página **http://pfam.xfam.org/search.**

PDB (www.pdb.org)

El PDB (Protein Data Bank) es un archivo de estructuras tridimensionales de proteínas conteniendo algo más de 100.000 estructuras tridimensionales obtenidas por cristalografía de rayos X o NMR (Resonancia Magnética Nuclear).

PIR (http://pir.georgetown.edu/)

Se trata de una base de datos de secuencias y estructuras de proteínas. La Protein Information Resource (PIR) se estableció en 1984 como resultado de la obra de la Dra. Margaret Dayhoff cuyo atlas de secuencias de proteínas y estructura fue la primera colección completa de secuencias de proteínas. En 1974, Dayhoff creó el concepto de la familia de proteínas y superfamilia, definido por la similaridad de secuencia, como un medio de organizar y clasificar las proteínas.

SCOP (http://scop.mrc-lmb.cam.ac.uk/scop/)

La base de datos SCOP es una clasificación jerárquica integral generada manualmente de estructuras de proteínas conocidas, organizado en función de sus relaciones evolutivas y estructurales. La base de datos se divide en cuatro niveles jerárquicos: Clase, Fold, Superfamilia y Familia. Una clase compartiría una arquitectura de dominios. Un fold presenta una importante similitud estructural. Una superfamilia de proteínas presenta un probable origen evolutivo común. Una familia de proteínas presenta un relación evolutiva clara.

GENE ONTOLOGY (GO) (www.geneontology.org)

Una ontología consiste en desarrollar un sistema jerárquico de vocabulario controlado y estructurado para describir con precisión conceptos y sus relaciones. En Biología molecular consiste en la descripción de los productos génicos utilizando términos controlados. Esto evita la anarquía que durante mucho tiempo ha existido en los descriptores de la función de las proteínas (anotación funcional). La ontología la ha venido desarrollando un consorcio (GOC) (Ashburner et al., 2000) que la actualiza mensualmente. La anotación de acuerdo con

GO se denomina GOA. GO representa un cuadruple descriptor: (a) *Biological process* en que participa el producto génico, con dos subniveles: *Broad* (p.e., cell growth, development) y *More specific* (p.e., pattern specification…); (b) *Molecular function* (= *biochemical activity*), también con dos subniveles, *broad* (p.e., enzyme, transporter) y *narrow* (p.e., adenylate cyclase); (c) *cellular components* (lugar de la célula en que el producto génico es activo) y (d) *biological phase* (periodo o etapa en un biological process or cycle). Existen además una serie de *Evidence codes* adicionales acerca del origen de la información:

- IMP = inferred from mutant fenotype
- IGI = from genetic interaction
- IPI = from physical interaction
- ISS = from sequence/structural similarity
- IDA = from direct assay
- IEP = from expression pattern
- IEA = from electronic annotation
- TAS = traceable author statement
- NAS = non-traceable author statement
- NR = not recorded

La base de datos GO contiene en 38.137 términos (23.928 procesos biológicos + 3.050 ubicaciones celulares + 9.467 funciones moleculares). El número de funciones moleculares resulta todavía muy bajo lo que dificulta, como se describirá más adelante, la comparación entre las anotaciones de las bases de datos de secuencias (p.e., ncbi), mayoritariamente con descriptores anárquicos y las bases de datos de interactómica, mayoritariamente con descriptores GO. Finalmente, al GO le faltaría un descriptor de niveles jerárquicos por encima del celular (fisiológico, fenotípico…). Un avance es su reciente incorporación del descriptor *biological phase*.

III.A.2 DISEÑO DE UNA BASE DE DATOS, MultitaskProtDB, DE PROTEÍNAS MOONLIGHT

Además de los ejemplos extraídos de la pequeña cantidad de artículos y revisiones sobre proteínas multifuncionales (Wool, 1996; Jeffery, 1999, 2003, 2004 y 2009; Piatigorsky 2007; Gancedo y Flores, 2008; Nobeli et al, 2009;. Huberts y van der Kiel, 2010; Copley, 2012) se recogieron unas 300 proteínas multifuncionales a partir de una inspección del servidor NCBI PubMed (**http://www.ncbi.nlm.nih.gov/pubmed**). La minería de datos de la literatura se ha realizado utilizando los siguientes términos y palabras clave: *moonlighting proteins; moonlight proteins; multitask proteins; multitasking proteins; moonlight enzymes; moonlighting enzymes; gene sharing*. Una serie de ejemplos se han encontrado por casualidad de algunas revisiones basadas en la función de la proteína, en la bibliografía de genomas secuenciados, en UniProt, etc.

La base de datos ha sido diseñada usando MySQL. El servidor web se ha diseñado con PHP y asistido por PHPRunner, una aplicación que ayuda a generar código PHP y crear archivos, informes, listas y formularios que facilitan el desarrollo de las partes importantes de la web. Estos archivos, informes, etc, también pueden ser generados usando un motor de búsqueda avanzada para permitir una búsqueda más precisa o restringida. Este tipo de procedimiento sirve para limitar la búsqueda al subconjunto de proteínas en las que uno realmente quiere centrar el estudio.

III.B. ALINEAMIENTO DE SECUENCIAS

III.B.1. ALINEAMIENTO MEDIANTE LOS ALGORITMOS BLAST Y PSI-BLAST Y REORDENAMIENTO MEDIANTE BYPASS

El alineamiento de secuencias es extraordinariamente útil para el descubrimiento de información funcional, estructural y evolutiva en las secuencias biológicas. Es importante obtener el mejor alineamiento posible o alineamiento "óptimo" para descubrir esta información. Las secuencias que son muy parecidas, o similares (en muchos casos también se describen como “homólogas”) en la terminología del análisis de secuencias, probablemente tienen la misma función, o una función bioquímica y estructura tridimensional similar en el caso de las proteínas. Aunque en sentido estricto para que dos secuencias de

dos diferentes organismos puedan ser definidas como homólogas han de presentar un ancestro común, lo cual es muy difícil de establecer en la realidad.

El alineamiento indica los cambios que podrían haber ocurrido entre las dos secuencias homólogas y una secuencia ancestro común durante la evolución. Los genes homólogos que comparten un ancestro común y la misma función en ausencia de cualquier evidencia de duplicación de genes se llaman ortólogos. Cuando existe una evidencia de la duplicación de genes, los genes en un linaje evolutivo derivado de una de las copias y con la misma función también se conocen como ortólogos. Las dos copias del gen duplicado y su progenie en el linaje evolutivo se conocen como parálogos. En otros casos, las regiones similares en secuencia pueden no tener un ancestro común, pero pueden haber surgido de forma independiente por dos caminos evolutivos que convergen en la misma función, llamada evolución convergente.

En el presente trabajo hemos utilizado como algoritmos de alineamiento de secuencias el Blast y en el caso de homología remota el PsiBlast (Figura 10), ambos en el servidor del ncbi (**http://www.ncbi.nlm.nih.gov/BLAST**). Al contrario que en el caso de las bases de datos de interactómica, en la búsqueda de funciones moonlight es conveniente utilizar bases de datos curadas en el sentido de no redundantes, de lo contrario el listado de salida presentará, de haberlas, numerosas secuencias homólogas o isoformas de proteínas que arrinconan las dianas interesantes a posiciones muy alejadas y obliga a los investigadores a rastrear largos listados de dianas. Debido a este problema el grupo desarrolló anteriormente un programa, ByPass (Gómez et al., 2008), en que mediante lógica borrosa reordena el listado y sube a posiciones superiores dianas que a pesar de ser verdaderos positivos han acabado en posiciones alejadas (Figura 11).

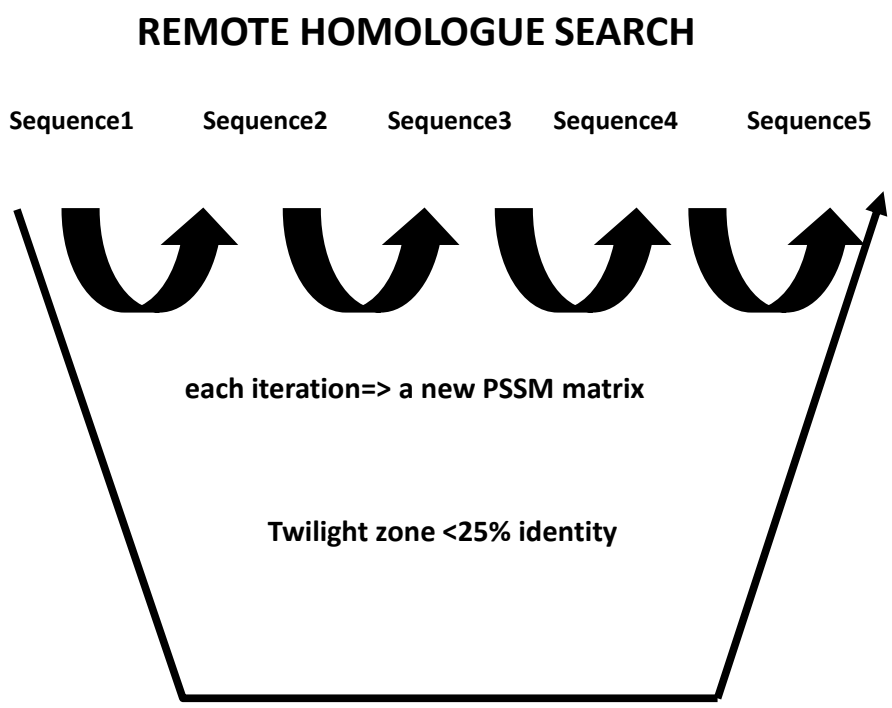


Figura 10. Esquema del procedimiento de iteraciones sucesivas del algoritmo de homología remota PsiBlast

Como se ha indicado, el análisis de homología remota en la base de datos no redundante NCBI se hizo utilizando PSI-BLAST (Altschul et al., 1997), accesible en **http://www.ncbi.nlm.nih.gov/BLAST**. La búsqueda se realizó con los siguientes ajustes y con un máximo de 5 iteraciones con los siguientes parámetros por defecto (*Filter*: "F"; *gap_extend*: "1", *expect*: 10, and *gap_open*: 11). Un problema clásico con los programas de homología es debido a la gran cantidad de secuencias que presentan algunas familias de proteínas, por ejemplo, las proteínas ribosomales, (y que son típicas proteínas multifuncionales) que saturan los "outputs" o listados de secuencias alineadas y dificultan, o incluso esconden, la posible segunda función. Por ello es conveniente utilizar una base de datos no redundante que permite ejecutar el PSI-Blast eliminando todas las entradas al banco de secuencias que comparten la misma secuencia. Por ejemplo en nuestro caso nos ha resultado muy útil el Swiss-Prot. Otro problema es que como Psi-Blast agrupa los resultados colocando los aciertos de acuerdo a sus puntuaciones matemáticas (e-values), y

no por sus “puntuaciones biológicas”, la secuencia diana correcta puede encuentrarse no en las primeras posiciones sino en las inferiores. Por ello, como ya se ha mencionado, el output de salida del PSI-Blast ha sido reordenado por medio del programa ByPass (Gómez et al., 2008) **http://bypass.uab.cat/wiki**. Bypass utiliza la lógica borrosa para reorganizar los resultados de salida del Blast o PSI-Blast y sube hacia posiciones superiores secuencias de proteínas cuya función se puede identificar como posibles verdaderos positivos. El algoritmo de lógica borrosa combina 4 características de la secuencia de las proteínas alineadas por un Blast o PsiBlast para analizar las secuencias del output. Estas características son: perfil de hidropaticidad (Kyte y Doolittle, 1982), perfil de flexibilidad (Karplus y Schulz, 1985), composición (porcentaje) de aminoácidos y longitud de la secuencia.

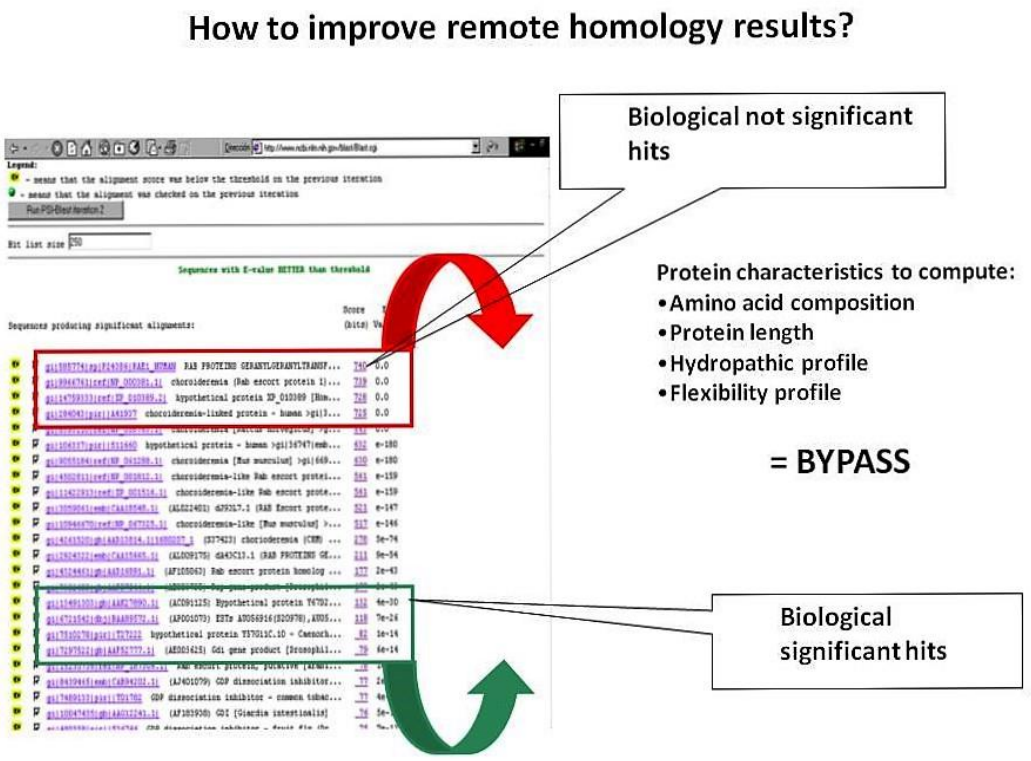


Figura 11. El algoritmo ByPass**,** a partir de un output de Blast o PsiBlast sube a posiciones superiores secuencias de posiciones inferiores, posiciones ambas establecidas previamente por sus e-values

Finalmente, y como alternativa al PsiBlast, se han realizado algunos análisis de homología remota basados en perfiles por cadenas ocultas de Markov mediante el programa HMMER

(Finn et al., 2011) **http://hmmer.janelia.org/**. El resultado es muy prometedor pero todavía no se ha extendido el análisis HMMER a toda la base de datos.

III.B.2. ALINEAMIENTO MULTIPLE DE SECUENCIAS

Los alineamientos múltiples son útiles para la predicción de estructuras de proteínas, motifs y dominios funcionales, identificar aminoácidos clave para la función de la proteína e indispensables para el análisis filogenético. En este sentido, en el presente trabajo se ha utilizado el alineamiento múltiple para sugerir que la conservación de secuencia entre diferentes especies, especialmente de regiones o motifs involucrados en cada una de las funciones, implicaría que la característica de la multifuncionalidad se conserva evolutivamente. Se ha utilizado el programa ClustalW en el servidor del European Bioinformatics Institute, **http://www.ebi.ac.uk/Tools/msa/clustalo/**.

III.C. RASTREO DE BASES DE DATOS DE INTERACTÓMICA

En el presente trabajo hemos propuesto que en las bases de datos de interactómica (PPI) hay mucha información que permitiría identificar proteínas multifuncionales, en muchos casos partners de interactómica descartados como Falsos Positivos por los investigadores experimentales (Gomez et al., 2011). También hemos propuesto que combinar el análisis de similitud de secuencia (por Blast o PsiBlast) con la información de las bases de datos de interactómica facilita la predicción de la función de las proteínas (Espadaler et al, 2008) y de la multifuncionalidad (Hernández et al. 2014b y 2015). Los partners de interactómica de una proteína pueden sugerir la función o funciones de la misma por lo que se denomina "culpabilidad por asociación" ("guilty-by-association"), por lo menos al nivel de "Biological Process" del GO. En el presente trabajo hemos considerado que las bases de datos de interactómica revelan la segunda función de una proteína si la PPI identifica una *Molecular Function* o, en algunos casos, un *Biological Process* de acuerdo con la anotación del GO y está de acuerdo con la función moonlight de la proteína de nuestra base de datos MultitaskProtDB. A continuación y para filtrar e incrementar la precision de la predicción, es aconsejable realizar un análisis de enriquecimiento GO. Para ello, para cada proteína moonlighting protein incluída en APID se capturan los terminos GO de los partners de interacción y se calcula el "GO term enrichment" mediante el programa "GOStat R package"

(Beissbarth and Speed, 2004). Esta función calcula valores p hipergeométricos por sobrerepresentación de cada término GO en la categoría específica entre las anotaciones GO. Este enriquecimiento se ha realizado en el servidor de GOstat, en **http://gostat.wehi.edu.au**, utilizando los parámetros por defecto que proponen los autores en la web del servidor. En nuestro caso seleccionamos como indicadores de verdadera función moonlight aquellos terminus GO con un p-value menor que 0.05, lo que nos permite eliminar bastantes descriptores GO inespecíficos (Gomez et al., 2011). La Figura 12 muestra el esquema del procedimiento utilizado y la Figura 8 muestra en detalle cuatro ejemplos de proteínas moonlighting.

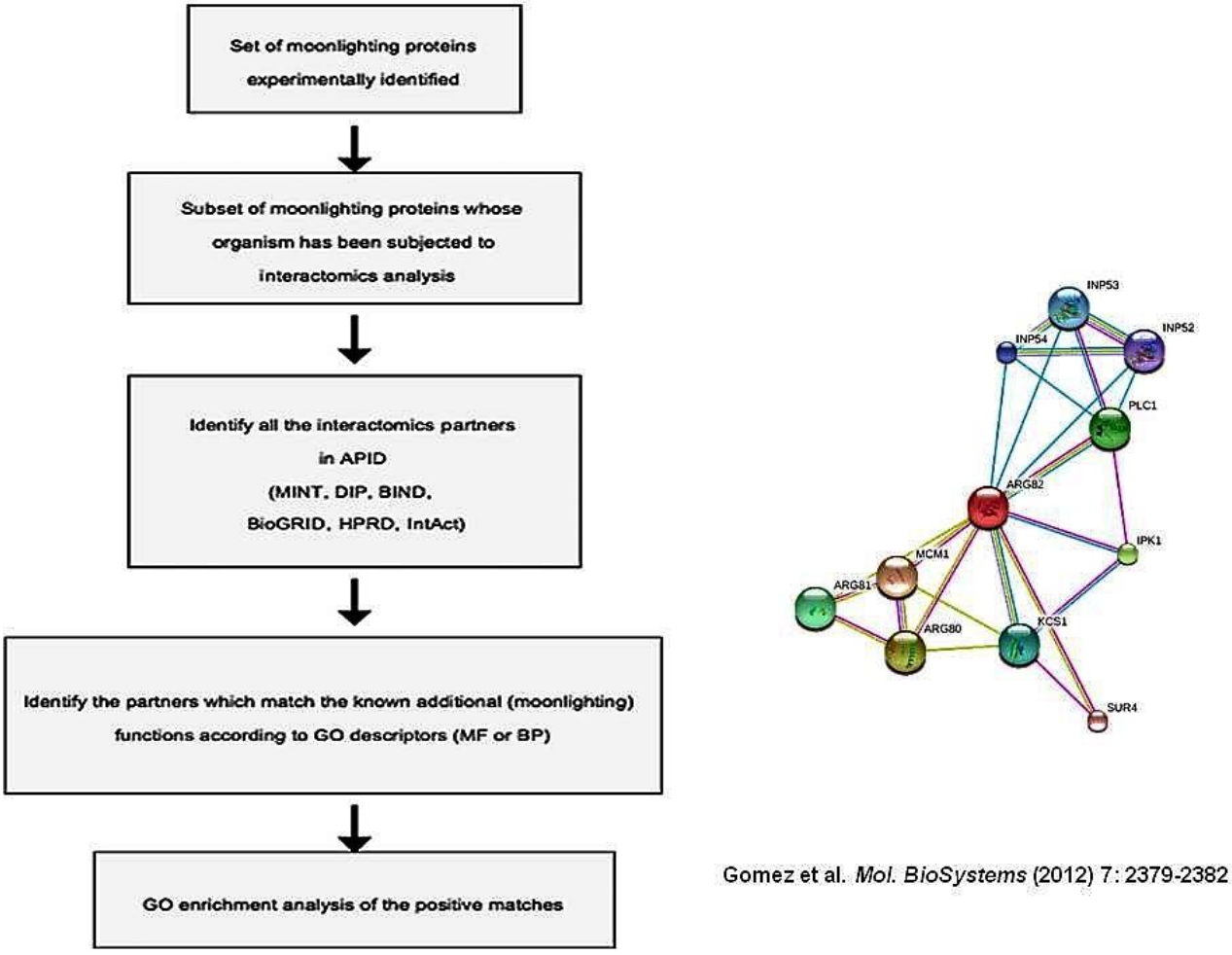


Figura 12. Esquema del procedimiento de predicción de posibles proteínas moonlighting a partir de la información existente en las bases de datos de interactómica (PPIs). En el presente trabajo se ha utilizado preferentemente el servidor de PPIs de APID

III.D. ANÁLISIS DE SECUENCIAS MEDIANTE PROGRAMAS DE IDENTIFICACIÓN DE MOTIVOS Y DOMINIOS (MOTIFS/DOMAINS) FUNCIONALES

La identificación de motivos y de dominios funcionales (Figura 13**)** se ha realizado mediante el servidor InterPro (Mitchell et al., 2015), accesible en **http://www.ebi.ac.uk/Tools/pfa/iprscan/** y descrito en el Apartado III.A.1. Se ha utilizado también el programa Blocks **http://blocks.fhcrc.org** (Henikoff et al., 1999). Como ya se ha mencionado anteriormente, aunque, en teoría, es más aconsejable utilizar bases de datos tipo InterPro; otras bases de datos como Blocks, que al tratarse de una base de datos no curada, que aunque no se hayan actualizado desde el año 2006, pueden servir para identificar como perteneciente al motif, secuencias más alejadas de un patrón canónico (que es lo que esperaríamos encontrar en muchos casos en las proteínas multifuncionales), pues de la semilla con la que se ha generado la PSSM tiene más diversidad de secuencias que aquellas bases de datos en la que la mano del experto ha descartado secuencias que se apartan excesivamente de lo que el "curador" considera una proteína ajustada al canon de la familia funcional. Pues las bases de datos curadas como InterPro se han programado para una identificación precisa del motivo/dominio principal reduciendo en la medida de lo posibles los falsos positivos. La mayoría de los programas de búsqueda de motivo/dominio se mueven en un delicado equilibrio entre especificidad y sensibilidad, que en caso de la búsqueda de proteínas multifuncionales interesa que esté más desplazado hacia la sensibilidad, aunque implique sacrificar algo de sensibilidad. Todos estos métodos construyen sus patrones o perfiles a partir de alineamientos múltiples, pero el algoritmo más corriente, Prosite, busca patrones altamente conservados en un conjunto reducido de proteínas. Es por eso que Prosite tiene dificultades en encontrar el patrón secundario correspondiente a la función moonlighting de la proteína, ya que estos patrones son muy diferentes de la secuencia de consenso canónico y se descartan automaticamente. Por lo tanto, un programa como Blocks que no es tan estricto en la selección de grupos de proteínas preseleccionados puede revelar más patrones relacionados con la función moonlighting. Otro programa, Pfam (Finn et al., 2011) (también lo contiene el servidor InterPro) identifica mediante perfiles HMMs patrones secuenciales comunes a diversas familias de proteínas. Pfam le da al usuario la posibilidad de buscar un conjunto más o menos restringido de dominios mediante dos subprogramas: PfamA y PfamB. PfamA da una salida más restringida, espurgada por eliminación de las dianas menos específicas. PfamB representa una salida menos restringida que contiene

resultados menos específicos por lo que permite identificar posibles funciones moonlight que son filtradas por PfamA. Sin embargo, y por defecto, InterPro sólo muestran la salida PfamA. PfamB tiene que ser activado por el usuario en **http://pfam.xfam.org/search** o calculado localmente.

También se ha utilizado dos servidores más: ELM y MinimotifMiner. ELM es un buscador de motifs secuenciales eucariotas (**http://elm.eu.org**) (Dinkel et al., 2014). Este servidor contiene 200 motifs de secuencia de aminoácidos relacionados con modificaciones post-traduccionales y su función y también con localización celular. El servidor MinimotifMiner (**http://minimotifminer.org**) (Mi et al., 2012) también identifica motifs cortos relacionados con interacción y modificación post-traduccional.

III.E. ANÁLISIS DE CORRELACIÓN DE MUTACIONES

Los análisis de coevolución de aminoácidos permiten identificar aminoácidos clave para la función y evolución de proteínas por la existencia de mutaciones coordinadas; por ejemplo identificar los aminoácidos clave para los sitios funcionales, etc. En el presente trabajo el análisis se llevó a cabo utilizando el servidor Mistic (Mutual Information Server To Infer Coevolution) **http://mistic.leloir.org.ar** (Simonetti et al., 2013). Mistic es un servidor web que proporciona una representación gráfica de la información contenida en un multialineamiento de secuencias. Este programa permite la estimación de la relación coevolutiva entre dos posiciones de aminoácidos en una familia de proteínas a partir de las correlaciones posicionales. De esta manera, el usuario puede identificar aquellas posiciones de los aminoácidos estructural o funcionalmente relevantes (Hernández et al., 2014b, 2015).

III.F. PREDICCIÓN DE QUE LAS PROTEÍNAS MOONLIGHTING PERTENEZCAN A LA CLASE DE LAS PROTEÍNAS INTRINSICAMENTE DESORDENADAS (IDP)

Diversos autores consideran que el pertenecer a la clase de las proteínas intrínsecamente desordenadas facilita la multifuncionalidad por poder adoptar diferentes conformaciones, generalmente locales, para interaccionar con los diferentes partners en cada función (Tompa et al., 2005). Para comprobar si las proteínas multifuncionales pertenecen a la clase de proteínas intrínsecamente desordenadas (IDP) hemos predicho las IDP a partir de sus

secuencias de aminoácidos para un número de proteínas multifuncionales conocidas. Algunas de estas proteínas tienen su estructura 3D resuelta pero sus Regiones Intrínsecamente Desordenadas (IDRs en inglés) no se encuentran bien definidas, precisamente por ser desordenadas, por ejemplo la proteína p53 (Figura 3). Por esta razón, los programas destinados a predecir IDP o IDRs pueden ser útiles para revelar datos estructurales que no se pueden observar a partir de la cristalografía. Hay varios programas para la predicción de IDP/IDR, y que hemos utilizado en el presente trabajo. PrDos (Ishida y Kinoshita, 2007); DisEMBL (Linding et al., 2003); DISOPRED (Ward et al., 2004) y IUPred (Dosztanyi et al, 2005). Estos programas se pueden encontrar en los siguientes servidores web:

PrDos: **http://prdos.hgc.jp/cgi-bin/top.cgi**

DisEMBL: **http://dis.embl.de**

Disopred: **http://bioinf.cs.ucl.ac.uk/disopred/**

Iupred: **http://upred.enzim.hu**

III.G. OTROS PROGRAMAS Y PREDICCIONES UTILIZADOS

Algunas predicciones adicionales que pueden colaborar o corroborar en la identificación de una función moonlight, aunque por ellos mismos no la identifiquen, son los siguientes

III.G.1. PREDICCIÓN DE LA SUBLOCALIZACIÓN CELULAR

Dado que muchas proteínas moonlighting presentan cada función alternativa en distinto compartimento celular se han utilizado programas cuyo output presenta las diferentes localizaciones ordenadas de acuerdo con las correspondientes puntuaciones del método de predicción. Concretamente, en nuestro caso se realizó mediante dos programas, PSORT ((Nakai y Horton, 1999); en http://psort.hgc.jp/) y ProtLoc (Cedano et al, 1997) en **http://bioinf.uab.es/cgi-bin/trsdb/protloc.cgi**). En principio, los dos resultados que aparecen como mejor predicción podrían estar relacionados con las localizaciones del par de funciones de la proteína moonlighting.

III.G.2. PREDICCIÓN DE PRESENCIA DE HÉLICES TRANSMEMBRANA

Los programas para la predicción de los tramos de secuencia de proteínas transmembrana también pueden ayudar a predecir la localización de proteínas y en algunos casos corroborar una posible función biológica. Estos programas son bastante precisos, más que aquellos para la predicción de la estructura secundaria de las proteínas en general. En el presente trabajo hemos utilizado el programa TranMem (Aloy et al., 1997) https://github.com/toniher/TransMem.

IV. RESULTADOS

IV.A. DISEÑO DE UNA BASE DE DATOS DE PROTEÍNAS MOONLIGHT

Las Figuras 14 y 15 muestran la “Home page” y una página del contenido de la base de datos MultitaskProtDB que hemos diseñado y que es la primera base de datos publicada (Hernández et al. 2014a). En Material suplementario S3 se encuentra el contenido total de la base de datos. Las diferentes funciones han sido denominadas como “canonical” o “moonlighting”, pero esto no indica relevancia funcional sino tan sólo el orden histórico del descubrimiento de cada función biológica.

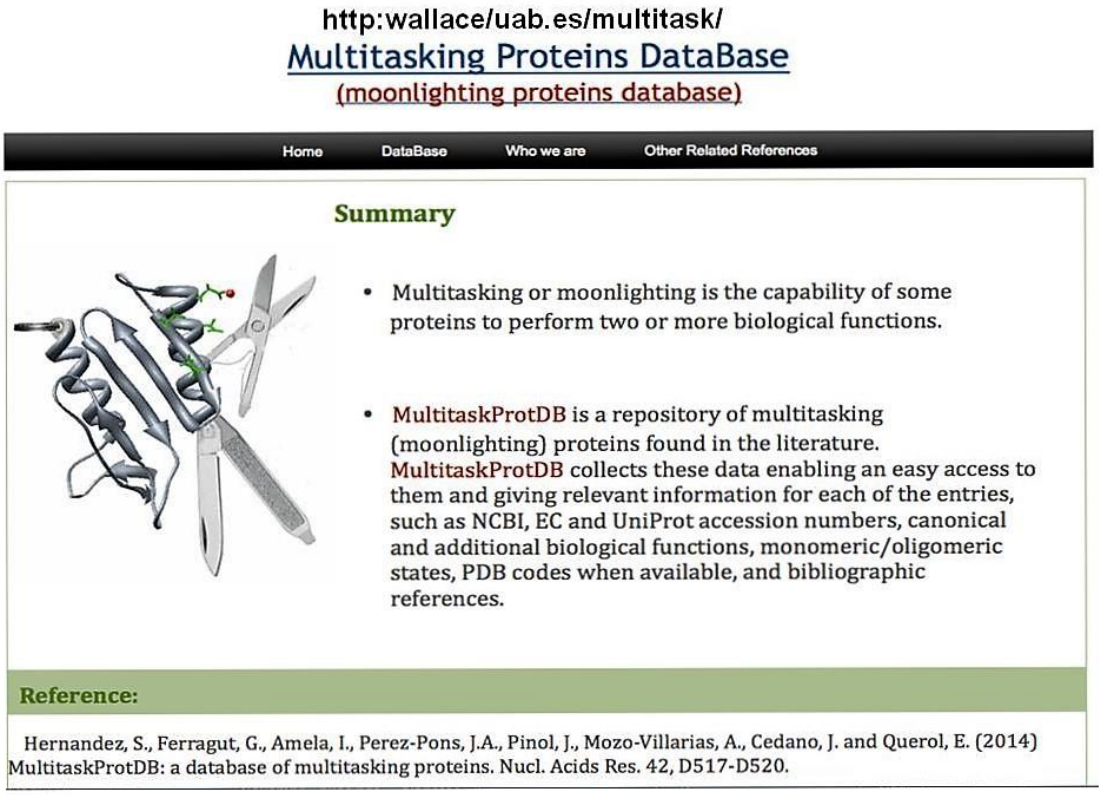
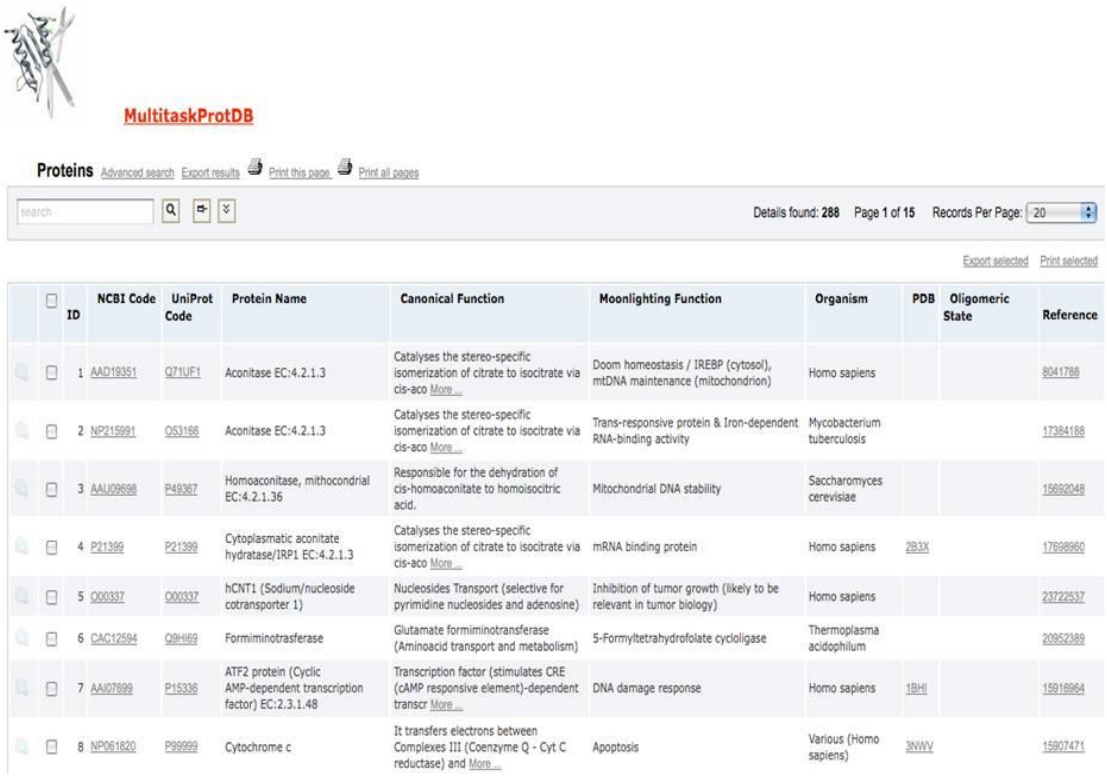


Figura 14. “Home Page” de la base de datos de proteínas moonlighting, MultitaskProtDB, utilizadas en el presente trabajo



Hernández et al. (2014) *Nucleic Acids Res.*42: D517-D520

Figura 15. Ejemplo de la información presentada en una página de la base de datos MultitaskProtDB

Al abrir la base de datos de la página web (**http://wallace.uab.es/multitask/**) se muestra una gran tabla que contiene 288 entradas de proteínas multifuncionales. La web por defecto aparece mostrando 20 entradas (dividido en 15 páginas), pero también permite seleccionar el número de entradas en 10, 30, 50, 100, 500 o todas. Independientemente del número de entradas que se seleccione la página muestra la información sobre todas las proteínas multifuncionales que contiene la base de datos. Hay 12 columnas de la tabla para caracterizar cada proteína. De izquierda a derecha muestra lo siguiente: Columna 1 es un botón clicando en el cual aparecen las características principales que definen la proteína. La Columna 2 (*ID*) permite la selección de entrada con el fin de exportar y manipular su contenido, si es necesario. La Columna 3 indica el número correlativo de la entrada en la tabla. Las Columnas 4 (*Código*) y 5 (UniProt), muestran los números de acceso de NCBI y UniProt respectivamente, que están vinculados a la información de las bases de datos correspondientes. La Columna 6 muestra el nombre de la proteína. Las columnas 7 (*Función*

Canónica) y 8 (*Función moonlight*) muestran las funciones canónicas y la función adicional respectivamente. La Columna 9 (*Organismo*) indica el organismo en el que la proteína moonlighting ha sido identificada. La Columna 10 (*PDB*) vincula la proteína a la correspondiente estructura tridimensional (3D) de la proteína en el PDB, si está disponible. La Columna 11 (*Estado Oligomérico*) indica su estado oligomérico dependiendo de si ha sido determinado (hay proteínas cuya función moonlighting depende de si están en estado mono u oligomérico, por ejemplo la GAPDH). La Columna 12 (*Referencia*) proporciona un enlace a la referencia bibliográfica en PubMed. La página web también facilta pantallas de visualización, impresión y búsqueda. Por otra parte, se puede la exportar fácilmente toda la base de datos, o entradas seleccionadas, mediante la obtención de un archivo en diferentes formatos de datos para facilitar su posterior análisis, por ejemplo en Excel, Word, CSV o XML.

Una visión general de la base de datos muestra que la mayoría de las proteínas multifuncionales presentan dos funciones biológicas. Como era de esperar, la mayoría de los pares de funciones de la proteína multifuncional corresponden a diferentes compartimentos celulares cuando se trata de proteínas eucariotas (Tabla 2a). La Tabla muestra las frecuencias de los 26 pares de funciones, canónica y moonlighting (de acuerdo a descriptores generales como el del Gene Onthology, por ejemplo enzima y factor de transcripción; enzima y adhesión celular, etc.). El par más frecuente es *enzima-factor de transcripción* (o *proteína de unión a ácidos nucleicos*) con 66 de 288 proteínas multifuncionales. Hay una falta de proteínas integrales de membrana, lo cual es lógico porque las proteínas monnlighting por lo general tienen cada función en diferentes compartimentos celulares, lo que sería problemático para las proteínas de membrana. Sin embargo, el segundo par de funciones más abundante corresponde a *enzima-proteína de adhesión* de microorganismos patógenos (46 de 288 proteínas multifuncionales). Como ya se ha mencionado en la Introducción, es un hecho bien conocido que muchos patógenos utilizan enzimas metabólicos que no son proteínas de membrana como elementos de adhesión al huésped. Esto requiere su secreción o localización en la membrana a través de diferentes mecanismos que se desconocen (Henderson y Martin, 2011 y 2013). Hay también un gran número de casos del par *enzima/enzima* y *enzima-proteína estructural*, en este último debido a las proteínas del cristalino de las que se conocen unas 30 que son ejemplos de proteínas multifuncionales. En el caso de *enzima/enzima* tan sólo se consideran aquellas proteínas

moonlighting que presentan dos actividades enzimáticas distintas en dos diferentes centros activos. Se conocen muchas enzimas con dos actividades enzimáticas en el mismo centro activo, pero no se suelen considerar casos de verdadero moonlighting, sino más bien de promiscuidad enzimática (Nobeli et. al., 2009).

Pero si en vez de considerar globalmente todas las proteínas moonlighting se consideran tan solo las 102 proteínas procariotas la tabla resultante, Tabla 2b, presenta algunas variaciones. El número total de proteínas moonlighting procariotas es de 102.

Como puede observarse las dos clases principales siguen siendo *enzima-factor de transcripción* y *enzima/proteína de adhesión* pero alterando las respectivas posiciones, cosa esperable dado el gran número de proteínas moonlighting bacterianas relacionadas con la adhesión al huésped.

Finalmente, se ha obtenido la distribución de los pares de funciones para las levaduras. Son tan solo 22 proteínas moonlighting y no se muestra la tabla. Tan solo indicar que el par más abundante es el de *enzima-factor de transcripción,* con 10 casos*,* seguido de *enzima-enzima* y *enzima-proteína estructural* con 3 casos cada uno.

Algunos pares de funciones en la Tabla 2a presentan un asterisco. Son aquellos en que la función moonlighting corresponde a lo que en el Gene Ontology (GO) se considera *Biological Process* o incluso de categoría superior, no incorporada por GO (por ejemplo organismo, sistema…). Corresponden a funciones supracelulares, fisiológicas y mecanismos más complejos con muchos componentes adicionales y por ello más difíciles de identificar y separar las funciones de sus componentes.

Más adelante, en el Apartado IV.E. se comentarán los aspectos evolutivos que aparecen tras el análisis de las proteínas de esta base de datos.

Tabla 2a. Número de proteínas agrupadas de acuerdo con las principales clases funcionales (Canónica/Moonlight) presentes en la base de datos de proteínas moonlighting MultitaskDB

PAR DE CLASES FUNCIONALES	NÚMERO DE PROTEÍNAS
Enzima/Factor de transcripción (o proteína DNA/RNA binding)*	66
Enzima/Proteína de Adhesión	46
Enzima/Enzima	41
Enzima/Proteína structural (ensamblaje citoesqueleto, cristalino…)	34
Ambas funciones Factor de transcripción o proteína NA binding*	27
Chaperona/Activador de CItoquina*	8
Proteína estructural/Proteína estructural	7
Enzima/Inhibidor enzimático	6
Receptor/Receptor *	5
Enzima/Chaperona*	5
Activador enzimático/Inhibidor enzimático	5
Proteína estructural/ Proteína de Adhesión	5
Enzima/Citoquina*	4
Enzima/Proteína apoptosis*	4
Enzima/Proteína de señalización*	4
Factor de transcripción /Proteína estructural	3
Factor de transcripción /Citoquina*	3
Enzima/Factor de Viruencia (no adhesina)	2
Chaperona/Proteína DNA binding	2
Chaperona/Toxina	2
Factor de transcripción/ Proteína de Adhesión	2
Proteína estructural/Proteína DNA binding*	2
Receptor/Enzima	2
Chaperona/Proteína de biofilm	1
Proteína estructural/Proteína de membrana*	1
Enzima/Proteína de halotolerancia*	1

Tabla 2b. Número de proteínas procariotas agrupadas de acuerdo con las principales clases funcionales (Canónica/Moonlight) presentes en la base de datos de proteínas moonlighting MultitaskDB

PAR DE CLASES FUNCIONALES	NÚMERO DE PROTEÍNAS
Enzima/Proteína de Adhesión	46
Enzima/Factor de transcripción (o proteína DNA/RNA binding)*	18
Chaperona/Activador de CItoquina*	8
Enzima/Enzima	3
Ambas funciones Factor de transcripción o proteína NA binding*	7
Enzima/Proteína structural (ensamblaje citoesqueleto, cristalino…)	5
Enzima/Chaperona*	3
Enzima/Proteína apoptosis*	2
Chaperona/Toxina	2
Enzima/Factor de Viruencia (no adhesina)	2
Factor de transcripción /Proteína estructural	1
Enzima/Inhibidor enzimático	1
Receptor/Receptor *	1
Enzima/Citoquina*	1
Proteína estructural/Proteína DNA binding*	1
Chaperona/Proteína de biofilm	1
Proteína estructural/Proteína estructural	0
Activador enzimático/Inhibidor enzimático	0
Proteína estructural/ Proteína de Adhesión	0
Enzima/Proteína de señalización*	0
Factor de transcripción /Citoquina*	0
Chaperona/Proteína DNA binding	0
Factor de transcripción/ Proteína de Adhesión	0
Receptor/Enzima	0
Proteína estructural/Proteína de membrana*	0
Enzima/Proteína de halotolerancia*	0

IV.B. ANÁLISIS BIOINFORMATICO A PARTIR DE LA SECUENCIA DE LAS PROTEÍNAS

Todos los siguientes resultados se refieren al análisis de las proteínas multifuncionales contenidas en la base de datos MultitaskProtDB. Como se ha mencionado anteriormente las diferentes funciones han sido etiquetadas como *canónica* (la primera identificada) o *moonlighting* (o *multitarea* o *multifuncional*), la función posteriormente determinada, pero esto no tiene relevancia biológica y simplemente se refiere al orden histórico del descubrimiento de la función biológica. Hay diferentes maneras de asignar una función a una secuencia de la proteína cuya función es desconocida, pero los métodos más utilizados lo hacen por la aplicación de la propiedad transitiva. Si una proteína de función desconocida tiene un cierto grado de similitud con una anotada, entonces se asume que comparten la misma función. Pero, si tenemos mucha información redundante (es decir, un gran conjunto de secuencias relacionadas) esta redundancia puede ser utilizada para inferir función por medio de la extracción de patrones o perfiles. En este caso, podemos usar estos patrones (los denominaremos como “motifs” y dominios) para inferir la función. El patrón extraído también se puede utilizar para identificar los aminoácidos esenciales para su función. Otra forma de identificar los aminoácidos importantes para la función de una proteína es por medio del modelado de su estructura a nivel tridimensional. Y también identificar sus partners de interacción. Todas estas estrategias (alineamiento de secuencias, identificación de motifs, identificación de partners, modelado 3D y sus combinaciones) han sido utilizadas en el presente trabajo para inferir la función de las proteínas multifuncionales.

IV.B.1. ANÁLISIS DE HOMOLOGÍA/ HOMOLOGÍA REMOTA

Los algoritmos Blast, y especialmente Psi-Blast, pueden detectar proteínas multifuncionales por presentar más de un tramo que se alinea a dos (o más) diferentes secuencias diana. Por ejemplo, la Figura 16 muestra como ejemplo la enzima bifuncional *dihydropteroate synthase* (DHPS) y *2-amino-4-hidroxi-6-hydroxymethyldihydropterine pyrophosphokinase* (HPPK). Este ejemplo también representa un caso de probable fusión de dos genes, o dominios, que conducen a una proteína multifuncional.

MAPPING THE SECOND FUNCTION: SEQUENCE SIMILARITY

Color	Protein name	Uniprot	Functions
Red	6-hydroxymethyl-7,8-dihydropterin pyrophosphokinase/ 7,8-dihydropteroate synthase	Q1ENB6	DHPS / HPPK
Black	Folic acid synthesis protein FO	P53848	DHPS / HPPK
Blue	Dihydropteroate synthase	Q81VW8	DHPS
Green	2-amino-4-hydroxy-6-hydroxymethyldihydropteridine pyrophosphokinase	P26281	HPPK

A

```
tr|Q1ENB6|Q1ENB6_ARATH
sp|P53848|FOL1_YEAST
Identity: 34,38 %
HPPK and DHPS functions

tr|Q1ENB6|Q1ENB6_ARATH  ------------------------------MDFTSLETT------TFEEV 14
sp|P53848|FOL1_YEAST    VGVSCIREPREIAMVNIPYLSSIHESSDIKFQLSSSQNTPIEGKNTWKRA 300
                                                    ::::* :.*      *::..
tr|Q1ENB6|Q1ENB6_ARATH  VIALGSNVGNRMNNFKEALRLMK-DYGISVTRHSCLYETEPVHVTDQPRF 63
sp|P53848|FOL1_YEAST    FLAFGSNIGDRFKHIQMALQLLSRERKTVKLRNISSIFESEPMYFKDQTPF 350
                        .:*:***:*:*::::: **:*:. :  :.: . *.::*:**::..**. *
tr|Q1ENB6|Q1ENB6_ARATH  LNAAIRGVTKLKPHELLNVLKKIEKEMGREENGLRYGPRPLDLDILFYG- 112
sp|P53848|FOL1_YEAST    MNGCVEVETLLTPSELLKLCKKIEYEELQRVKHFDNGPRTIDLDIVMFLN 400
                        :*..:.  * *.* ***:: **** *  :. : :  ***.:****:::
tr|Q1ENB6|Q1ENB6_ARATH  ----KHKIISDKLIIPHERIWERPFVLAPLVDLLGTEDIDNDKIVAYWHS 158
sp|P53848|FOL1_YEAST    SAGEDIIVNEPDLNIPHPRMLERTFVLEPLCELISPVHLHPVTAEPIVDH 450
                          .  : . .* *** *: **.*** ** :*:.. .:.  .  .  .
tr|Q1ENB6|Q1ENB6_ARATH  LSMHSGGIFQAWERLGGESLLGKDGIIQRVIPIG--DHLWDFS------- 199
sp|P53848|FOL1_YEAST    LKQLYDKQHDEDTLWKLVPLPYRSGVEPRFLKFKTATKLDEFTGETNRIT 500
                        *.  .  .:      .*  :.*:  *.: :    :* :*:
tr|Q1ENB6|Q1ENB6_ARATH  -KKTYVMGILNLTPDSFSDGG-KFQSVDTAVSRVRSMISEG-----VDII 242
sp|P53848|FOL1_YEAST    VSPTYIMAIFNATPDSFSDGGEHFADIESQLNDIIKLCKDALYLHESVII 550
                          . **:*.*:* ********* :* .::: :. : .: .:.      **
tr|Q1ENB6|Q1ENB6_ARATH  DIGAQSTRPMASRISSQEEIDRLIPVLKVVRGMAEMKG--KLISVDTFNS 290
sp|P53848|FOL1_YEAST    DVGGCSTRPNSIQASEEEEIRRSIPLIKAIRESTELPQDKVILSIDTYRS 600
                        *:*. **** : : *.:*** * **::*.:*  :*:      ::*:**:.*
tr|Q1ENB6|Q1ENB6_ARATH  EVALEAIRNGADILNDVSGGSLDENMHKVVADS-DVPYMIMHMRGDPCTM 339
sp|P53848|FOL1_YEAST    NVAKEAIKVGVDIINDISGGLFDSNMFAVIAENPEICYILSHTRGDISTM 650
                        :** ***: *.**:**:*** :*.**. *:*:. :: *:: * *** .**
tr|Q1ENB6|Q1ENB6_ARATH  Q-------------------------NKENLEYNEICKDVATELYERVR 363
sp|P53848|FOL1_YEAST    NRLAHYENFALGDSIQQEFVHNTDIQQLDDLKDKTVLIRNVGQEIGERYI 700
                        :                    .:  : . : ::*. *: **
tr|Q1ENB6|Q1ENB6_ARATH  EAELSGIPAWRIMIDPGIGFSKGIDHNLDIVMELPKIR-------EEMAK 406
sp|P53848|FOL1_YEAST    KAIDNGVKRWQILIDPGLGFAKTWKQNLQIIRHIPILKNYSFTMNSNNSQ 750
                        :*  .*:  *:*:****:**:*  .:**:*:  .:* ::          : ::
tr|Q1ENB6|Q1ENB6_ARATH  KSIGLSHAPILIGPSRKRFLGDICGRPEASERDAATVACVTAGILKGANI 456
sp|P53848|FOL1_YEAST    VYVNLRNMPVLLGPSRKKFIGHITKDVDAKQRDFATGAVVASCIGFGSDM 800
                          :.* : *:*:*****:*:*.*   :*.:** ** * *:: *  *:::
tr|Q1ENB6|Q1ENB6_ARATH  IRVHNVRDNVDAARLCDAMMTKRFKNVD 484
sp|P53848|FOL1_YEAST    VRVHDVKNCSKSIKLADAIYKGLE---- 824
                        :***:*::  .: :*.**: .
```

B

```
tr|Q1ENB6|Q1ENB6_ARATH          tr|Q1ENB6|Q1ENB6_ARATH
sp|P26281|HPPK_ECOLI            tr|Q81VW8|Q81VW8_BACAN
Identity: 32,91%                Identity: 40%
HPPK function                   DHPS function

tr|Q1ENB6|Q1ENB6_ARATH  MDFTSLETTTFEEVVIALGSNVGNRMNNFKEALRLMKDYGIS-VTRHSCL 49
sp|P26281|HPPK_ECOLI    ----------MTVAYIAIGSNLASPLEQVNAALKALGDIPESHILTVSSF 40
                                :  . **:***:.. :::.: **: : *   * :   *.:
tr|Q1ENB6|Q1ENB6_ARATH  YETEPVHVTDQPRFLNAAIRGVTKLKPHELLNVLKKIEKEMGREENGLRY 99
sp|P26281|HPPK_ECOLI    YRTPPLGPQDQPDYLNAAVALETSLAPEELLNHTQRIELQQGRVRKAERW 90
                        *.* *:   *** :****:   *.* *.****  ::** : ** .:. *:
tr|Q1ENB6|Q1ENB6_ARATH  GPRPLDLDILFYGKHKIISDKLIIPHERIWERPFVLAPLVDLLGTEDIDN 149
sp|P26281|HPPK_ECOLI    GPRTLDLDIMLFGNEVINTERLTVPHYDMKNRGFMLWPLFEIAPELVFPD 140
                        ***.*****:::*:. * :::* :**  : :* *:* **.::     : :
tr|Q1ENB6|Q1ENB6_ARATH  DKIVAYWHSLSMHSGGIFQAWERLGESLLGKDGIIQRVIPIGDHLWDFSK 200
sp|P26281|HPPK_ECOLI    GEMLR--QILHTRAFDKLNKWD--------------YDLRCGEYTLNLNE 21
                        .:::   : *  :: . :: *                :  *::  ::.:
tr|Q1ENB6|Q1ENB6_ARATH  KTYVMGILNLTPDSFSDGGKFQSVDTAVSRVRSMISEGVDIIDIGAQSTR 250
tr|Q81VW8|Q81VW8_BACAN  KTLIMGILNVTPDSFSDGGSYNEVDAAVRHAKEMRDEGAHIIDIGGESTR 71
                        ** :*****:*********.::.**:** :.:.* .**..*****.:***
tr|Q1ENB6|Q1ENB6_ARATH  PMASRISSQEEIDRLIPVLKVVRGMAEMKGKLISVDTFNSEVALEAIRNG 300
tr|Q81VW8|Q81VW8_BACAN  PGFAKVSVEEEIKRVVPMIQAVSKEVKLP---ISIDTYKAEVAKQAIEAG 118
                        *  :::* :***.*::*:::.*   .::     **:**:::*** :**.
tr|Q1ENB6|Q1ENB6_ARATH  ADILNDVSGGSLDENMHKVVADSDVPYMIMHMRGDPCTMQNKENLEYNEI 350
tr|Q81VW8|Q81VW8_BACAN  AHIINDIWGAKAEPKIAEVAAHYDVPIILMHNR---------DNMNYRNL 159
                        *.*:**: *.. : :: :*.*. *** ::** *         :*::*.::
tr|Q1ENB6|Q1ENB6_ARATH  CKDVATELYERVREAELSGIPAWRIMIDPGIGFSKGIDHNLDIVMELPKI 400
tr|Q81VW8|Q81VW8_BACAN  MADMIADLYDSIKIAKDAGVRDENIILDPGIGFAKTPEQNLEAMRNL--- 206
                          *: ::**: :: *: :*:   .*::******:*  ::**: : :*
tr|Q1ENB6|Q1ENB6_ARATH  REEMAKKSIGLSHAPILIGPSRKRFLGDICGRPEASERDAATVACVTAGI 450
tr|Q81VW8|Q81VW8_BACAN  ------EQLNVLGYPVLLGTSRKSFIGHVLDLP-VEERLEGTGATVCLGI 249
                              :.:.   *:*:*.*** *:*.: . * ..**  .* * *  **
tr|Q1ENB6|Q1ENB6_ARATH  LKGANIIRVHNVRDNVDAARLCDAMMTKRFKNVD 484
tr|Q81VW8|Q81VW8_BACAN  EKGCEFVRVHDVKEMSRMAKMMDAMIGKGVK--- 280
                          **.::*:***:*::    *:: ***: * .*
```

Figura 16. Ejemplo de que el programa Psi-Blast puede detectar proteínas multifuncionales por alineamiento a dos (o más) diferentes secuencias diana. En este caso se trata de la

enzima bifuncional *dihydropteroate sinthase* (DHPS) y *2-amino-4-hidroxi-6-hydroxymethyldihydropterine pyrophosphokinase* (HPPK)

El programa de homología remota Psi-Blast es especialmente adecuado para identificar las proteínas multifuncionales porque, debido al algoritmo que utiliza PSSM e iteraciones sucesivas rediseñando en cada iteración la matriz de alineamiento, puede identificar tramos de secuencia de aminoácidos conservados de diferentes dominios (Gómez et al., 2003; Khan et al, 2012). Por supuesto también puede llevarnos a un resultado erróneo debido a la "corrupción" de la PSSM. Como también ocurre en las búsquedas en bases de datos de interactómica (PPIs, ver la siguiente sección), el resultado del análisis de Blast o PsiBlast se presenta como una larga lista de proteínas dianas y, a priori, el investigador no sabe cuál o cuáles de ellas serán verdaderos positivos. Esto requiere por parte del investigador un posterior cuidadoso análisis de las diferentes predicciones y de los datos experimentales de los que disponga.

En nuestro caso, como se ha indicado en métodos se alinearon las secuencias de las proteínas de nuestra base de datos mediante el algoritmo PsiBlast. Se realizaron cinco iteraciones, y las anotaciones funcionales se inspeccionaron con el fin de comprobar si las dianas, o los tramos con alta similitud a diferentes proteínas diana, contenían anotaciones que correspondían a la función canónica y la función moonlighting de las proteínas de nuestra base de datos. Un problema clásico con los programas de homología se debe al gran número de secuencias de familias de proteínas, por ejemplo, de las proteínas ribosomales que colapsan los resultados de los outputs y trasladan la función moonlighting a posiciones muy alejadas en el listado (o incluso las ocultan si el investigador pone por defecto un umbral de corte razonable, por ejemplo 100 secuencias). Utilizar bases de datos no redundantes y manualmente curadas como SWISS-PROT, permite una mejora en el análisis Psi-Blast. O en el caso de no realizar un análisis automático como el que hemos llevado a cabo con toda la base de datos, que el investigador seleccione las secuencias a utilizar en la segunda iteración. Como se ha descrito en Métodos, el programa Bypass (que traslada dianas a posiciones superiores del output) también ayuda a identificar verdaderos aciertos, ya que hemos encontrado que entre las secuencias que se han "movido hacia arriba" con mejores puntuaciones por ByPass existen aciertos correspondientes a las funciones moonlight; sin embargo, en la mayoría de los casos ByPass no las desplaza a exactamente a la primera y

segunda posiciones. Por ello, utilizar Bypass para reordenar el output de PsiBlast, aunque representa una ayuda no nos dispensa de tener que hacer un análisis manual cuidadoso del listado del output. Hay ejemplos, tales como las enolasas, en el que la adaptación a realizar la función adicional (la unión al plasminógeno del huésped en este caso) puede ser un proceso más común de lo esperado y a menudo sólo implica el rediseño de una pequeña porción de la secuencia de la proteína. Por ello proteínas que presentan secuencias muy similares pero sólo con pequeños cambios locales una puede tener una función moonlighting y otra no. En estos casos programas como Bypass que hacen un cálculo global de la similitud entre las secuencias de aminoácidos de las proteínas analizadas, pueden fallar en la detección de pequeños cambios locales en ellas. En conclusión, dependiendo del tipo de criterios utilizados por Bypass, no siempre podemos garantizar que el enriquecimiento en las primeras posiciones en el listado de resultados corresponderá a los casos de verdaderos positivos de las funciones moonlighting.

En el presente trabajo hemos considerado como resultados positivos de Psi-Blast aquellos que describen la función en un sentido amplio en cualquier posición de la lista de resultados obtenidos. Para los casos en los que la proteína moonlight es enzima y factor de transcripción (el par más abundante que encontramos en MultitaskProtDB), podemos considerar como una buena predicción que la función moonlighting se prediga como *Transcription factor*, en general, o incluso como *Zinc finger domain*. En la Columna 4 de la Tabla 3 se muestran algunos ejemplos de proteínas identificadas por homología remota (en el Material suplementario se encuentra la Tabla S4 con todas las proteínas analizadas). De las 288 proteínas moonlighting de la base de datos MultitaskProtDB, Psi-Blast identifica la segunda función en aproximadamente el 41% de los casos cuando se considera un resultado positivo. Como ya se ha mencionado consideramos una identificación positiva en un sentido más amplio que en el estricto de la anotación GO para *Molecular function*, en muchos casos identificar el *Biological process* aporta información interesante sobre una posible función biologica. Esta flexibilidad en la identificación es debida a que las anotaciones en las bases de datos de secuencias como NCBI, etc, son bastante ambiguas y anárquicas (por ejemplo existen anotaciones como *17kDa protein*…). Sin llegar a la anomalía de este ejemplo, el análisis de las anotaciones presenta dificultades. Es de esperar que la especificidad en la identificación de la función moonlighting mejorará utilizando Blast/PsiBlast con anotación GO (por ejemplo existe Blast2GO (Gotz et al., 2008)). Aunque la anotación GO compacta los

aciertos a unas pocas categorías funcionales, reduciendo el número de entradas, hemos encontrado que la sensibilidad en la detección de algunas funciones moonlighting bajan. Entre otros motivos hay que mencionar que las anotaciones *Molecular function* existentes en la base de datos GO actual contienen tan sólo unas 9500 funciones moleculares, por lo que muchas o la mayoría de las 40.000 funciones esperables no se pueden encontrar en GO y además muchos descriptores son muy vagos.

Por último, cabe mencionar que tan sólo una tercera parte de las 249 proteínas de nuestra base de datos para las que hay datos de interactómica son identificadas como moonlighting por ambos métodos, PsiBlast e Interactómica (ver Tabla 4a y b y Apartado IV.B.3)

Por supuesto hay un cierto número de casos en que tan sólo predice la función moonlight mediante uno de los dos métodos, homología remota e interactómica, o por ninguno, en la Tabla 4 se presentan algunos ejemplos. Uno de los motivos es que las bases de datos de interactómica, PPIs, son muy incompletas, de hecho no existen datos para numerosas especies. En nuestro caso de las 288 proteínas de la base de datos MutitaskProtDB tan solo 249 presentaban datos de interactómica. Por otra parte en un cierto número de casos los métodos de interactómica no pueden detectar algunas clases de proteínas, por ejemplo las de membrana, las muy grandes o muy pequeñas, etc, (Jensen & Bork, 2008). Y ya se ha mencionado anteriormente las limitaciones de la anotación por homología así como los numerosos descriptores que no han seguido las normas semánticas.

TABLA 3. Ejemplos de predicción de las proteínas moonlighting combinando interactómica y análisis de secuenciasmediante PsiBlast/ByPass

CANONICAL FUNCTION	MOONLIGHTING FUNCTION	PPI PARTNERS (only some hits are shown)	PsiBlast/ByPass OUTPUT (only some hits are shown)
Phosphoglucose isomerase	- Neurotrophic factor - Neuroleukin - Autocrine motility factor - Nerve growth factor	- GO:4842 Autocrine motility factor receptor 2 - GO: 31994 Insulin-like growth factor binding protein 3	gil17380385 - Glucose 6 Phosphate isomerase - Autocrine motility factor - Neuroleukin
Pyruvate kinase	Tyroid hormone-binding rotein	- GO:3707 Nucelar hormone receptor member nhr-111 - GO: 9914 Sex hormone binding globulin - GO: 5179 Atrial natriuretic factor	gil20178296 - Pyruvate kinase isozymes - Cytosolic yhyroid hormone-binding protein
Ribosomal protein S3 (human)	Apurinic/apirymidinic endonuclease	- GO: 31571 DNA damage binding protein 1 - GO: 3735 S27 ribosomal protein	gil290275 - Ribosomal protein S3 - AP endonuclease DNA repair
Ure2	Glutathione peroxidase	GO: 6808 Nitrogen regulatory protein	gil173152; gi449015276 - Glutathione transferase-like protein - Nitrogen catabolite repression transcriptional regulator
P0 ribosomal protein	DNA repair	GO:6281, FACT complex subunit SSRP1	
Vhs3 - phosphopantothenoylcysteine decarboxylase subunit Vhs3	Regulator of serine/threonine protein phosphatase	GO:4724, Serine/threonine-protein phosphatase PP-Z1	gi|254572327|ref|XP_002493273.1| Negative regulatory subunit of the protein phosphatase 1 Ppz1p
Epsin	Organizing mitotic membranes/influencing spindle assembly	GO:7067, Cell division control protein 2 homolog	gi|2072301|gb|AAC60123.1| mitotic phosphoprotein 90
alpha-crystallin A chain	Heat-shock protein	GO:6986, Heat shock protein beta-1	gi|1706112|sp|P02489.2|CRYAA_HUMAN RecName: Full=Alpha-crystallin A chain; AltName: Full=Heat shock protein beta-4
Hexokinase	Transcriptional regulation	GO:16563, Metallothionein expression activator	gi|254573908|ref|XP_002494063.1| Non-essential protein of unknown function required for transcriptional induction
Ribosomal protein L7	Autogenous regulation of translation	GO:6414, 60S ribosomal protein L7a	gi|339256006|ref|XP_003370746.1| eukaryotic translation initiation factor 2C 2
PIAS1 (E3 SUMO-protein ligase PIAS1)	Activation of p53	GO:7569, Cellular tumor antigen p53	gi|58176991|pdb|1V66|A Chain A, Solution Structure Of Human P53 Binding Domain Of Pias-1

TABLA 4. Algunos ejemplos de predicción simple (la función moonlighting es identificada sólo por PsiBlast/ByPass o por interactómica)

CANONICAL FUNCTION	MOONLIGHTING FUNCTION	PPI PARTNERS (only some hits are shown)	BYPASS OUTPUT (only some hits are shown)
DNA primase	DNA and RNA polymerase	GO:5658=alpha DNA polymerase: primase complex	No results
Thymidine phosphorylase	Platelet-derived endothelial cell growth factor	No results	gi|67477361|sp|P19971.2|TYPH_HUMAN RecName: Full=Thymidine phosphorylase; Short=TP; AltName: Full=Gliostatin; AltName: Full=Platelet-derived endothelial cell growth factor; Short=PD-ECGF; AltName: Full=TdRPase; Flags: Precursor; gi|123981106|gb|ABM82382.1| endothelial cell growth factor 1 (platelet-derived) [synthetic construct]; gi|148672399|gb|EDL04346.1| endothelial cell growth factor 1 (platelet-derived) [Mus musculus]
Dihydrofolate-reductasethymidylate synthase	Nucleotide biosynthesis	No results	gi|254572255|ref|XP_002493237.1| Thymidylate synthase, required for de novo biosynthesis of pyrimidine deoxyribonucleotides [Komagataellapastoris GS115]
cPrxI (Peroxiredoxin TSA1)	Chaperon & Phospholipase aiPLA2	GO:6457=protein folding GO:6950=response to stress	No results
Acetohydroxy acid reductoisomerase (ILV5)	Maintenance of rho and mitochondrial DNA (& implication in the distribution of mtDNA molecules into nucleoids)	GO:2=mitochondrial genome maintenance	No results
Inositol P kinase	Transcriptional regulation	GO:0001104= RNA polymerase II transcription cofactor activity	No results
Mycolyl transferase	Fibronectin binding proteins	No results	gi|13431273|sp|O52972.1|A85C_MYCAV RecName: Full=Diacylglycerol acyltransferase/mycolyltransferase Ag85C; Short=DGAT; AltName: Full=Acyl-CoA:diacylglycerol acyltransferase; AltName: Full=Antigen 85 complex C; Short=85C; Short=Ag85C; AltName: Full=Fibronectin-binding protein C; Short=Fbps C; Flags: Precursor; gi|433633148|ref|YP_007266775.1| Secreted antigen 85-C FbpC (85C) (antigen 85 complex C) (Ag58C) (mycolyl transferase 85C) (fibronectin-binding protein C) [Mycobacterium canettii CIPT 140070017];
Paramyxovirus hemagglutinin	Neuraminidase	No results	gi|151935431|gb|ABS18756.1| hemagglutinin-neuraminidase [Sendai virus]; gi|56378307|dbj|BAD74223.1| Hemagglutinin-Neuraminidase protein [Sendai virus] gi|152002457|dbj|BAF73483.1| Hemagglutinin-Neuraminidase protein [Sendai virus]; gi|193888390|gb|ACF28540.1| hemagglutinin-neuraminidase [Human parainfluenza virus 3];

Cabe resaltar que algunos de los casos en que el programa de homología remota PsiBlast no identifica en su output la función moonlighting si lo hace el programa HMMER basado en perfiles de modelos de Markov. Por ejemplo, en la Tabla 4 para la proteína *I-TevI endonuclease*, o la *Thyimidine phosphorylase* la función moonlighting (*transcriptional regulator*) es predicha por HMMER ya en la primera iteración. En el presente trabajo no se han analizado por HMMER el conjunto de proteínas de la base de datos, cosa que será realizada en el futuro. En parte es porque el servidor (hmmer.janelia.org) desde hace meses presenta muchos problemas y muestra la siguiente información: *Search Failed. We're sorry, it looks like something went wrong with our search system. It may be a transient error, so please feel free to try the search again. Alternatively, please contact us.*

Es posible que la triple combinación de PsiBlast, HMMER y bases de datos de interactómica incremente la sensibilidad y especificidad de la predicción bioinformática de las proteínas multifuncionales.

IV.B.2. BÚSQUEDAS EN BASES DE DATOS DE INTERACTÓMICA

En trabajos previos propusimos que combinar las búsquedas en bases de datos de interacción proteína-proteína (PPI) con el análisis de similitud de secuencias pueden ayudar a predecir la función, canónica, de las proteínas (Espadaler et al, 2008) y que las bases de datos de PPI deben contener información sobre las proteínas moonlighting y proporcionar sugerencias para un posterior análisis experimental con el fin de demostrar sus propiedades multifuncionales, Figura 17**,** (Gómez et al., 2011). Se ha descrito abundantemente que la interactómica puede ayudar a predecir la función de una proteína por sus partners, de acuerdo con la consideración de "culpable por asociación" (*guilty-by-association*). Por ello nosotros propusimos que las proteínas asociadas a otra proteína por interactómica podrían sugerir también la función moonlighting de una proteína al menos a nivel de GO: *Biological Process* (Gómez et al., 2011). Hay que resaltar que, en contraste con la anarquía existente en las anotaciones presentes en las bases de datos de secuencias, y por ser de diseño más reciente, en las base de datos de interactómica (PPI) predomina la anotación GO. En el presente trabajo hemos considerado que las bases de datos de interactómica revelan correctamente una segunda función para la proteína moonlight si la base de datos PPI identifica una función molecular (GO: *Molecular Function*) o, en algunos casos, un proceso

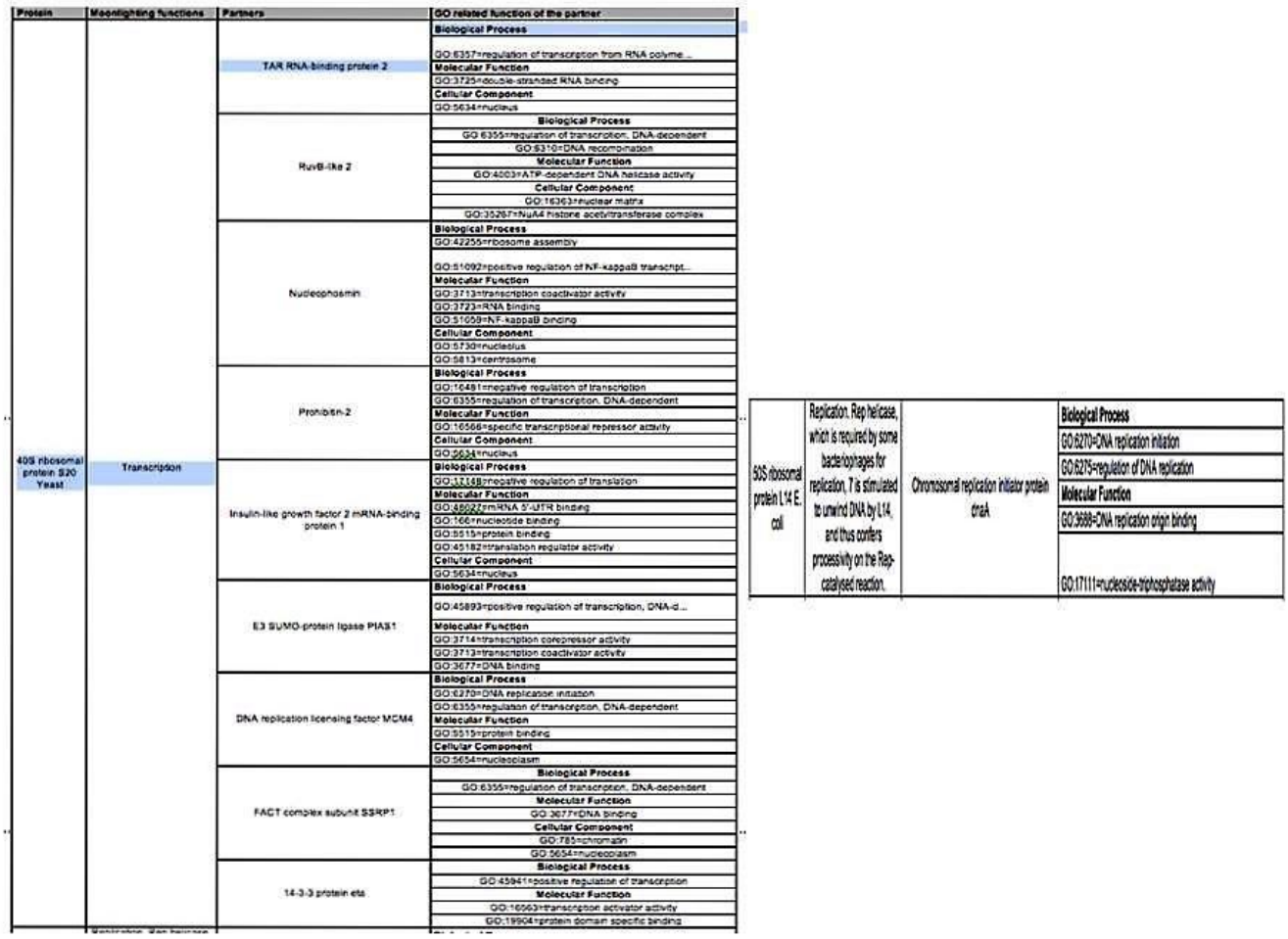
biológico (GO: *Biological Process*) que esté de acuerdo con la función moonlighting que se describe en nuestra base de datos MultiutaskProtDB. Luego, con el fin de filtrar los aciertos y mejorar la precisión, es recomendable realizar un análisis de enriquecimiento de ontología de genes utilizando GOstat, como se ha descrito en el Apartado III.C. de Métodos.

Los términos GO asociados a la proteína diana son claramente significativos e identifican la función moonligting

Protein	Known moonlighting functions	Database interacting partners	GO related functions	GO enrichment *P*-value
Aconitase	mtDNA maintenance	ATP-dependent DNA helicase MER3	GO:0017111: nucleoside-triphosphatase activity	0.00461
			GO:0030554: adenyl nucleotide binding	0.00648
			GO:0001883: purine nucleoside binding	0.00664
			GO:0001882: nucleoside binding	0.00685
			GO:0008135: translation factor activity, nucleic acid binding	0.00017
Aldolase	Vacuolar H+-ATPase assembly	V-type proton ATPase subunit E 1	GO:0008553: hydrogen-exporting ATPase activity	0.00361
			GO:0042623: ATPase activity, coupled	0.00615
			GO:0051117: ATPase binding	0.00677
			GO:0046961: proton-transporting ATPase activity, rotational	0.00857
			GO:0016887: ATPase activity	0.00857
Enolase	Bind to cytoskeletal structures	Actin	GO:0034621: cellular macromolecular complex organization	7.54×10^{-5}
			GO:0032506: cytokinetic process	0.0053
			GO:0007109: cytokinesis, completion of separation	0.0021
		Microtubule-associated protein 4	GO:0007017: microtubule-based process	0.00286
			GO:0051488: activation of anaphase-promoting complex	0.00314
			GO:0000920: cytokinetic cell separation	0.00418
Glyceraldehyde-3-phosphate dehydrogenase	Microtubule bundling	Tubulin polymerization-promoting protein	GO:0051015: actin filament binding	0.0071
			GO:0001948: beta-catenin binding	0.00594
			GO:0008017: microtubule binding	0.00251
	Phosphate group transfer	Phosphoglycerate kinase 1	GO:0017111: nucleoside-triphosphatase activity	0.00222
			GO:0016462: pyrophosphatase activity	0.00316
			GO:0016772: transferase activity, transferring phosphorus-containing groups	0.00104

Figura 17. Algunos ejemplos de identificación de proteínas moonighting a partir de la información de bases de datos de interactómica (en este caso utilizando el servidor APID (http://bioinfow.dep.usal.es/apid/index.htm)

El problema con las bases de datos de PPIs, aparte del hecho de que en muchas especies no se han realizado experimentos de interactómica, es la heterogeneidad de casos. Por ejemplo la Figura 18 muestra dos ejemplos, en que identifica la función moonlight, pero que en uno (proteína ribosomal S20) hay numerosas posibilidades y a priori no podemos saber cuál o cuáles son verdaderos positivos, y otro (proteína ribosomal 50) en que identifica directamente la función moonlight. Por ello proponemos combinar los datos de PPIs con el análisis de homología remota (ver siguiente apartado y la Tabla 3).



Protein	Moonlighting functions	Partners	GO related function of the partner
40S ribosomal protein S20 Yeast	Transcripson	TAR RNA-binding protein 2	**Biological Process**
			GO:6357=regulation of transcription from RNA polyme...
			Molecular Function
			GO:3725=double-stranded RNA binding
			Cellular Component
			GO:5634=nucleus
		RuvB-like 2	**Biological Process**
			GO:6355=regulation of transcription, DNA-dependent
			GO:6310=DNA recombination
			Molecular Function
			GO:4003=ATP-dependent DNA helicase activity
			Cellular Component
			GO:16363=nuclear matrix
			GO:35267=NuA4 histone acetyltransferase complex
		Nucleophosmin	**Biological Process**
			GO:42255=ribosome assembly
			GO:51092=positive regulation of NF-kappaB transcript...
			Molecular Function
			GO:3713=transcription coactivator activity
			GO:3723=RNA binding
			GO:51059=NF-kappaB binding
			Cellular Component
			GO:5730=nucleolus
			GO:5813=centrosome
		Prohibitin-2	**Biological Process**
			GO:10481=negative regulation of transcription
			GO:6355=regulation of transcription, DNA-dependent
			Molecular Function
			GO:10566=specific transcriptional repressor activity
			Cellular Component
			GO:5634=nucleus
		Insulin-like growth factor 2 mRNA-binding protein 1	**Biological Process**
			GO:17148=negative regulation of translation
			Molecular Function
			GO:48027=mRNA 5'-UTR binding
			GO:166=nucleoside binding
			GO:5515=protein binding
			GO:45182=translation regulator activity
			Cellular Component
			GO:5634=nucleus
		E3 SUMO-protein ligase PIAS1	**Biological Process**
			GO:45893=positive regulation of transcription, DNA-d...
			Molecular Function
			GO:3714=transcription corepressor activity
			GO:3713=transcription coactivator activity
			GO:3677=DNA binding
		DNA replication licensing factor MCM4	**Biological Process**
			GO:6270=DNA replication initiation
			GO:6355=regulation of transcription, DNA-dependent
			Molecular Function
			GO:5515=protein binding
			Cellular Component
			GO:5654=nucleoplasm
		FACT complex subunit SSRP1	**Biological Process**
			GO:6355=regulation of transcription, DNA-dependent
			Molecular Function
			GO:3677=DNA binding
			Cellular Component
			GO:785=chromatin
			GO:5654=nucleoplasm
		14-3-3 protein eta	**Biological Process**
			GO:45941=positive regulation of transcription
			Molecular Function
			GO:16563=transcription activator activity
			GO:19904=protein domain specific binding

50S ribosomal protein L14 E. coli	Replication. Rep helicase, which is required by some bacteriophages for replication, T is stimulated to unwind DNA by L14, and thus confers processivity on the Rep-catalysed reaction.	Chromosomal replication initiator protein dnaA	**Biological Process**
			GO:6270=DNA replication initiation
			GO:6275=regulation of DNA replication
			Molecular Function
			GO:3688=DNA replication origin binding
			GO:17111=nucleoside-triphosphatase activity

Figura 18. Algunos ejemplos de algunas proteínas moonlighting y sus partners de interactómica presentes en las bases de PPIs. Aparte del problema de que en muchos casos no hay experimentación en interactómica para la proteína o la especie de interés, la información puede ser de muy diferente calidad, con muchos falsos positivos y falsos negativos. Por otra parte hay proteínas con numerosos partners (proteína ribosomal 40s) o con muy pocos (proteína ribosomal 50s).

La Columna 3 de la Tabla 3 muestra algunos ejemplos de identificación de funciones moonlighting a partir de las bases de datos PPIs. Debido a que el número de proteínas de las PPI asociadas por interacción a la nuestra (la problema) puede ser alta, escoger los Verdaderos Positivos no es una tarea fácil si el investigador no tiene pistas adicionales. La lista de aciertos tiene que reducirse adecuadamente al tomar en cuenta las otras predicciones bioinformáticas como se describe a continuación o con la ayuda de los datos experimentales o clínicos que sugieran correlaciones interesantes. Hemos encontrado que

mediante la combinación de información de bases de datos de PPI y búsquedas de homología remotas, la predicción moonlighting mejora positivamente, en el sentido de que incrementa la especificidad de la predicción (menor número de Falsos Positivos), aunque sea a costa de una cierta perdida de sensibilidad (mayor número de Falsos Negativos). Un problema adicional es que muchas especies no han sido analizadas por interactómica, por lo tanto, una serie de proteínas de la base de datos MultitaskProtDB no tiene proteínas asociadas interaccionando en las bases de datos PPI (249 proteínas de MultitaskProtDB corresponden a las especies con interactómica experimental descrita).

En nuestra opinión, el principal límite del nivel de predicción de la multifuncionalidad de las proteínas a partir de las bases de datos PPIs se debe principalmente a la baja sensibilidad de los métodos de interactómica (es decir, dan lugar a muchos Falsos Negativos, especialmente el método del doble híbrido) en lugar de una baja especificidad (es decir, Falsos Positivos).

IV.B.3. RESULTADO DE COMBINAR LA BÚSQUEDA EN BASES DE DATOS DE INTERACTÓMICA CON EL ANÁLISIS DE HOMOLOGÍA PSI-BLAST/BYPASS

Se realizaron búsquedas de aquellas proteínas de la base de datos de proteínas multifuncionales MultitaskProtDB de las que se tenían datos de interactómica en el servidor APID. Como se indicó anteriormente, cada proteína moonlighting puede presentar un gran número de partners de interactómica y, a su vez, una gran lista de resultados de homólogos remotos a partir del algoritmo de PSI-BLAST. Hemos inspeccionado manualmente ambos tipos de resultados de salida para comprobar si la intersección de los dos conjuntos reduce la lista de aciertos de candidatos y mejora la predicción de proteínas multifuncionales de la base de datos. Esta inspección manual es necesaria porque, como ya se ha descrito anteriormente, existe un problema relacionado con los diferentes descriptores de anotación representados por los dos tipos de salida. La mayor parte de salida de resultados Blast/PSI-Blast no corresponden a anotaciones semánticas sino que son ambiguas e incluso anárquicas, mientras que muchas bases de datos de PPI, por ser recientes, ya utilizan anotaciones GO. Este hecho complica la comparación automática de los resultados de salida. En nuestro caso hemos considerado como resultados positivos aquellos en los que se identifica una función, canónica y/o moonlighting, en cualquier posición del output del Psi-Blast/ByPass y que se corresponde con un partner de la base de datos PPI, como se

muestra en los ejemplos de la Tabla 3, columnas 3 y 4 (y S4 en Material suplementario). Actualmente estamos diseñando un programa que sea capaz de encajar e identificar automáticamente dos o más salidas de resultados. Por otra parte, también se ha mencionado que el conjunto de anotaciones GO para *Molecular function* contiene tan sólo alrededor de 9500 funciones moleculares, por lo que muchas de las anotaciones funcionales de las bases de datos de secuencias no se pueden encontrar utilizando GO. De todos modos, el análisis de homología con anotación GO, por ejemplo el Blast2GO (Gotz et al., 2008), facilitarán el análisis ya que no se tendrán en cuenta las anotaciones ambiguas y los descriptores de baja calidad presentes en las bases de datos de secuencias actuales.

En nuestra opinión, la combinación de los resultados de los *outputs* de homólogos remotos a partir de PSI-BLAST con los partners de interacción contenidos en las PPIs es el mejor enfoque para reducir los en general largos outputs de ambos servidores y mejorar la predicción bioinformática de posibles proteínas moonlighting. Como se ha mencionado anteriormente este solapamiento tan sólo representa la tercera parte de las 249 proteínas analizadas, pero lo ha hecho con un alto nivel de especificidad (Tablas 5a,b). En algunos casos tan sólo se identifica el GO: *Biological Process* pero esto puede sugerir pistas para revelar la función moonlighting experimentalmente o computacionalmente. Estas Tablas muestran los resultados obtenidos y esperados de las predicciones simples (la función moonlight es identificada tan solo por homología remota o por interactómica) y dobles (la función moonlight es identificada tanto por homología remota como por interactómica). Es sobre las 249 proteínas de MultitaskProtDB que de las que existen resultados de interactómica, por falta de experimentación en algunas especies o por escasez de datos para algunas clases de proteínas (membrana, pequeñas…), etc.

Tabla 5a. Frecuencias observadas (sobre 249 proteínas moonlight de base de datos MultitaskDB)

	PsiBlast SI identifica	**PsiBlast NO identifica**	**Suma**	
PPIs SI identifica	66	69	135	0.54216867
PPIs NO identifica	35	79	114	0.45783133
Suma	101	148	249	
	0.40562249	0.59437751		

Tabla 5b. Frecuencias esperadas (sobre 249 proteínas moonlight de base de datos MultitaskDB)

	PsiBlast SI identifica	PsiBlast NO identifica	Suma
PPIs SI identifica	54.759	80.240	135
PPIs NO identifica	46.240	67.759	114
Suma	101	148	249

p = 0.0035912

Como puede observarse en la tabla, aproximadamente el 40% de las proteínas moonlighting de la base de datos MutitaskProtDB tan sólo son identificadas por el PsiBlast/ByPass y el 54% tan solo por interactómica. Por ambos a la vez el 27%. Un análisis mediante un test de contingencia usando el test de X2 muestra que PPI y PsiBlast no están identificando el mismo subconjunto de proteína moonlighting ya que podemos observar una p de aproximadamente 0.0036. Esto nos indicaría que la forma en la que ambos algoritmos funcionan a la hora de encontrar la función moonlighting sigue vías muy diferentes, pues el subgrupo de funciones encontradas son significativamente diferentes.

IV.B.4. BÚSQUEDAS DE PATRONES DE SECUENCIA DE PROTEÍNAS (MOTIFS/DOMINIOS) ESPECÍFICOS DE FUNCIÓN

La identificación de diferentes motivos/dominios (motifs/domains) de secuencia de proteínas vinculados a función usando InterPro u otros algoritmos debería, en principio, ayudar a identificar las proteínas moonlighting (Figura 19). Sin embargo, hay dos problemas principales: (a) el número relativamente bajo de motivos, dominios y firmas conocidas en la actualidad (1300 Prosite y 1008 ProDom, solapando un cierto número de ambos) y (b), la versión actual de programas como Prosite, etc, han sido diseñados para una predicción más precisa de los motivos/dominios principales y más comunes (en general asociados a la función canónica) pero no para identificar patrones secundarios y menos específicos (en general asociados a la función moonlighting). Esto explicaría el hecho de que el uso de InterPro sobre las proteínas de la base de datos MultitaskProtDB revela la función canónica de alrededor el 80% de ellos, pero la función moonlighting en sólo el 8% de los casos. Por ejemplo, una proteína muy multifuncional es la *glyceraldehyde-3-phosphate dehydrogenase*, en que tanto el análisis por PSI-BLAST como de partners de interactómica sí que identifican varias funciones con buena puntuación (“score”) y anotación GO (Figura 20). Sin embargo la actual versión de InterPro sólo identifica un motivo para la función canónica de esta proteína

(Figura 21). En cambio el programa Blocks identifica más funciones moonlighting, por ejemplo en el caso de la proteína Arg 2, Blocks identifica las funciones canónicas y moonlighting como las dos mejores puntuaciones del output (Figura 22). El hecho de que un programa como Blocks, que no ha sido actualizado desde el año 2006, sea mejor para la detección del patrón de una función secundaria, moonlighting, que programas más recientes como InterPro, nos hizo pensar que este fenómeno puede ser debido a un problema de la relación entre sensibilidad y especificidad. Las herramientas de detección de patrones se han desarrollado tradicionalmente para tener una buena relación entre la especificidad y sensibilidad. Cuando se diseña un conjunto de datos que representen un buen gold-standard para entrenar una herramienta como InterPro, generalmente se asume que todas las proteínas incluidas en la base de datos sólo tienen una función única. Por lo tanto, si esta suposición no es cierta, como es el caso de las proteínas multifuncionales, el programa comienza sesgado en términos de pérdida de la sensibilidad, de modo que las herramientas tienden a detectar un bajo número de funciones secundarias. En este sentido, la tendencia de la utilización de muchas secuencias escogidas ("curadas") para construir estos patrones podría explicar por qué las herramientas obsoletas como Blocks son más eficaces en la detección de las funciones secundarias, multifuncionales. Si es así, esto indicaría que para detectar este tipo de funciones secundarias, herramientas como Blast o PSI-Blast puede ser más apropiada, ya que no dependen de la pre-existencia de patrones previamente construidos con un conjunto de secuencias funcionalmente sesgadas. También pueden existir otros factores adicionales, tales como el hecho de que muchas herramientas nuevas, además de un conjunto de proteínas con función conocida, incorporan un conjunto de falsos positivos (la secuencia comparte el motivo, pero no tiene una función asociada). O sea, este conjunto de falsos positivos contienen proteínas que llevan el patrón asociado con la función de la proteína, pero que en realidad no realizan esta función. Para comprobar si algunos de los falsos positivos son en realidad erróneamente descartados para las funciones moonlighting hemos comparado todo el conjunto de secuencias de falsos positivos en la base de datos Prosite con nuestra base de datos de proteínas multifuncionales (Tablas S5 y S6 en Material suplementario). Luego se comprobó si los patrones correspondientes a las secuencias de falsos positivos mostraban un alto grado de homología secuencial con nuestras proteínas multifuncionales y si tenían una similitud con la función moonlighting de estas proteínas. Este cálculo nos llevó a la conclusión de que, al menos para Prosite, los

falsos positivos son verdaderos falsos positivos, porque ninguna de esas funciones coincide con la función moonlighting de la proteína.

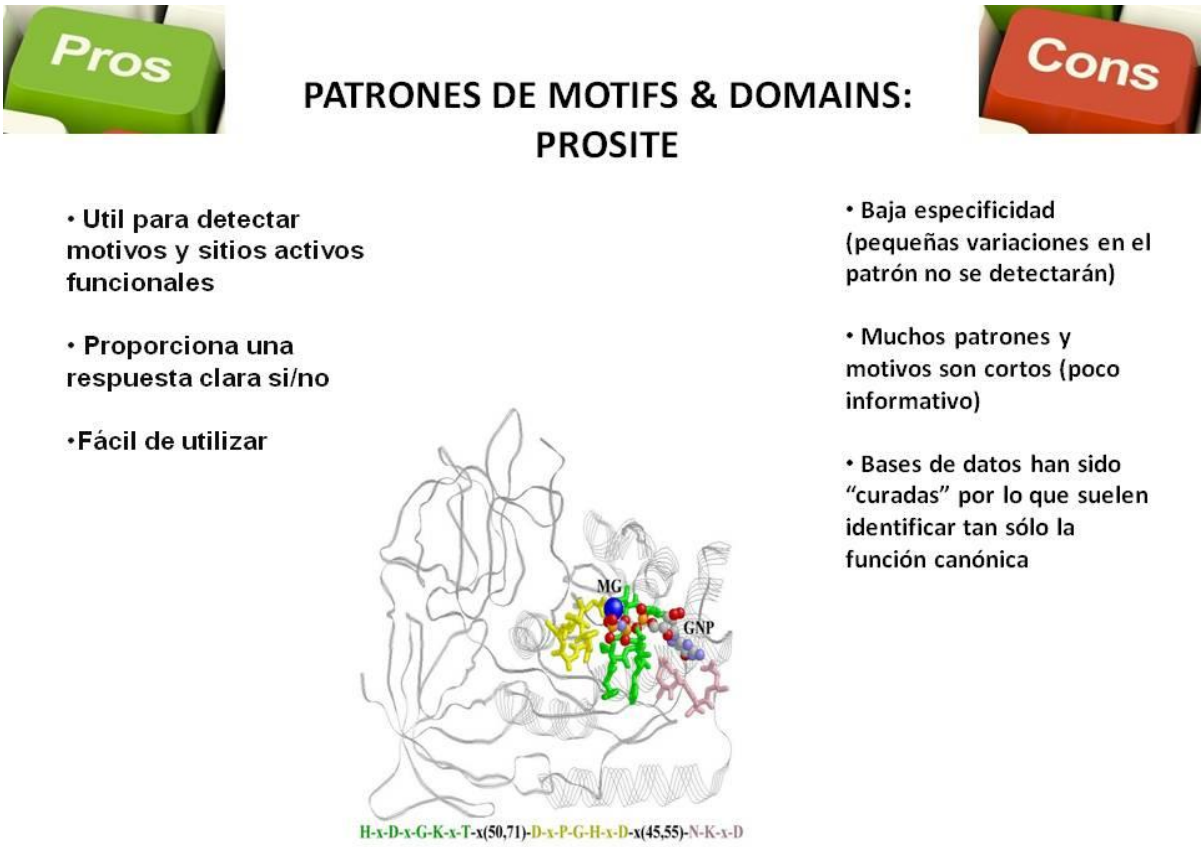


Figura 19. Pros y contras del primer programa existente de identificación de motifs funcionales, el Prosite (actualmente puede ejecutarse con varios más, Pfam, ProDom, etc, en la base de datos de InterPro (**www.ebi.ac.uk/interpro/Tools/pfa/iprscan/**)

		Microtubule-associated protein 4	GO:0007017: microtubule-based process	0.00286
			GO:0051488: activation of anaphase-promoting complex	0.00314
			GO:0000920: cytokinetic cell separation	0.00418
Glyceraldehyde-3-phosphate dehydrogenase	Microtubule bundling	Tubulin polymerization-promoting protein	GO:0051015: actin filament binding	0.0071
			GO:0001948: beta-catenin binding	0.00594
			GO:0008017: microtubule binding	0.00251
	Phosphate group transfer	Phosphoglycerate kinase 1	GO:0017111: nucleoside-triphosphatase activity	0.00222
			GO:0016462: pyrophosphatase activity	0.00316
			GO:0016772: transferase activity, transferring phosphorus-containing groups	0.00104
	Binds to RNA, RNA polymerase	Heterogeneous nuclear ribonucleoprotein Q	GO:0003727: single-stranded RNA binding	0.00788
			GO:0008266: poly(U) RNA binding	0.00094
			GO:0003723: RNA binding	5.11×10^{-1}
	Decrease blood insulin levels	Growth factor receptor-bound protein 2	GO:0043567: regulation of insulin-like growth factor receptor signaling pathway	0.00593
	Nuclear tRNA export	Ataxin-1	GO:0050658: RNA transport	0.001658
			GO:0050657: nucleic acid transport	0.001658
			GO:0051236: establishment of RNA localization	0.001658
	Significant role in apoptosis	TNF receptor-associated factor 1	GO:0042981: regulation of apoptosis	4.03×10^{-6}
			GO:0006915: apoptosis	1.93×10^{-5}
			GO:0043065: positive regulation of apoptosis	0.00053
			GO:0006917: induction of apoptosis	0.00199

Figura 20. Una de las proteínas más mulifuncionales conocidas es la *glyceraldehyde-3-phosphate dehydrogenase* (GAPDH). Las bases de datos de interactómica identifican varias de sus funciones moonlighting

A partir de los programas individuales ejecutados en conjunto o por separado por InterPro, hemos encontrado y publicado previamente (Gómez et al., 2003) que ProDom tiene el mejor rendimiento para identificar los dominios tanto canónico como moonlighting (Figura 7 de la Introducción). Esto es probablemente debido al hecho de que se trata de una base de datos construida a partir de perfiles secuenciales, que resultan ser más flexible que los algoritmos de búsqueda de patrones específicos. Se han generado por un procedimiento automático que conserva una importante fuente de variabilidad y, además, tiene una mayor representación de familias de proteínas. La Figura 23 muestra una predicción ProDom de los dos dominios relacionados con ambas funciones, canónica y moonlighting, de una proteína muy multifuncional en diversas especies, la aconitasa. Los programas de búsqueda de perfiles (es decir, Blocks, ProDom) proporcionan una buena puntuación, y por otra parte, el patrón no se limita tan sólo al evolutivamente conservado sitio activo, sino que se extiende

incluso lejos de las regiones conservadas. Sin embargo, estos programas como Blocks y ProDom presentan un mal tratamiento de los huecos (*gaps*). Los programas de búsqueda de patrones, como Prosite, son más útiles para la detección de sitios activos funcionales pero presentan baja especificidad, por lo tanto no se detectarán pequeñas variaciones de un patrón. El potencial de identificar las funciones moonlighting por InterPro no supera el 10%, incluso sumando todas las aplicaciones incluidas en él y teniendo en cuenta las anotaciones correctas derivadas de los materiales suplementarios presentes en los descriptores de los patrones. Cuando la posible función anotada corresponde a una nueva función, la probabilidad de fallar en la predicción de aplicar este método es muy alta, ya que ir desde el nivel molecular hasta los niveles superiores (celulares, organismo, etc.) es arriesgado.

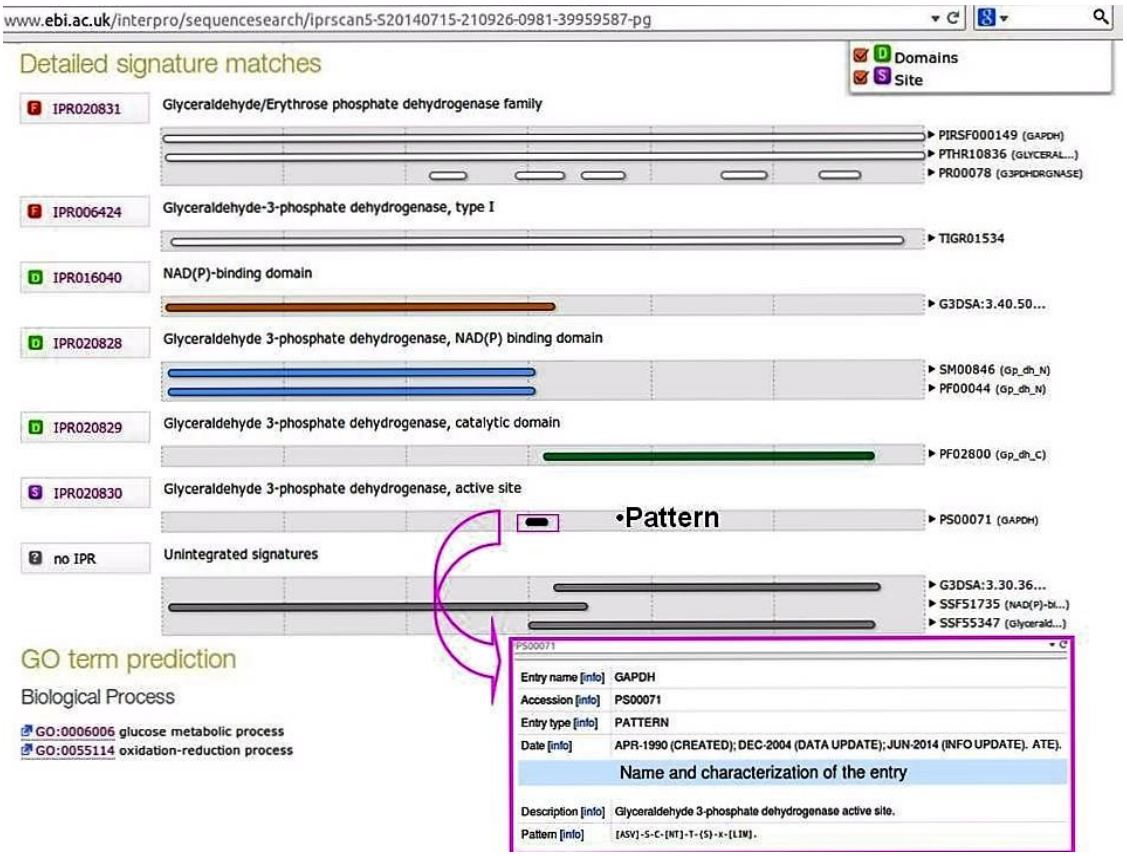


Figura 21. Una de las proteínas más mulifuncionales conocidas es la *glyceraldehyde-3-phosphate dehydrogenase* (GAPDH). El programa InterPro, al menos en su versión actual, tan sólo identifica la función GAPDH. El que muchos programas de identificación de motifs, como la versión actual de InterPro, hayan sido espurgados de los motifs de menor *score* hace que, en muchos casos, se pierda la posibilidad de identificar dos o más funciones

Pfam es uno de los programas que ejecuta InterPro. Pfam está basado en el patrón de búsqueda mediante Hidden Markov Models (HMM) Figura 24). Las familias incluidas en Pfam (dominios de proteínas agrupados utilizando los Modelos Ocultos de Markov) se construyen a partir de múltiples alineamientos secuenciales, en muchos casos divididos en dominios Pfam separados. La actividad biológica de estas familias podría ser descrita como múltiples dominios que realizan juntos la función principal. A priori, estas características podrían implicar que los dominios Pfam sería una mejor herramienta para identificar las funciones moonlighting. Nuestros resultados, a partir del análisis Pfam (por ejemplo vía InterPro) aplicado a nuestra base de datos muestran que los dominios Pfam son cuatro veces más eficaces en la detección de la función moonlighting que otros métodos de motivos y de dominios, pero la significación estadística de esta diferencia es baja, el valor p proporcionado por una prueba de X^2 es 0,02.

Block Searcher Results

Go to hits

Hits

Query=Unknown Unknown Size=355 Amino Acids Blocks Searched=27288
Alignments Done= 10398239 Cutoff combined expected value for
hits= 1 Cutoff block expected value for repeats/other= 1

Combined Family			Strand	Blocks	E-value
IPB005522	Inositol polyphosphate kinase	1	5 of 5		7.4e-70
IPB005612	CBF/Mak21 family	1	1 of 11		4.6e-08
IPB007759	DNA-directed RNA polymerase delta s	1	1 of 2		2.6e-07
IPB004855	Transcription factor IIA, alpha/bet	1	1 of 4		2.7e-06

Figura 22. Programas que han sido menos espurgados de los motifs de menor *score* como Blocks (que no ha sido actualizado desde el año 2006) permiten en muchas ocasiones identificar mejor las dos funciones de una proteína moonlighting. La figura presenta un ejemplo, la proteína Arg82, en que Blocks identifica las dos funciones en las dos posiciones superiores por su puntuación

Otra consideración importante es que parte de la mejora en la predicción de la función moonlighting por Pfam es debido a la información adicional de la función de un dominio dado como documentación complementaria (Figura 25).

>PD349383 (ProDom release)
Number of domains in family: 63
Commentary (automatic):
SUBNAME: BIOSYNTHESIS FULL=HOMOACONITASE MITOCHONDRIAL FULL=HOMOACONITASE
LYASE EC=4.2.1.36 HYDRATASE FLAGS: IRON
Length = 65
Score = 208 (84.7 bits), Expect = 2e-16
Identities = 36/47 (76%), Positives = 39/47 (82%)

```
Query:    18 LKGQNLTEKIVQSYAVNLPEGKVVHSGDYVSIKPAHCMSHDNSWPVA 64
             L+GQ LTEKIVQ YAV LP GK V SGDYV+I P HCM+HDNSWPVA
Sbjct:    19 LRGQTLTEKIVQRYAVGLPPGKYVRSGDYVTISPHHCMTHDNSWPVA 65
```

>PDB1H055 (ProDom release)
Number of domains in family: 4
Commentary (automatic):
SUBNAME: LYASE METAL-BINDING IRON RECNAME:
Length = 57
Score = 179 (73.6 bits), Expect = 6e-13
Identities = 34/53 (64%), Positives = 44/53 (83%), Gaps = 33/53 (62%)

```
Query:   576 GSSREQAATALLAKGINLVVSGSFGNIFSRNSINNALLTLEIPALIKKLREKY 628
             GSSREQAAT++LAK + LVV GS GN FSRN++NNAL  LE+P L+++LRE +
Sbjct:     2 GSSREQAATSILAKQLPLVVCGSIGNTFSRNAVNNALPLLEMPRLVERLREAF 54
```

Figura 23. Ejemplo de identificación de los dos dominios funcionales de la aconitasa por el programa ProDom. Este programa es un buen predictor de multifuncionalidad en los casos en que la proteína presenta cada función en dominios bien definidos y presentes en la base de datos de ProDom

Otro de los puntos que hemos explorado es la diferencia entre las bases de datos PfamA y PfamB. PfamA es una base de datos manualmente curada que contiene un conjunto de Modelos Ocultos de Markov de más de 14.000 familias. La base de datos PfamB se construye de forma automática con grupos de secuencias producidas por el algoritmo ADDA (Heger et al., 2005), y sus familias suelen provenir de alineamientos que contienen proteínas con funciones bastante heterogéneas. Esta característica nos animó a probar si PfamB era una herramienta adecuada para predecir funciones moonlighting. Hemos probado las dos

versiones utilizando el conjunto de proteínas de la base de datos MultitaskProtDB. PfamA predice el 78% de las funciones canónicas, pero sólo el 6% de las funciones moonlighting. Con PfamB, encontramos 58 proteínas a partir del conjunto de proteínas de la base de datos MultitaskProtDB que tienen alta homología con al menos una familia PfamB, y el programa caracterizó adecuadamente el 60% de las funciones canónicas y el 14% de las funciones de moonlighting. Sin embargo, este método es difícil de automatizar, ya que el número de anotaciones a testar es muy alta, incluso seleccionando previamente los mejores ejemplos. De esta manera, hemos hecho una pequeña lista de anotaciones en cada familia PfamB, dando prioridad a las secuencias de cadena más larga con respecto a las de cadena corta incluidas en el grupo original de secuencias semilla utilizadas para generar familias PfamB. También hay que resaltar que alrededor del 80% de las proteínas identificadas por PfamB como proteínas multifuncionales no fueron identificadas por PfamA, lo que nos habla que realmente están explotando conjuntos distintos de datos para inferir la función. Obviamente, si tenemos alguna ligera idea de la función de las proteínas, la exploración de los resultados de salida de Pfam A o B puede proporcionar sugerencias sobre el proceso para encontrar la función moonlighting de nuestra proteína problema. En S7 del Material suplementario hay la información Pfam recopilada para las proteínas de MultitakProtDB.

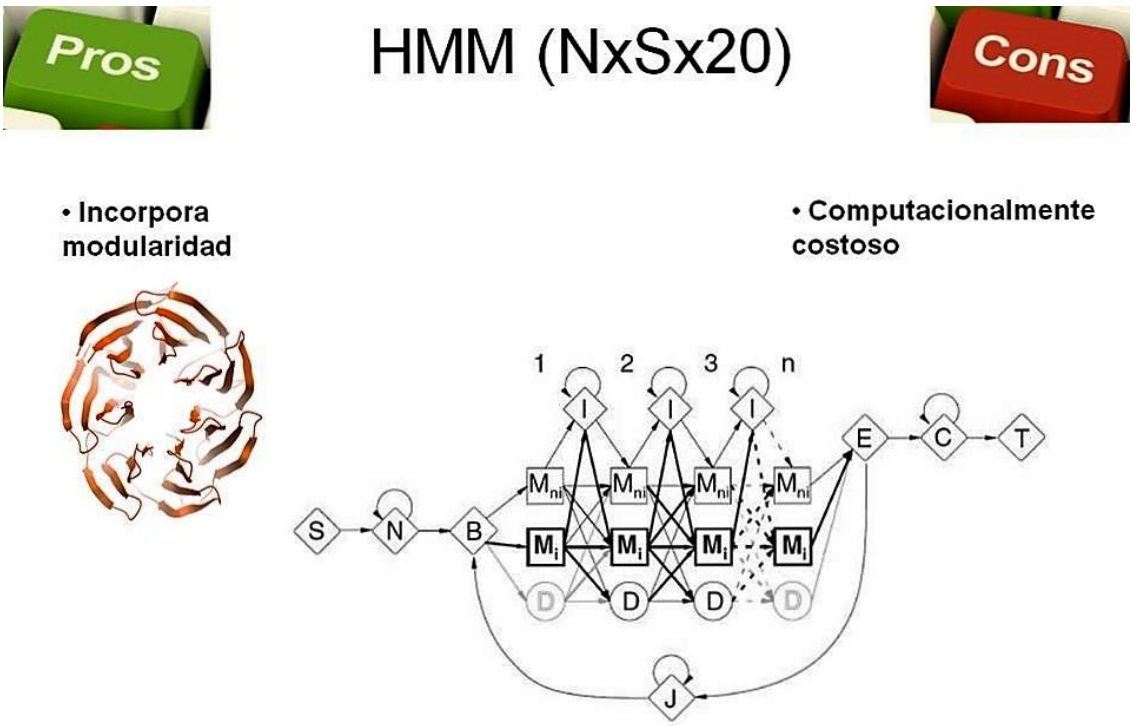


Figura 24. Pros y contras del programa de identificación de familias funcionales Pfam basado en cadenas de Markov

En la Figura 7 de la Introducción, correspondiente al primer trabajo sobre proteínas moonlighting del grupo, se muestra una serie de ejemplos en que a partir de programas de identificación de motifs y dominios en algunos casos se predicen las funciones moonlighting, aunque en aquel momento el número de proteínas moonlighting conocidas era muy pequeño. En la Figura 39 (en la Discusión General del presente trabajo) se muestran y discuten detalles adicionales sobre Pfam y otros algoritmos, programas y bases de datos.

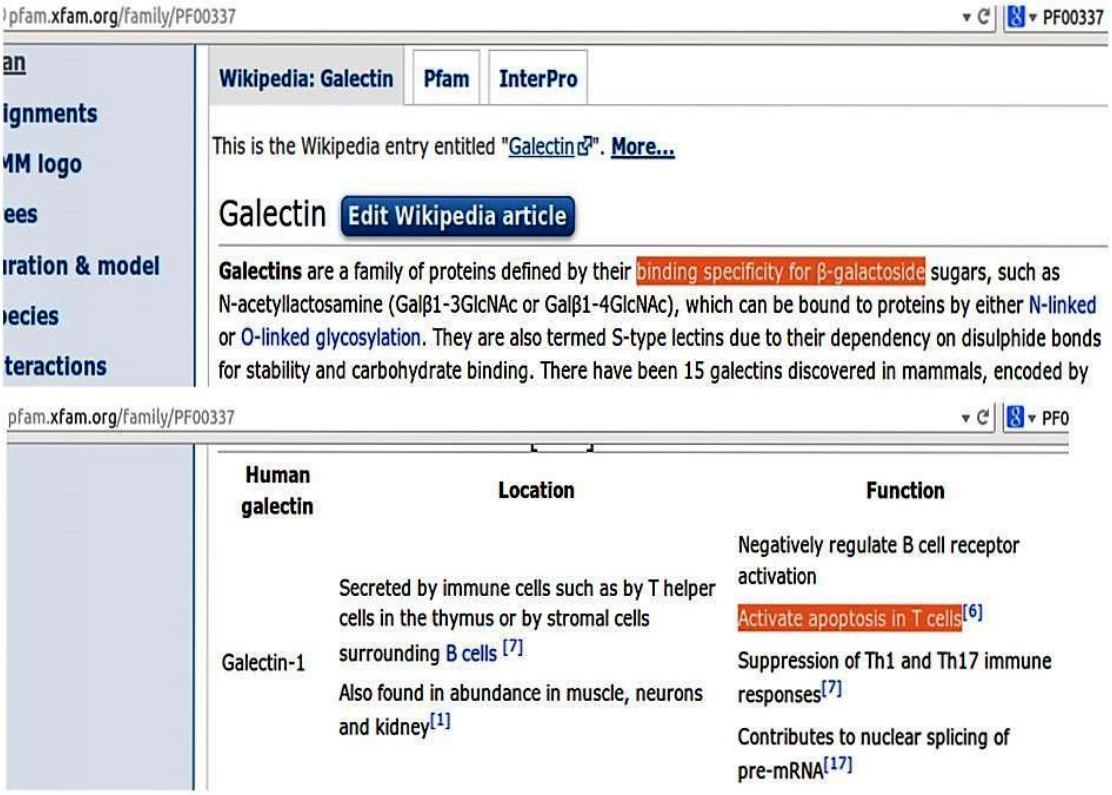


Figura 25. El programa Pfam presenta información suplementaria que puede facilitar la identificación de una segunda función moonlighting, como es el caso de la galactin

En cuanto al análisis mediante el servidor de motifs de secuencia de aminoácidos eucariotas ELM (http://elm.eu.org) (Dinkel et al., 2014) relacionados con modificación post-traduccional, función y localización celular de proteínas no nos ha representado un incremento en la predicción de la función moonlighting respecto a InterPro, Blocks, etc. Tampoco el servidor Minimotif Miner (http://minimotifminer.org) (Mi et al., 2012) que predice cortos motifs relacionados con interacción y modificación post-traduccional. En ambos casos da lugar a un enorme listado de dianas, lo que no facilita identificar las que pueden estar relacionadas con la función moonlighting. Hay filtros para reducir el número de falsos positivos. En todo caso el gran número de motifs cortos de interacción que muestran los servidores ELM y Minimotif Miner cuando se pasa una proteina indica la gran facilidad para interaccionar que ofrece la superficie de una proteína lo que facilitaría la adquisición de nuevas funciones. De hecho Keskin y Nussinov, 2007, ya mostraron que incluso sitios de unión similares permiten la interacción con diferentes Partners.

Ya se ha mencionado en la Introducción una interesante paradoja y es que diferentes proteínas del metabolismo primario tengan igual función moonlight en relación con la virulencia. Por ejemplo la enolasa de más de 30 diferentes especies se une al plasminógeno del huésped, pero en este caso son todas enolasas con una alta homología de secuencia. Pero a su vez la *glyceraldehyde-3-phosphate dehydrogenase* de 18 especies también une plasminógeno. O la *phosphoglycerate mutase* de 4 microorganismos. Y la *triosephosphate isomerase* de 3 microrgansmos, etc. Hay más ejemplos en nuestra base de datos. Esto implica que estas enzimas que no presentan similitud de secuencia comparten alguna característica conformacional, o algún motif no identificado. Tanto las enolasas, GAPDH, etc son proteínas que pertenecen a complejos (glicolisis, Krebs…) por lo que están adaptadas a facilitar la interacción con otras. También ocurre con las proteínas ribosomales y también son muy propensas a la multifuncionalidad. Esto facilitaría la fácil adaptación a una segunda función, que obviamente pasa por una nueva interacción con otras proteínas y biomoléculas.

IV.B.5. ANÁLISIS DE CORRELACIÓN DE MUTACIONES

Los estudios de redes de correlación de mutaciones o de evolución conjunta de los aminoácidos catalíticos se han usado para predecir los residuos catalíticos clave de las enzimas. Cualquiera que haya trabajado en ingeniería de proteínas sabe que pequeños cambios en la secuencia de las proteínas a veces pueden tener resultados catastróficos. El hecho de que tan sólo se suelen publicar los mutantes que han prosperado lleva a la falsa impresión de que rediseñar u obtener otras funciones por mutaciones es un proceso sencillo. Pero suele ocurrir que otros aminoácidos, a veces alejados del objeto de estudio, también contribuyen al proceso de plegamiento y a la conformación funcional final. Los programas de análisis de correlación de mutación, como Mistic, pueden ayudar a identificar estos aminoácidos.

Hemos comprobado si el algoritmo Mistic puede ayudar a predecir las proteínas moonlighting. La principal limitación de algoritmos como MISTIC es que requieren un gran número de secuencias alineadas mediante multi-alineamiento (Figura 26). Sin embargo el número de de proteínas moonlighting perteneciente a la misma familia es escaso, siendo excepciones las enolasas y aldolasas. En el presente trabajo se analizó la matriz de correlación de los aminoácidos de las enolasas que tienen la función adicional de unirse a

plasminógeno, creando la matriz de correlación de todas las enolasas contenidas en nuestra base de datos (Figura 27). Para ello se alinearon un conjunto de secuencias enolasa con menos del 35% de identidad de secuencia aminoacídica. Al mismo tiempo, hemos comparado el mismo conjunto de enolasas pero eliminando todas las que se unen a plasminógeno. Se utilizó la primera entrada en este alineamiento múltiple como una referencia de secuencia con el fin de facilitar la comparación entre los dos alineamientos múltiples (Figura 28).

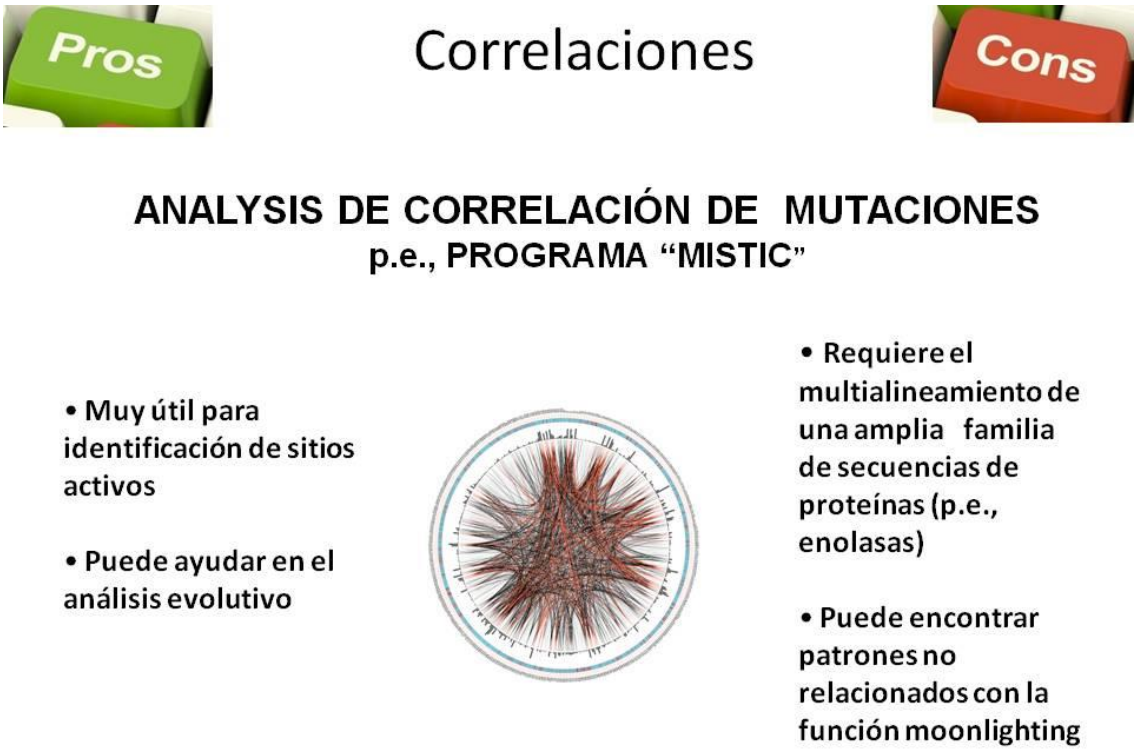


Figura 26. Pros y contras de los programas de correlación mutacional, en este caso Mistic, que permiten identificar los aminoácidos clave para la multifuncionalidad por comparación entre las proteínas que presentan una única función con las que son moonlighting. El programa requiere introducir el problema como un gran número de secuencias multialineadas

Como se muestra en las interacciones relacionadas con la unión del plasminógeno en las Figuras 29 y 30, las interacciones introducidas por una nueva funcionalidad distorsionan la red previa de dependencias mutuas entre los residuos de aminoácidos. Aún así, algunas distorsiones también se extienden alrededor de la posición 250, otra región implicada en la

interacción con plasminógeno. Este patrón alterado de correlación se propaga a la posición 280. El análisis de la estructura tridimensional de la proteína muestra que esta región está flanqueando claramente el lazo que interacciona con el plasminógeno. Es decir, la adquisición de nuevas funciones no parece estar confinada a los aminoácidos que normalmente se asocian con el patrón de unión, pero puede implicar, en algunos casos, los cambios más globales en la proteína. Estas observaciones abren una metodología para encontrar posiciones específicas donde se ha producido un cambio asociado a la adquisición de una nueva función.

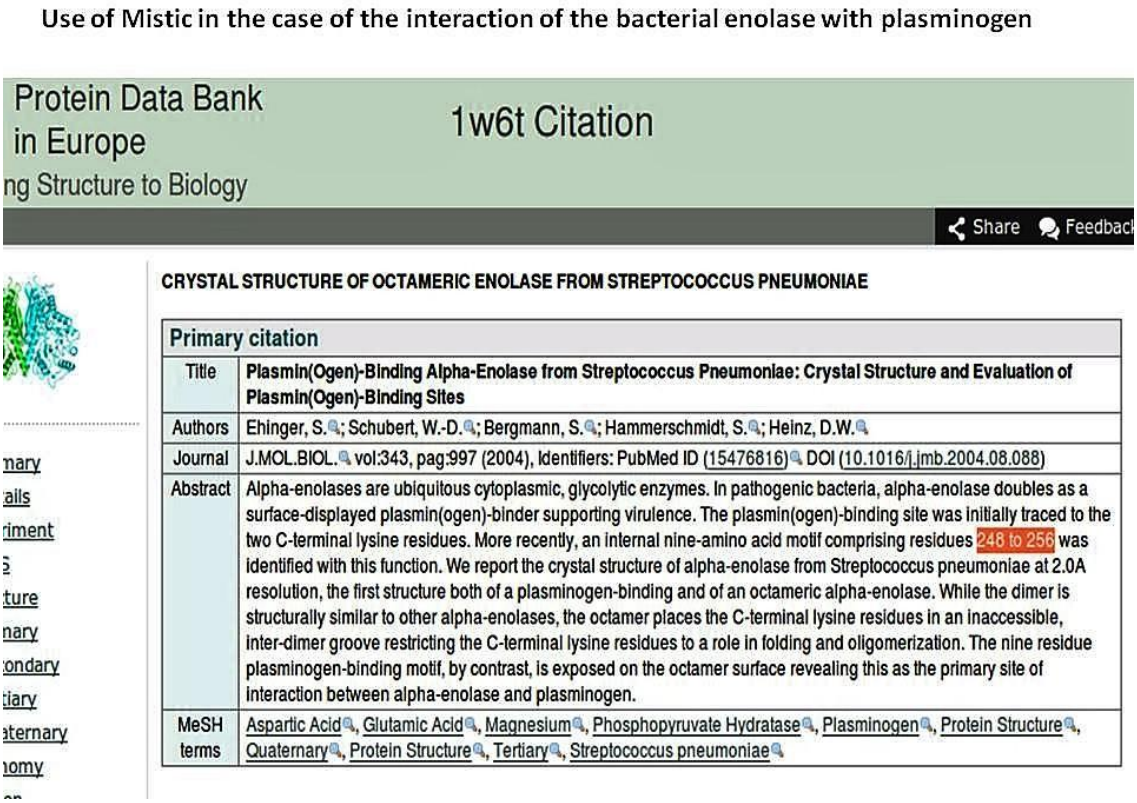


Figura 27. La enzima de la glicólisis *enolase* presenta en un cierto número de casos una segunda función como factor de virulencia de microorganismos por unión al plasminógeno del huésped. Se conocen los dominios estructurales involucrados en esta función por lo que representa ser un buen ejemplo para comprobar si el programa Mistic permite identificarlos

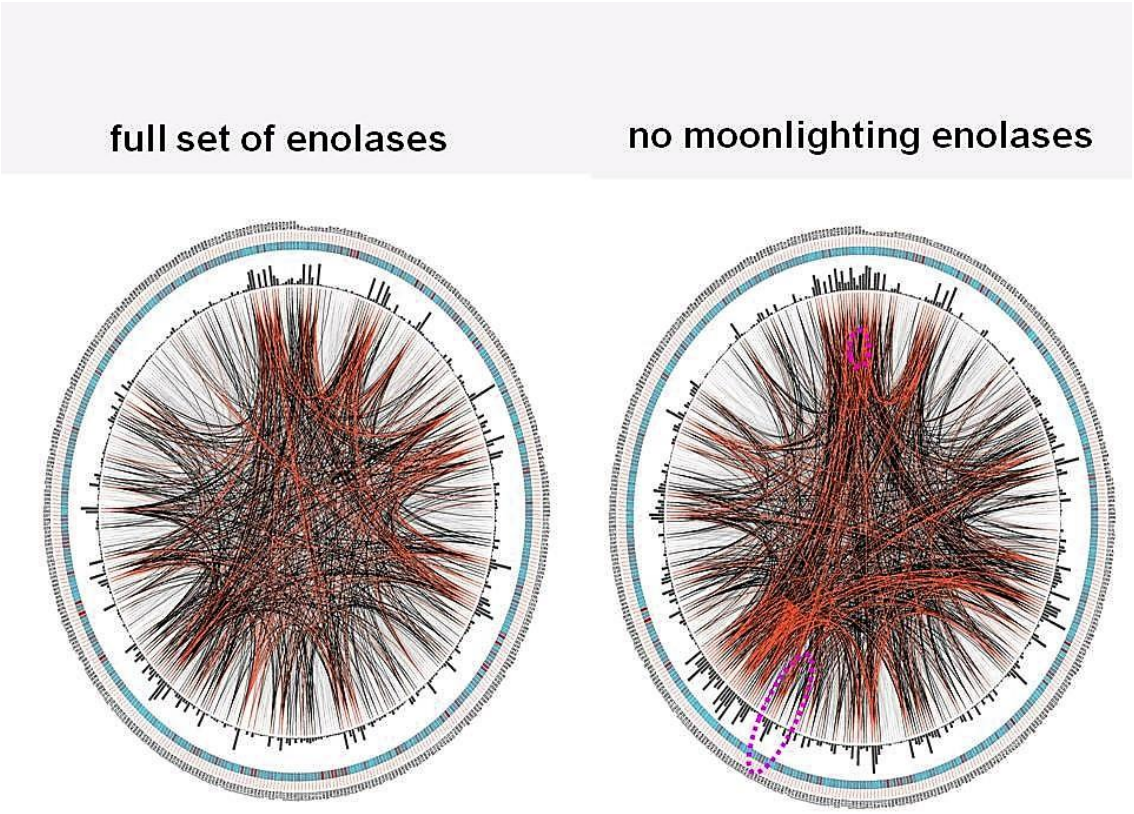


Figura 28. El *output* de Mistic muestra como un circulo la secuencia de la proteína, en este caso de numerosas enolasas alineadas. El círculo de la izquierda se ha obtenido con el conjunto de todas las enolasas alineadas y el de la derecha tan sólo con las enolasas que presentan la función moonlighting (la unión al plasminógeno del huésped)

La *enolase* presenta la conformación para complementar al plasminógeno, aunque probablemente la enolasa original carece de algunos de los aminoácidos correctos para permitir una unión suficientemente fuerte. En este caso, el análisis mediante MISTIC muestra que necesitamos solamente entre 5 a 8 mutaciones simultáneas para que ocurra la adaptación, pero esto requerería eventualmente involucrar a la reestructuración de otras regiones de la proteína que no están directamente relacionadas con la función recién adquirida, a la vez que manteniendo la estructura y el plegamiento. La *enolase* tiene una función secundaria, la adhesión al plasminógeno del huésped, que aparece muy a menudo en diferentes microorganismos patógenos y esta proteína nos permite probar si la adquisición de una nueva multifuncionalidad es un fenómeno frecuente o no. Si esto se produce sólo muy de vez en cuando, la repetición de la misma función estará vinculada a la similitud entre las diferentes proteínas, lo que indicaría que esta función ha surgido a partir de un ancestro

común de los microorganismos que contienen estas enolasas. En el caso contrario, si ninguna de las proteínas comparten esta función extra con cualquier secuencia estrechamente relacionada podría significar que la multifuncionalidad es un evento frecuente en la evolución. El resultado de estos análisis es más consistente con la segunda hipótesis. Este no es un resultado concluyente, pero es una pista interesante en el sentido de que la lista actual de proteínas multifuncionales es sólo una representación mínima de lo que podemos a esperar. Y en todo caso confirma lo que se ha mencionado anteriormente (ejemplos de la GroEL, Cpn60, GAPDH…) de que para adquirir una segunda función se requieren muy pocas mutaciones en la secuencia de una proteína.

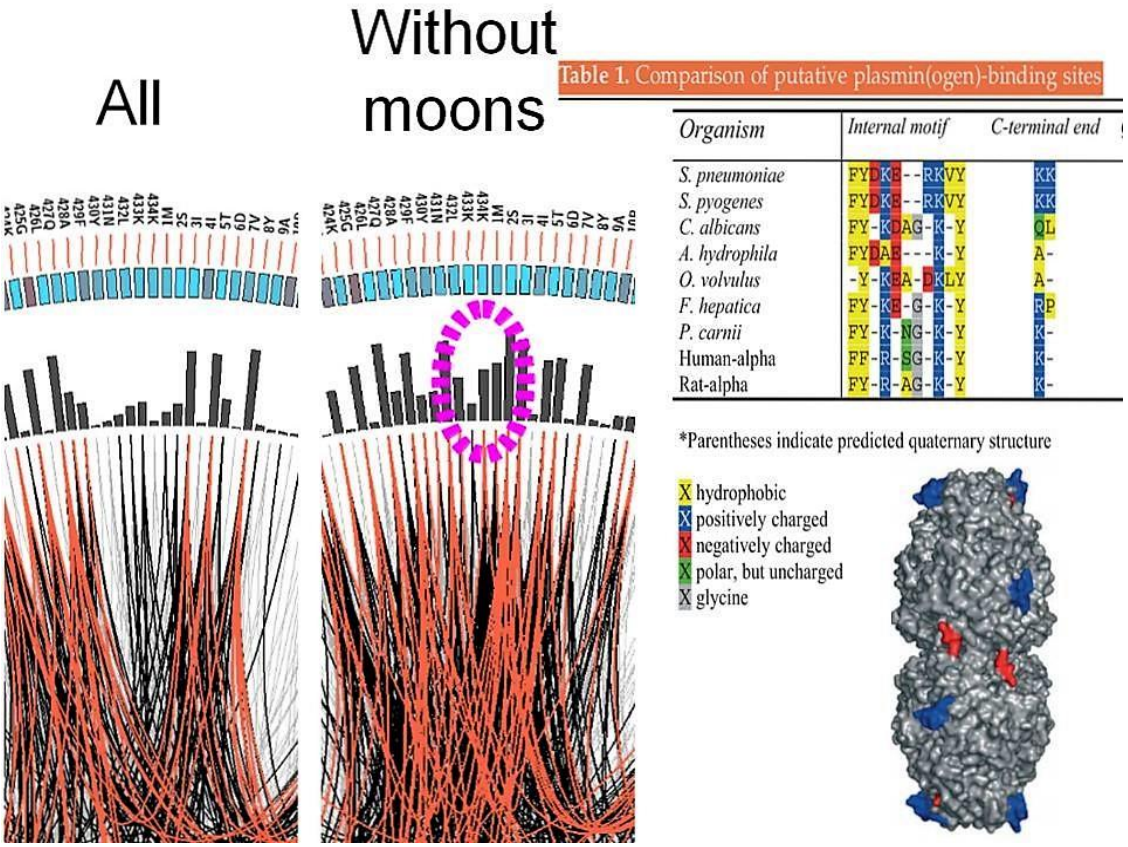


Figura 29. Ampliación de la representación de los círculos de la figura anterior en que se muestra que el programa de correlación mutacional Mistic permite identificar las regiones en los distintos dominios de las enolasas relacionadas con su función moonlighting

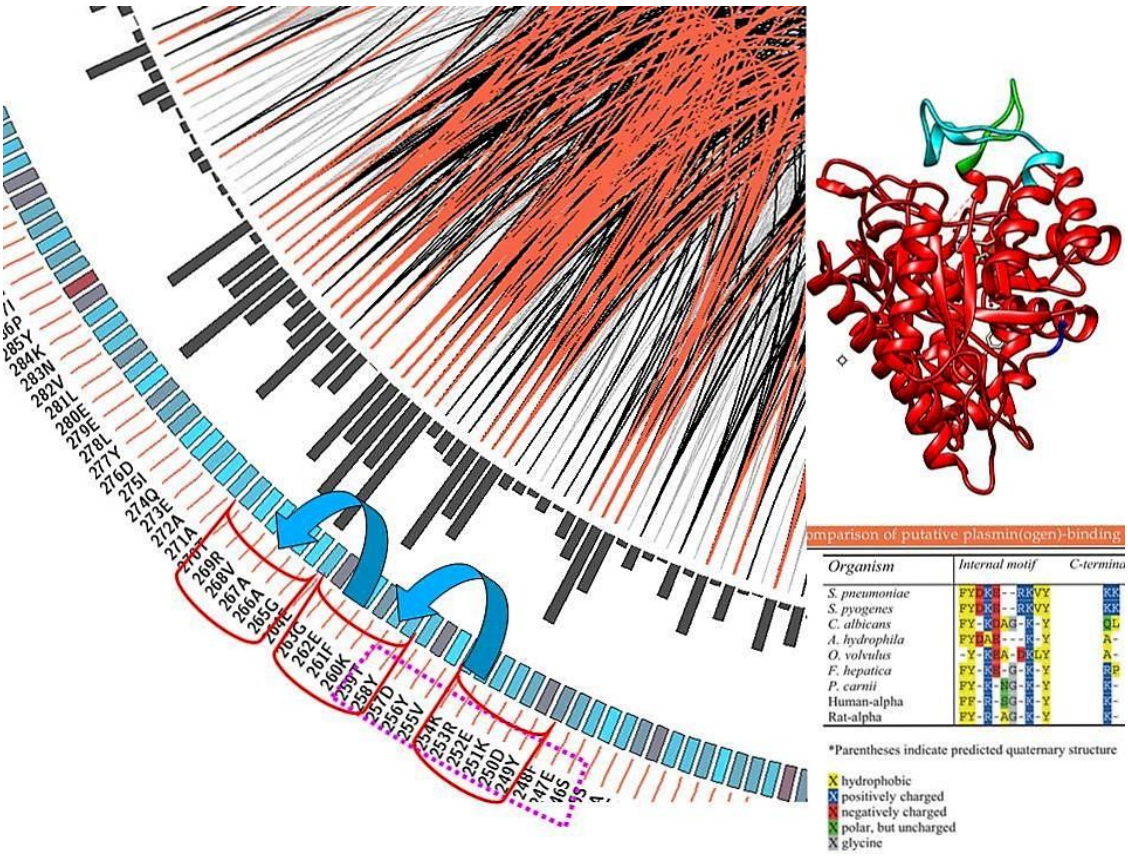


Figura 30. Ampliación de la representación de los círculos output de la figura 28 en que se muestra que el programa de correlación mutacional Mistic permite identificar las regiones en los distintos dominios de las enolasas relacionados con su función moonlighting)

IV.B.6. OTROS MÉTODOS DE ANALISIS ESTRUCTURAL 3D NO DETALLADOS EN EL PRESENTE TRABAJO

En los artículos publicados por el grupo pueden encontrarse dos aproximaciones estructurales adicionales a la predicción y mapado de las 2 funciones moonlighting mediante programas de modelado estructural como el Pisite (Higurashi et al., 2009) y Phyre (Kelley & Strernberg, 2009). No se describen en el presente trabajo dado que son objeto de otra tesis doctoral (L. Franco). Si se desea obtener información adicional sobre estos métodos y su aplcación a la identificación de proteínas moonlighting puede encontrase en los siguientes artículos: Hernández et al., 2014b y 2015.

IV.C. PREDICCIÓN DE LA LOCALIZACIÓN CELULAR DE UNA PROTEÍNA Y DE LA PRESENCIA DE HÉLICES TRANSMEMBRANA

En numerosos de casos de proteínas moonlighting (en la mayoría de casos en los organismos eucariotas) presentan cada función en un compartimiento celular diferente. Por lo tanto, los programas para la predicción de la localización celular a partir de la secuencia de la proteína pueden ayudar a predecir o corroborar una segunda función. Utilizamos dos de estos programas: PSORT (Nakai y Horton, 1999) y ProtLoc (Cedano et al., 1997). El output de ambos progamas predice posibles diferentes localizaciones en su salida de acuerdo con sus puntuaciones respectivas. Esperábamos que las dos mejores puntuaciones podrían sugerir diferentes funciones en diferentes localizaciones y verificamos si las dos principales predicciones de localización correspondían a las funciones canónicas y moonlighting. Aunque en algunos casos predicen correctamente las localizaciones correspondientes a ambas funciones, los resultados no son lo suficientemente buenos para considerarlos fiables.

La Figura 31 muestra algunos ejemplos. En S8 de Material suplementario se pueden encontrar otros muchos. En general no son predicciones en las que se pueda confiar para predecir, ayudar, la segunda función. Una de las causas principales de este resultado es que, como en el caso de los motifs Prosite, etc., al diseñar los programas de predicción se buscaba optimizar la predicción para la función canónica, en ningún caso se asumía que podía haber más de una localización válida.

La presencia de hélices transmebrana también predice localización celular y en muchos casos relacionable con función (por ejemplo la presencia de 7 hélices transmembrana y pertener a la clase GPRC), etc. Por otra parte, aunque son estructuras que se predicen muy bien, de hecho muchas proteínas secretadas no las llevan, por ejemplo las moonlighting relacionadas con virulencia. Por otra parte, como ya se ha mencionado en el Apartado IV.A. hay muy pocas proteinas multifuncionales que sean proteínas integrales de membrana.

PREDICCIÓN LOCALIZACIÓN CELULAR

PROTEÍNA E IDENTIFICADORES (Accession number, GO)	FUNCIÓN MOONLIGHT	PREDICCIÓN PSORT	PREDICCIÓN PROTLOC
Glycerol kinase GO: Cellular Component; GO:5737=cytoplasm; GO:5634=nucleus; GO:5886=plasma membrane	GK ATP-Stimulated translocation protein (ASTP) rat liver - ASTP enhances nuclear binding of the activated glucocorticoid-receptor complex. Bind to histones	cytoplasm Certainty= 0.450(Affirmative) < succ> microbody (peroxisome)Certainty= 0.362(Affirmative) < succ> mitochondrial matrix spaceCertainty= 0.100(Affirmative) < succ> lysosome (lumen) Certainty= 0.100(Affirmative)	Anchored => 4.21307800435645; Intracellular => 4.45391934777734; Nuclear => 4.46492409260957; Extracellular => 6.17035570017687; Membrane => 7.0038254966959
Tuf E. Coli Gene Ontology: Cellular Component GO:5737=cytoplasm GO:5886=plasma membrane	Receptor for host proteins	bacterial cytoplasm --- Certainty= 0.180(Affirmative) < succ> bacterial periplasmic space --- Certainty= 0.000(Not Clear) < succ> bacterial outer membrane --- Certainty= 0.000(Not Clear) < succ> bacterial inner membrane --- Certainty= 0.000(Not Clear) < succ>	Intracellular => 10.6818195604223; Anchored => 14.6427428301124; Extracellular => 16.0950740945896; Membrane => 16.2776373795691; Nuclear => 18.9136333342684
S10 E. Coli Gene Ontology: Cellular Component 47 - GO:5737=cytoplasm 12 - GO:5886=plasma membrane 7-GO:5739=mitochondrion	Antiterminator	bacterial cytoplasm --- Certainty= 0.180(Affirmative) < succ> bacterial periplasmic space --- Certainty= 0.000(Not Clear) < succ> bacterial outer membrane --- Certainty= 0.000(Not Clear) < succ> bacterial inner membrane --- Certainty= 0.000(Not Clear) < succ>	Intracellular => 10.6818195604223; Anchored => 14.6427428301124; Extracellular => 16.0950740945896; Membrane => 16.2776373795691; Nuclear => 18.9136333342684

Figura 31. Ejemplo de predicción de localización celular mediante los programas Psort y ProtLoc para tres proteínas moonlighting. En rojo se indica las predicciones correctas.

IV.D. ¿PERTENECEN LAS PROTEÍNAS MULTIFUNCIONALES A LA CLASE DE LAS PROTEÍNAS INTRINSICAMENTE DESORDENADAS?

Algunos autores han señalado que existe una relación entre las fluctuaciones conformacionales y las funciones promíscuas de las proteínas. Esta promiscuidad sería posible debido a las propiedades conformacionales de las regiones estructuralmente desordenadas. En solución las proteínas existen en una variedad de conformaciones y las regiones estructuralmente desordenadas pueden alterar sus propensiones de estructura secundaria, así como la flexibilidad conformacional en respuesta a diferentes entornos o a los partners con que interactúan (Tsai et al, 1999, 2001 y 2009; Ma et al. , 1999; Tompa et al, 2005; Amitai et al, 2007).

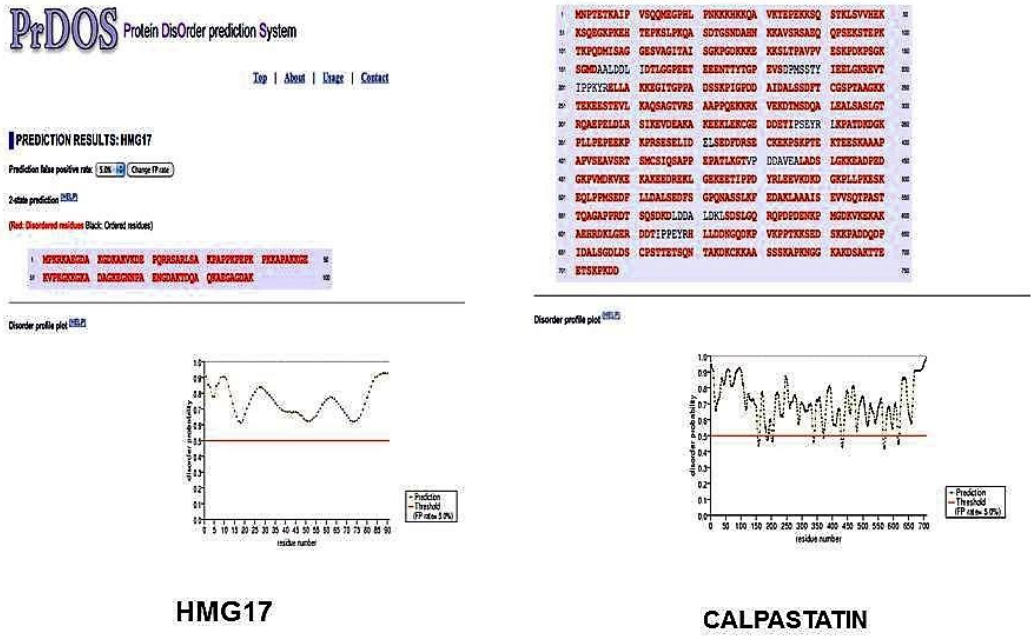
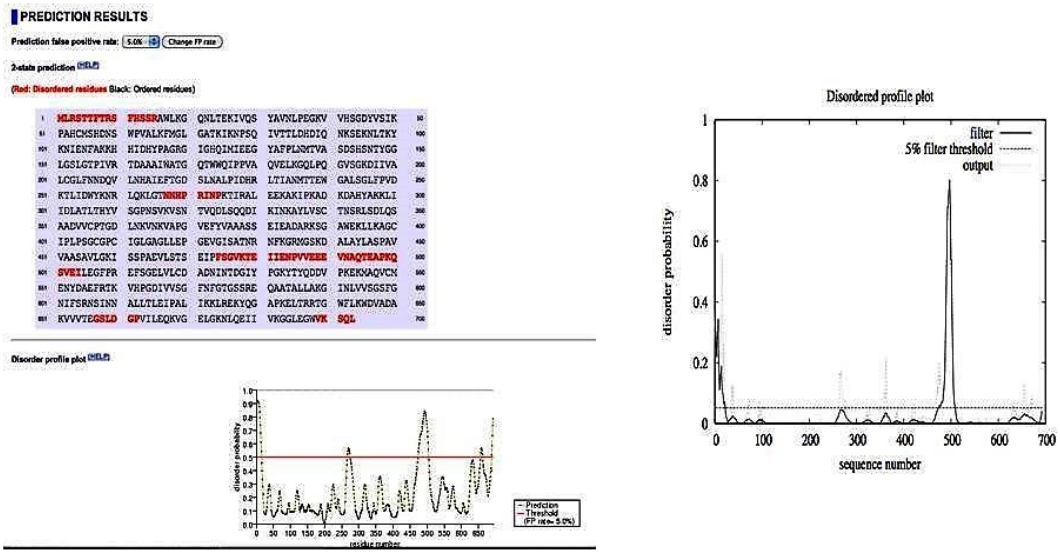


Figura 32. Ejemplos de proteínas verdaderamente IDps como control. Predicción de pertenecer una proteína a la clase de las proteínas intrinsecamente desordenadas (IDP) mediante el programa PrDOS. En este caso se trata de dos proteínas conocidas por ser IDPs, la proteína no histona HMG17 y la calpastatin. Como puede observarse la predicción es correcta

Antes de la predicción IDR del conjunto de proteínas moonlighting mencionadas anteriormente nos hicimos dos preguntas clave para determinar la fiabilidad de los programas de predicciónn de proteínas IDP/IDR: (a) ¿Son las proteínas conocidas por ser desordenadas correctamente predichas por estos programas? (b) ¿Producen los diferntes programas de predicción de IDP/IDR resultados similares? Ambas preguntas tienen respuestas positivas. Por ejemplo, como se puede ver en la Figura 32, dos proteínas desordenadas bien conocidas, la *calpastatine* y la proteína no histona HMG17, se predicen correctamente como proteínas completamente desordenadas. Las Figuras 33 y 34 muestran que los perfiles de IDP/IDR predichos utilizando diferentes programas para cuatro proteínas moonlighting son los mismos. Estos últimos cuatro ejemplos también sugieren que las proteínas moonlighting no son más internamente desordenadas que las proteínas no moonlighting. Sólo 3 de las 24 proteínas moonlighting ensayadas (en Material suplementario S9 se pueden encontrar las predicciones) muestran, hasta cierto punto, que podrían

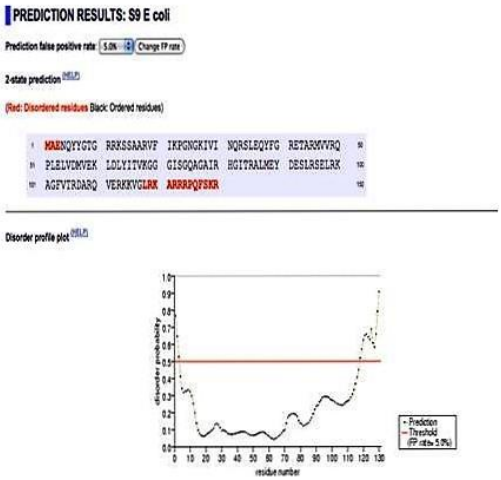
pertenecer a la clase de proteínas intrínsecamente desordenadas. De hecho, si se considera que para pertenecer a esta clase se requieren tramos de al menos 40 aminoácidos de las regiones desordenadas (Dyson, 2011) el número aún sería menor. La Figura 35 muestra que algunas de las principales proteínas multifuncionales, por ejemplo la proteína GAPDH, carecen prácticamente de regiones desordenadas, mientras que la proteína p53 sí que presenta grandes tramos de regiones IDR predichas, lo que permite incluir esta proteína en la clase de proteínas IDP. Se puede concluir que la mayoría de las proteínas moonlighting no pertenecen a la clase de proteínas IDP ya que los tramos de aminoácidos desordenados son bastante cortos y, en muchos casos, están localizados en las regiones N- y C-terminal del polipéptido que, además, suelen corresponder con regiones muy móviles. Además de esto, hemos comprobado si las regiones IDRs predichas coinciden con los lazos o tramos *coil*. La mayoría de los IDRs coinciden con lazos y regiones *coil* y, de hecho, el programa de predicción de IDP/IDRs llamado DisEMBL predice tanto regiones IDP/IDR y lazos, y por lo general coinciden. Las conformaciones locales alternativas pueden lograrse sin un gran cambio en la estructura de la proteína. A favor de nuestra hipótesis de que las proteínas moonlighting no son en general IDPs cabe mencionar que el análisis y mapado de las estructuras de rayos X de cuatro proteínas moonlighting indica que utilizan diferentes regiones para cada una de las actividades y que estas regiones corresponden a dominios o motivos bastante complejos, no a tramos de aminoácidos desordenados (Jeffery, 2004). Por supuesto, hay algunos ejemplos de proteínas moonlighting que son intrínsecamente desordenadas, como la quimioquina linfotactina humana, *human lymphotactin* (Tuinstra et al, 2008, y ver los reviews de Tompa et al 2005; Tsai et al, 2009 y Copley 2012 para algunos ejemplos), pero nuestro análisis sugiere que la mayoría de las proteínas moonlighting no pertenecen a la clase de proteínas intrinsecamente desordenadas. De hecho, una proteína moonlighting, la Proteína Ribosomal S10 de *E. coli* que fue considerada de la clase de las IDPs en que sus conformaciones alternativas le permitirían interaccionar con el Factor de Antiterminación NusB, se ha demostrado recientemente que adopta el mismo pliegue global en complejo con NusB y en el ribosoma (Figura 6). Este hecho excluye la posibilidad de que su estructura esté extensivamente remodelada y, por lo tanto, S10 se une a ARN y a NusB por diferentes regiones de la proteína. La unión a ARN se lleva a cabo por un largo bucle que es la única región desplegada (Luo et al., 2008).



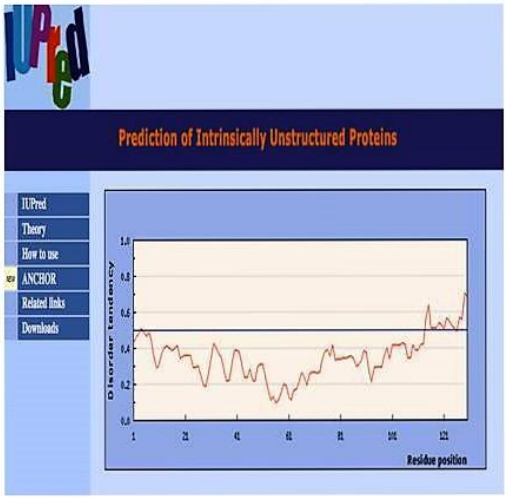
Aconitase por PrDOS **Aconitase por Disopred**

Figura 33. Puede observarse que diferentes programas de predicción de proteínas IDPs, en este caso PrDOS y Disopred, dan una predicción comparable, muy similares: en general las proteínas moonliught no son IDPs.

Se sabe que las proteínas pueden evolucionar y adoptar nuevas funciones sin cambios significativos en la secuencia. En este sentido, los resultados anteriores sugieren también que puede que no se requiera la existencia de regiones completamente desordenadas. Los lazos son lo suficientemente flexibles como para permitir la adaptación a diferentes interacciones. Las nuevas funciones podrían estar preferentemente relacionadas con establecer interacciones adicionales utilizando secuencias existentes que mediante la incorporación de nuevos tramos de aminoácidos. En Material suplementario S9 se encuentran los resultados de las predicciones de IDP/IDR.



S9 *E. coli* por PrDos



S9 *E. coli* por Iupred

Figura 34. Diferentes programas de predicción de IDPs dan resultados muy similares: en general las proteínas moonlight no son IDPs. Puede observarse que la aplicación de dos programas de predicción de proteínas IDPs, en este caso PrDOS y Iupred, aplicados a dos proteínas moonlighting apuntan a que éstas no pertenecen a la clase IDP

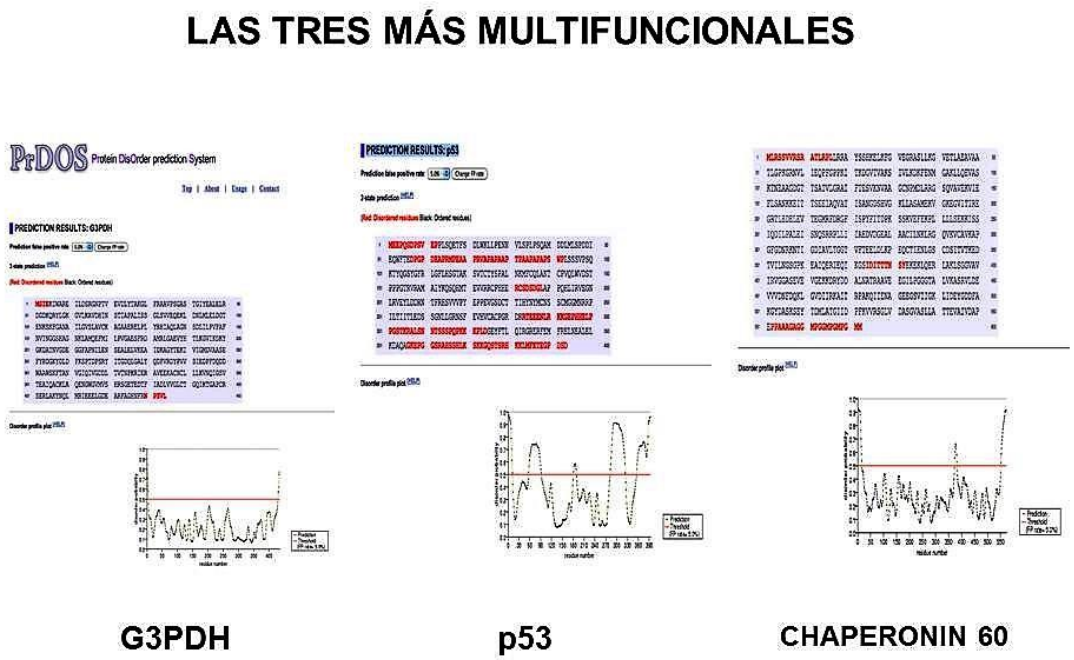


Figura 35. Puede observarse que la predicción de proteínas IDPs a las tres proteínas más moonlghting conocidas (GAPDH, p53 y chaperonin 60) apuntan a que éstas no pertenecen a la clase IDP

IV.E. ¿SE CONSERVA EVOLUTIVAMENTE LA MULTIFUNCIONALIDAD?

Dos preguntas importantes que surgen de las características estructurales y funcionales de las proteínas moonlighting son: (a) si han incorporado las segundas funciones mediante la incorporación de dominios adicionales o tramos de aminoácidos durante el proceso evolutivo y, (b) si una proteína es moonlighting en una especie implica que también sea moonlighting en otra especie filogenéticamente cercana. Aunque el bajo número de proteínas moonlighting experimentalmente demostradas impide una respuesta precisa se pueden sugerir algunas respuestas. Se llevaron a cabo una serie de análisis de multi-alineamiento de la secuencia de una serie de proteínas moonlighting conocidas con proteínas homólogas de otras especies en busca de diferencias importantes en sus estructuras (reflejados como diferencias en la longitud de la secuencia, en el grado de identidad o similitud de secuencia de aminoácidos). Concretamente se alinearon las proteínas moonlighting aconitase, alpha crystallin A y B,

neuropilin, carbinolamine dehydratase, thymidilate synthase de los organismos eucariotas *Rattus norvegicus, Homo, Mus musculus, Bos Taurus.* Asimismo las proteínas moonlighting *aconitase*, *enolase*, *aldolase*, G3PDH, *glycogen synthase*, *pyruvate kinase*, *chaperone ClpB*, OmpR, *thymidine phosphorilase*, BirA, *Lon protease*, *ribosomal protein S10* de diversos organismos procariotas *E. coli, B. subtilis, S. enterica, A. pleuropneumoniae, Thermus thermophilus.* La Figura 36 muestra como ejemplo el multialineamiento de cuatro aconitasas eucariotas que presentan un altísimo nivel de conservación de la secuencia en las diversas especies, en tanto que experimentalmente la función moonlighting tan sólo ha sido demostrada en la *aconitase* humana. Cabe decir que como muchas proteínas moonlighting presentan como función canónica una correspondiente al metabolismo primario y las enzimas de éste suelen estar muy conservadas, incluso entre taxones alejados, no es descartable que a pesar de la conservación evolutiva de secuencia no conserve la función moonlighting.

Figura 36. El Multialineamiento mediante ClustalW de diversas aconitasas eucariotas muestra la existencia de un altísimo nivel de similitud. La humana es la única en que se ha demostrado experimentalmente que sea moonlight, ¿lo son las demás?

En todo caso estos resultados corroboran trabajos anteriores (Jeffery, 1999 y 2003), que establecen que el proceso de evolución hacia la multifuncionalidad en general no se debe a la adquisición de nuevos motivos o dominios funcionales, y también sugieren que las proteínas pueden evolucionar y obtener nuevas funciones sin cambios significativos en la secuencia (ya se han mencionado varias veces a lo largo del presente trabajo los ejemplos de la GroEL, Cpn60, GAPDH, etc). Si las proteínas multifuncionales no son más conservadas que otras, la adquisición de nuevas funciones no será fácilmente detectable bioinformáticamente por el análisis de sus ortólogos. Las nuevas funciones podrían estar relacionadas más con el establecimiento de interacciones adicionales a las secuencias anteriores que mediante la incorporación de nuevos tramos de aminoácidos. Esto concuerda con los resultados de Kühner et al., 2009, para *Mycoplasma pneumoniae*, que sugiere que las proteínas que interactúan tienden a acomodar a más ligandos por superficie de interacción y que la multifuncionalidad involucra interacciones mutuamente excluyentes. Y el de Keskin y Nussinov, 2007, que muestran que un mismo motif de interacción puede servir para hacerlo con diferentes proteínas.

Un ejemplo interesante dado que se ha mapeado con precisión las regiones implicadas en sus dos funciones (su estructura 3D está resuelta) es la Proteína Ribosomal S10 de *E. coli*. Además de proteína ribosomal interviene en la regulación génica vía la interacción con el Factor de Antiterminación NusB. La proteína S10 se une al ARNr y al Factor NusB por diferentes regiones de la proteína. La unión a ARN se lleva a cabo por un largo bucle que es la única región desplegada (Luo et al., 2008). La Figura 37 muestra la estructura 3D de las dos regiones funcionales y el alineamiento de los dos motivos implicados en 10 especies bacterianas diversas (Gram positivo y Gram negativo; enterobacterias, etc). Como puede verse el nivel de conservación del Proline motif GPIPLPT correspondiente a la función moonlighting es mucho mayor que el de unión a rRNA 16s, de la función canónica, y eso pese a corresponder a un lazo de la proteína, que suelen corresponder con regiones más variables en secuencia. Esto sugiere fuertemente que las proteínas S10 ortólogas de las 10 especies presentarán también la función moonlighting. Sin embargo no se puede afirmar con seguridad mientras no se demuestre experimentalmente, entre otras cosas por que también depende del partner. Este caso representa un buen ejemplo de lo que la bioinformática puede aportar, la sugerencia de conservación de multifuncionalidad para luego ser analizada

experimentalmente. Pero nuestra hipótesis es que la multifuncionalidad se conserva evolutivamente.


```
--MEKIRLKLKAYDHRVLDRSVVAIVEAVKRSGSEIRGPIPLPTKNKRYTVLRSPHVNKDSREQFEI
MQNQRIRIRLKAFDYKLIDASTAEIVETAKRTGAQVRGPIPLPTRKERFTVLISPHVNKDARDQYEI
MQNQRIRIRLKAFDHRLIDQATAEIVETAKRTGAQVRGPIPLPTRKERFTVLISPHVNKDARDQYEI
MQNQRIRIRLKAFDHRLIDQATAEIVETAKRTGAQVRGPIPLPTRKERFTVLISPHVNKDARDQYEI
MQNQRIRIRLKAFDHRLIDQSTAEIVETAKRTGAQVRGPIPLPTRKERFTVLISPHVNKDARDQYEI
MQNQRIRIRLKAFDHRLIDQSTAEIVETAKRTGAQVRGPIPLPTRKERFTVLISPHVNKDARDQYEI
MQNQRIRIRLKAFDHRLIDQSTAEIVETAKRTGAQVRGPIPLPTRKERFTVLISPHVNKDARDQYEI
MQNQRIRIRLKAFDHRLIDQSTAEIVETAKRTGAQVRGPIPLPTRKERFTVLISPHVNKDARDQYEI
MQNQRIRIRLKAFDHRLIDQSTAEIVETAKRTGAQVRGPIPLPTRKERFTVLISPHLNKDARDQYEI
MAGQKIRIRLKAYDHEAIDASARKIVETVVRTGASVVGPVPLPTEKNVYCVIRSPHKYKDSREHFEM
MAKQKIRIRLKAFDHSLLDQSALKIVETAKTTGAKVAGPVPLPTEKDIVTILRAPHKYKDAREQFEI
MAKQKIRIRLKAYDHRILDQSAEKIVETAKRSGASVSGPIPLPTEKSVYTILLAVHKYKDSREQFEM
   ::**::***:*:  :* :.  ***:.  :*:.: **:****.:.   :: : *  **:*:::*:
                                     NusB  loop interacción RNA 16s

RVYSRLIDIISATPETVDSLMKLDLAPEVDVEVTSMETK 104 Helicobacter pylori
RTHKRLIDIVEPTDKTVDALMRLDLAAGVDVQISLG--- 103 Vibrio cholerae
RTHLRLVDIVEPTEKTVDALMRLDLAAGVDVQISLG--- 103 Escherichia coli
RTHLRLVDIVEPTEKTVDALMRLDLAAGVDVQISLG--- 103 Shigella dysenteriae
RTHKRLVDIVEPTEKTVDALMRLDLAAGVDVQISLG--- 103 Salmonella enterica
RTHKRLVDIVEPTEKTVDALMRLDLAAGVDVQISLG--- 103 Actinobacillus pleuropneumoniae
RTHKRLVDIVEPTEKTVDALMRLDLAAGVDVQISLG--- 103 Yersinia pestis
RTHKRLVDIVEPTEKTVDALMRLDLAAGVDVQISLG--- 103 Serratia marcescens
RTHKRLVDIVEPTEKTVDALMRLDLAAGVDVQISLG--- 103 Klebsiella pneumoniae
RTHKRLIDIIDPTPKTVDALMRIDLPASVDVNIQ----- 101 Mycobacterium tuberculosis
RTHKRLIDIISPSPKTVDALMRLDLPAGVDIEIKL---- 102 Clostridium botulinum
RTHKRLIDIVNPTPQTVDALMRLDLPSGVDIEIKL---- 102 Bacillus subtilis
*.: **:**:. : :***:**::**   **:::
```

El "Proline motif" GPIPLPT de la función moonlight está mucho más conservado que el motif de unión a rRNA de la función canónica

Figura 37. Estructura 3D de la proteína ribosomal S10 de *E.coli* que muestra los dos motivos responsables de sus dos funciones (unión a RNAR 16s y al factor de transcripción NusB) y su alineamiento con las secuencias ortólogas de diversas especies. Puede verse que el motif de la función moonlighting (unión al factor terminador de transcripción NusB) está más conservado que el de unión al rRNA 16s

Otro hecho que contribuye a confirmar la conservación de la multifuncionalidad está en que una serie de enzimas como la *enolase*, la *aconitase*, etc, presentan similar función moonlighting como proteínas de unión al huésped en diferentes patógenos. La tabla 1 muestra el caso de la enolasa y el Material suplementario S1 y S2 todos los casos conocidos hasta la fecha.

Finalmente cabe sugerir otra estrategia para proponer el que haya conservación de multifuncionalidad: la interactómica comparada. Si existe conservación del interactóma, o sea de los “partners” de interacción para una proteína moonlighting de dos o más especies,

probablemente hay conservación de multifuncionalidad (es la ortología de interactoma). Lo hemos tratado de analizar pero el escaso número de especies con las que se ha realizado experimentación en interactómica, y el que en muchos casos estén filogenéticamente muy alejadas, nos han impedido sacar las conclusiones pertinentes.

IV.F. PROTEÍNAS MOONLIGHTING EN ENFERMEDADES INFECCIOSAS

Ya hemos descrito que una numerosa clase de proteínas moonlighting son las relacionadas con la patogenicidad y virulencia de microorganismos patógenos (para una revision ver Henderson et al, 2011, 2013a, 2014). En nuestro caso estamos especcialmente interesados en identificar genes involucrados en patogenicidad y virulencia para el diseño de vacunas. En este sentido hemos realizado la anotación funcional de una serie de proteínas hipotéticas de un cierto número de patógenos respiratorios y predicho aquellas candidatas a ser moonlighting de acuerdo a los métodos bioinformáticos descritos en el presente trabajo. La anotación sugiere que algunas de las mismas serían posibles factores de virulencia (adhesins, haemolysins, biofilm-related genes…).

Cuando el patógeno infecta al huésped expresa una serie de genes específicos, algunos de los cuales pueden ser factores de virulencia, y por ello presentan interés para el diseño de vacunas o kits de diagnóstico. Entre las estrategias experimentales para identificar estos genes expresados diferencialmente se han desarrollado las de DNA arrays, proteómica diferencial, In Vivo Expression Technology (IVET) (Mahan et al., 1993), Signature-Tagged Mutagenesis (STM) (Hensel et al., 1995), Differential Fluorescence Induction (DFI) (Valdivia and Falkow, 1997), Selective Capture of Transcribed Sequences (SCOTS) (Baltes and Gerlach, 2004). Hemos escogido patógenos respiratorios por su importancia, tanto en humanos como en veterinaria y su facilidad de infectar otros huéspedes. Concretamente hemos analizado las proteínas hipotéticas de las siguientes especies e las que se han identificado los genes relacionados con el proceso de infección: *Actinobacillus pleuropneumoniae* (Fuller et al., 2000b; Sheehan et al., 2003; Baltes and Gerlach, 2003, 2004; Moser et al., 2004; Hodgetts et al., 2004; Jenner and Young, 2005; Jacobsen et al., 2005a,b,c; Deslandes et al., 2007; Wagner and Mulks, 2007; Hedegaard et al., 2007); *Pasteurella multocida* (Fuller et al., 2000a; Hunt et al., 2001; Paustian et al., 2002; Harper et

al., 2003; Boucher et al., 2005); *Bordetella avium* (Hot et al.,2003; Spears et al., 2003); *Staphylococcus aureus* (Palmqvist et al.,2002; Benton et al.2004); *Haemophilus influenzae* (Herbert et al.,2002; Gilsdorf et al.,2004); *Legionella pneumophila* (Edelstein et al.,1999; Polesky et al.,2001); *Pseudomonas aeruginosa* (Lehoux et al.,2002; Wang et al.,1996; Woods et al.,2004); *Streptococcus pneumoniae* (Marra et al.,2002; Orihuela et al., 2004); *Chlamydia pneumoniae* (Mahony et al.,2002); *Yersinia pseudotuberculosis* (Karlyshev et al.,2001).

En las publicaciones que describen los análisis mediante las técnicas IVET etc, antes mencionadas se describen 819 genes expresados en el proceso de infección (el listado completo puede encontrase en nuestra página http://bioinf.uab.es/JCSB). La mayoría presentan anotación funcional (desconocemos si correcta o no) pero aproximadamente el 10% son *unknown* o *hypothetical proteins*. Los hemos reanotado funcionalmente y en 24 de ellos podemos predecir el ser proteínas moonlighting, y la mitad de éstas son relacionables con factores virulencia (las anotadas como porin, adhesin, transferring binding protein, sensor histidin kinase, drug resistance transporter, etc.). Las Tablas 7 y 8 muestran los resultados.

Tabla 7: Anotación funcional de diversas proteínas hipotéticas de diversos microorganismos patógenos respiratorios

NCBI CODE	REANNOTATION
AAK19158	ref|ZP_00874620.1| Pneumococcal vaccine antigen A
AAK74324	ref|YP_336405.1| Outer membrane porin, OprD family
AAK74615	ref|YP_441197.1| Sensor histidine kinase
AAK74938	ref|YP_139191.1| Protease
AKK75508	ref|NP_391047.1| Two-component sensor histidine kinase
AAK76264	gb|AAK94503.1| Glucan-binding protein B
NP_244983	gb|AAF68414.1|AF237928_1 Putative filamentous hemagglutinin
NP_872680	gb|AAC43485.1| Transferrin binding protein 1
ZP_00135039	ref|YP_823923.1| Acriflavin resistance protein
HI0894	ref|ZP_00156756.2| COG0845: Membrane-fusion protein
P75830	ref|AP_001509.1| Macrolide transporter subunit, membrane fusion protein (MFP)
NP_878961	ref|NP_406637.1| Beta-lactamase induction signal transducer AmpG.
NP_879219	ref|YP_776657.1| Haemin-degrading family protein
NP_879893	ref|NP_882467.1| Adhesin
YP_001628663	ref|ZP_00944928.1| Tricarboxylate-binding protein
YP_001630768.	ref|YP_550179.1| Twin-arginine translocation pathway signal
YP_001630856	ref|YP_550179.1| Twin-arginine translocation pathway signal
YP_001632413	ref|YP_548917.1| Twin-arginine translocation pathway signal
SP0288	ref|ZP_01047380.1| Abortive infection protein protein
SP0703	ref|NP_698984.1| Intimin/invasin family protein
SP0958	ref|NP_835060.1| Bacitracin transport permease protein BCRB
SP1240	ref|ZP_00874693.1| Abortive infection protein
SP1926	ref|NP_979219.1| Drug resistance transporter, EmrB/QacA family
SP2143	ref|ZP_00602653.1| Glycoside hydrolase

Tabla 8: Predicción de posibles proteínas moonlighting de entre las proteínas hipotéticas de diversos microorganismos patógenos

NCBI CODE	PUTATIVE MOONLIGHTING FUNCTIONS	BYPASS	PRODOM
AAK74326	Protease	x	x
	Sensor Histidine Kinase	x	x
AAK74089	Metalloprotease	x	x
	Fibronectin-binding	x	x
AF109148	Antigenic glutamine-binding	x	x
	Sensory/regulatory system		x
NP_002036	Receptor alpha interferon chain	x	x
	Calmodulin-binding membrane	x	x
NP_002189	Interferon regulatory factor	x	x
	Phosphate/Sulphate sodium permease		x
NP_246423	Transmembrane permease	x	x
	Oxidoreductase	x	
NP_246553	Membrane outer P2 precursor	x	x
	Hydrolase AD-ribose pyrophophatase	x	x
NP_879578	Hemolysin	x	x
	Adenylate cyclase	x	x
NP_879580.	cyclolysin secretion protein	x	x
	RTX toxin transporter	x	x
NP_880309	Ppgpp synthetase II	x	x
	Guanosine-3',5'-bisdiphosphate 3'-pyrophosphohydrolase	x	x
NP_881353.	proline dehydrogenase	x	
	Delta-1-pyrroline-5-carboxylate dehydrogenase	x	x
NP_881358	penicillin-binding protein	x	
	Glycosyl transferase	x	x
NP_881613	hemolysin	x	
	4-hydroxyphenylpyruvate dioxygenase	x	x
NP_357945	Choline binding protein	x	
	N-acetylmuramoyl-L-alanine amidase	x	x
SP0288	Uridine kinase	x	x
	Sodium extrusion protein	x	x
	Protease	x	x

IV.G. ¿SON LAS PROTEÍNAS MULTIFUNCIONALES PROPENSAS A ESTAR ASOCIADAS CON PATOLOGÍAS HUMANAS?

Se han descrito una serie de ejemplos de proteínas moonlighting asociadas a enfermedades humanas (Sriram et al., 2005;Jeffery, 2011). Nuestra base de datos MultitaskProtDB contiene más de un centenar de proteínas humanas, y en una próxima actualización de la misma unas 132. Una búsqueda de cuantas proteínas moonlighting humanas corresponden con ejemplos de proteínas ligadas a enfermedades descritas en las bases de datos *Online Mendelian Inheritance in Man* (OMIM) (http://www.ncbi.nlm.nih.gov/omim o en http://omim.org/) y *Human Gene Mutation Database* (HGMD) (http://www.hgmd.org/) muestra que el 77% de ellas lo están (Figura 38). Esto sugiere que las proteínas moonlighting son muy propensas que la media a estar involucradas en enfermedades. Aunque se desconoce todavía el alcance del proteoma humano es poco probable que el 76% del mismo esté involucrado en patologías, pero si las proteínas multifuncionales. También hemos encontrado que el 47% de las dianas farmacéuticas conocidas son proteínas moonlighting.

Estas cuestiones serán tratadas más ampliamente en otra tesis doctoral en desarrollo en el grupo (L. Franco).

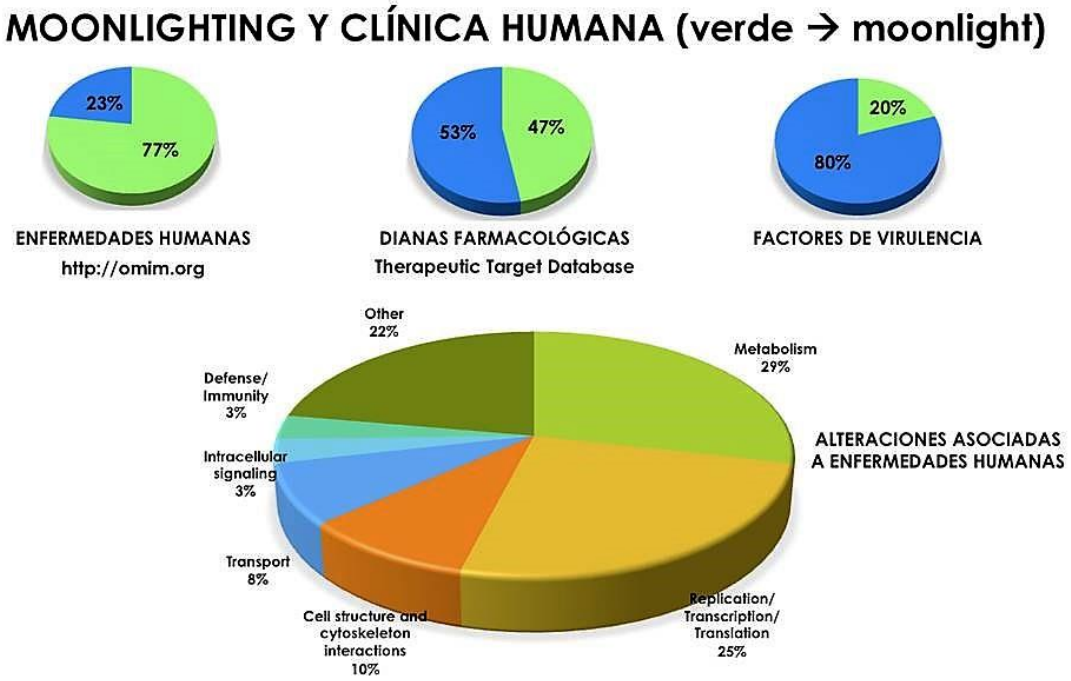


Figura 38. Esquema mostrando que el 77% de las proteínas moonlighting humanas están relacionadas con enfermedades, que un 47% son dianas de fármacos existentes y que el 20% de las proteínas moonlighting en general son factores de virulencia de microorganismos patógenos

V. DISCUSION GENERAL

La predicción de la función de una proteína es una tarea dificil. Tan solo vale la pena recordar que tres proteínas con una gran importancia clínica, investigadas a lo largo de muchos años en numerosos grupos acádemicos o industriales, las proteínas relacionadas con el Alzheimer, Prion y Corea de Huntington, a estas alturas se desconoce en detalle su función biológica. Y todavía es más difícil cuando se trata de una proteína multifuncional (Gómez et al, 2003 y 2011; Khan et al, 2012 y 2014a,b; Hernández et al., 2014b, 2015). La predicción de proteínas multifuncionales es muy útil para los investigadores en tareas como la anotación funcional de nuevos genomas, la interpretación de los experimentos de noqueo de genes en los que la eliminación de un gen produce resultados fenotípicos inesperados, y como hemos descrito, también para la identificación de dianas terapéuticas y diseño de fármacos y vacunas, los efectos secundarios de los fármacos, etc.

En el desarrollo del presente trabajo hemos creado la primera base de datos de proteínas multifunciones existente (MultitaskProtDB) cuya actualización (más del doble de proteínas y añadiendo más características) esperamos enviar a publicar próximamente. La base de datos MultitaskProtDB puede ser de gran ayuda a los investigadores para identificar las características de las proteínas multifuncionales y profundizar en su función biológica y en su evolución.

Desde el objetivo de la predicción de las proteínas multifuncionales podemos indicar que a nivel global, el algoritmo de homología remota Psi-Blast es una muy buena herramienta, si bien en la práctica y utilizándo tan sólo este algoritmo es difícil para el investigador identificar las mejores dianas candidatas de la larga lista de dianas que salen en el output. Por otra parte las anotaciones de los bancos de datos de secuencias de proteínas no siempre son precisas y abundan las ambigüedades y anotaciones de baja calidad. Aunque no lo hemos podido aplicar a toda nuestra base de datos, el algoritmo HMMER de análisis de secuencias es muy prometedor y complementario al PsiBlast. Como hemos descrito, la combinación de diferentes algoritmos y bases de datos bioinformáticos y experimentales para el análisis de secuencias de proteínas puede ayudar a reducir el número de secuencias candidatas y revelar posibles proteínas moonlighting. En nuestra opinión, en este momento el mejor

enfoque es combinar los análisis de tipo PsiBlast (y probablemente HMMER) con los resultados existentes en las bases de datos de interactómica (PPIs). Por ahora este análisis se ha de realizar por inspección manual, pero otro miembro del grupo (M. Huerta) está tratando de automatizarlo mediante una herramienta que está diseñando (e incorporando más bases de datos, por ejemplo las de expresión génica, etc). Hemos determinado (Tabla 5a,b) que la combinación de PsiBlast+PPIs (y en un cierto número de casos Pfam y ProDom) conduce a la predicción correcta de alrededor del 30% de las proteínas moonlighting, con un buen nivel de especificidad (un menor número de falsos positivos) y sensibilidad (un menor número de falsos negativos). Pero si consideramos los casos en que cualquiera de los dos por separado, o PsiBlast o PPIs, identifica la proteína la predicción es muy alta pero en estos casos aumenta la sensibilidad a costa de la especificidad. También hay que tener en cuenta que en nuestro análisis hacíamos uso de una base de datos de proteínas en que previamente se ha demostrado que son moonlighting; llevar a cabo esta tarea para proteínas desconocidas es más difícil y como ya se sabe, en el actual "estado del arte" requiere finalmente la determinación experimental. En todo caso PsiBlast identifica mejor las proteínas moonlighting que son multidominio y en las que cada dominio presenta una función independiente que aquellas en que las dos funciones están solapándose. La Figura 39 de una publicación previa del grupo (Gómez et al., 2003) muestra una serie de ejemplos de proteínas moonlighting multidominio y las predicciones boinformáticas de las mismas.

Table V.3: Moonlighting proteins that are multi-domain proteins.

MOONLIGHTING PROTEIN	PSI-BLAST		PRODOM	BYPASS
	Positives	Accesion number		
a.FtsH protease (*T.thermophilus*) b.Chaperone activity	+	gi5231279	+	+
	+	gi12724524	+	+
a.Uracil-DNA-glycosylase (*Homo*) b.Glyceraldehyde-3-phosphate DH	+	gi35053		+
	+	gi11436283	+	+
a.CFTR chloride channel (*Homo*) b.Regulator of Na+ channels	+	gi453286	+	+
	+	gi1405353	+	+
a.Thymidine phosphorylase (*Homo*) b.PD-ECGF	+	gi136588	+	+
	+	gi136588		+
a.Neuropilin (*Homo.*) b.VEGFR, regulation of angiogenesis	+	gi2978560	+	+
	+	gi2978560	+	+
a.Aconitase (*Homo*) b.IRE-BP	+	gi8659555	+	+
	+	gi896473	+	+
a.Carbinolamine dehydratase (Rat) b.Dimerization factor	+	gi423757	+	+
	+	gi3127813	+	+
a.Thymidylate synthase (*Homo*) b.DHFR	+	gi4507751	+	+
	+	gi1169423	+	+
a.BirA biotin synthetase (*E.coli*) b.Bio operon repressor	+	gi13364376	+	+
	+	gi11497694	+	+
a.Lon protease (*E.coli*) L12349 b.Chaperone activity	+	L12349	+	+
	+	sp083985	+	+

Symbols: + true positive; - true negative; false +; false -.
PSI-BLAST: default parameters (BLOSUM62, expected 10, inclusion threshold: 0.002, database: non redundant (NCBI)
a, b: Functions of the moonlighting proteins. a) One function, b), additional function

Figura 39. Tabla de una publicación previa del grupo (Gómez et al., 2003) mostrando una serie de ejemplos de proteínas moonlighting multidominio correctamente predichas por PsiBlast, Bypass o ProDom

El equilibrio adecuado entre la sensibilidad y la especificidad es especialmente difícil en el caso del análisis bioinformático de las proteínas moonlighting. El grupo de Kihara (Khan et al, 2014a,b) ha enfocado el problema de encontrar las funciones moonlighting utilizando tan sólo información de secuencias por anotación en los términos de GO. Utilizando el algoritmo PsiBlast, el método PFP (Protein Function Prediction) realiza una búsqueda única con un e-value no restrictivo. Por el contrario, el método de ESG (Extended Similarity Group) utiliza la primera búsqueda para lanzar nuevas búsquedas diferentes a partir de las secuencias encontradas ampliando enormemente el espacio de secuencias exploradas en el siguiente paso de la búsqueda. Estos autores utilizan PSI-Blast con la matriz BLOSUM45 porque consideran que puede capturar mejor los homólogos remotos en las iteraciones iniciales, por lo que esto incrementa la eficiencia de la búsqueda de funciones moonlighting y evita la degeneración PSSM (Kahn et al, 2012;. 2014a, b). En nuestro caso, hemos querido utilizar el algoritmo Psi-Blast para detectar las funciones moonlighting, pero con el fin de aumentar la

capacidad de búsqueda de homólogos remotos hemos utilizado un mayor número de iteraciones, cinco por defecto. Nuestra estrategia es diferente, pero ambos tratamos de aumentar la sensibilidad sin comprometer la especificidad. Por los resultados obtenidos utilizando PSI-Blast, podemos ver que hay una hiper-saturación de las funciones canónicas que son, por otra parte, anotadas en las más diversas formas. Estos descriptores tienen que ser analizados mediante la concentración de ellos en un número reducido de expresiones, aunque hay algunos descriptores muy abundantes (Figura 40, panel B). Hay tantas variantes en la anotación de la función de la proteína, que este hecho complica el análisis.

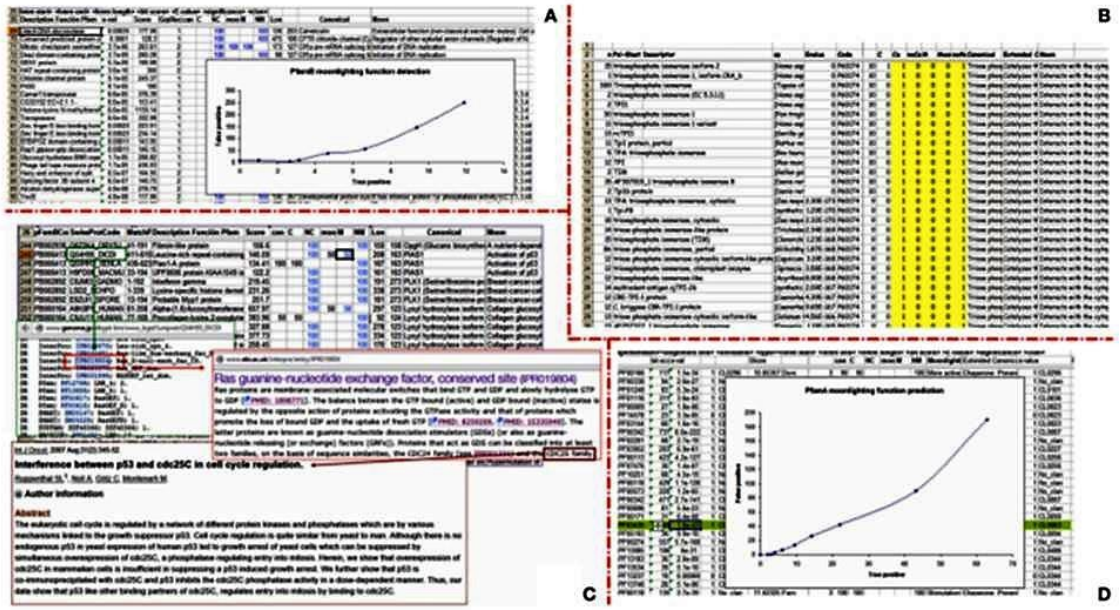


Figura 40. El problema de la predicción da las proteínas moonlighting. (A) El panel muestra que las bases de datos que contienen una alta heterogeneidad de funciones como PfamB permiten la identificación de las funciones no canónicas que no se identificarían utilizando las bases de datos de patrones con alta homología secuencial (por ejemplo PfamA). Sin embargo, esto implica que da lugar a una mayor tasa de falsos positivos comparando con PfamA (como puede verse en el panel D). (B) Un intento de explotar la variabilidad de anotación en bases de datos no redundantes también presenta sus costes dado que la hipersaturación de anotaciones de funciones canónicas contiene toda clase de sinónimos. (C) Revisar la documentación suplementaria puede ayudar a identificar detalles relevantes relacionados con la segunda función, moonlighting. (D) La relación entre falsos positivos y verdaderos positivos nos da una idea del compromiso entre especificidad y sensibilidad. Como puede observarse cuando los *scores* se relajan, aunque todavía podamos encontrar

nuevas funciones moonlighting, el número de falsos positivos incrementa abruptamente

Para ilustrar el problema de la especificidad y la sensibilidad en la búsqueda de las proteínas monlighting, se utilizó el programa PfamA y B. Estos dos algoritmos representan muy bien todos los métodos que se basan en perfiles o HMM. En el caso de Pfam A, las puntuaciones más restrictivas son capaces de encontrar algunas funciones moonlighting con un bajo número de falsos positivos, pero la relajación del umbral de corte aumenta la proporción de falsos positivos. En contraste con las familias de proteínas PfamA, los perfiles PfamB están compuestos por muchas más familias de proteínas y que representan diferentes funciones. Esta diversidad en las funciones que se encuentra en las familias de proteínas PfamB, y el hecho de que las secuencias incluidas en las familias PfamB no están presentes en los perfiles PfamA, aumenta la sensibilidad del método PfamB en comparación con la detección de la función por PfamA (Figura 40, panel A). Hay que tenerse en cuenta que este aumento de la sensibilidad tiene un cierto coste en la especificidad, porque con valores de corte restrictivos se disminuyen el número de falsos positivos detectados (Figura 40, panel D). En el análisis con PfamB, y como un parámetro para describir la significación estadística, hemos considerado la relación entre el e-value de la familia PfamB dividido por la puntuación de los miembros de la familia PfamB, que proporciona el descriptor utilizado para la función. Lo que se observa es que con valores más restrictivos, mayor es la acumulación de un elevado número de verdaderos positivos. Por el contrario, cuando este parámetro se relaja, el número de falsos positivos aumenta rápidamente, creciendo cerca de diez veces más rápido que en el caso de los análisis realizados usando PfamA. En otras palabras, con el fin de conseguir una mayor sensibilidad en la detección de las funciones de moonlighting con PfamB, hemos tenido que sacrificar mucho de especificidad. Encontramos que cuando se relajó este parámetro de corte el número de falsos positivos aumentó casi diez veces más rápido que en el caso de los análisis realizados usando PfamA. En todo caso, a partir de la experiencia obtenida con el presente trabajo sugerimos utilizar ambos, PfamA y PfamB, y el investigador que analiza manualmente los resultados puede tomar una decision basada en su experiencia o en datos experimentales adicionales.

No hay que olvidar que el objetivo de este trabajo ha sido fundamentalmente explorar diferentes metodologías bioinformáticas que se pueden utilizar para buscar las funciones moonlighting de las proteínas, no el calcular los parámetros estadísticos. Al mismo tiempo,

estamos tratando de encontrar posibles pistas que nos puedan hacer entender los mecanismos subyacentes al proceso de evolución de estas funciones moonlighting. Uno de los problemas que surgen a partir del análisis de los datos es determinar si existe alguna correlación entre la puntuaciónn estadística y la importancia biológica. En este sentido hay que mencionar lo que dicen Bork y Koonin: *Statistical and biological significance are not equivalent (sometimes, biological important matches can have relatively low scores. The first hit probably isn't the most informative (in terms of function) (*Bork & Koonin, 1998). Sin embargo es importante determinar algunos parámetros estadísticos para resumir la distribución de los datos sobre la importancia biológica. Como ya ha descrito Kihara, el método estándar, curvas ROC, tiene un problema fundamental, que es el cálculo del número de verdaderos negativos. Esta no es una tarea fácil, especialmente si no se está trabajando con una base de datos diseñada especialmente para calcular estos parámetros sin problemas. Además, el cálculo de una curva ROC es interesante si se está desarrollando un nuevo método en el que se desea establecer un punto de corte en la que el compromiso entre sensibilidad y especificidad sea el principal objetivo. En nuestro caso, todavía no estamos en esa fase sino que sólo estamos utilizando los métodos ya existentes para determinar su potencial para identificar las funciones moonlighting. En este sentido, es clave analizar la relación entre los verdaderos positivos y los falsos positivos mediante la implementación de un parámetro de puntuación (*score*) más restrictivo. Aunque todas las proteínas incluidas en la base de datos utilizada tienen al menos una función moonlighting, podríamos considerar que si no se ha encontrado ni una sola función (canónica o moonlighting) podría considerarse como un falso negativo, sin embargo, realmente no sabemos si esa suposición es, de hecho, verdadera.

Por ejemplo, cuando se utiliza PfamA, no sabemos si una familia con la misma función de ese perfil en realidad existe en la base de datos PfamA. Aunque un descriptor similar está presente, no podemos estar seguros de que esta función realmente corresponde a la misma familia de proteínas, ya que también es posible que ésta llegara a la misma función, pero utilizando un plegamiento y una conformación 3D diferente, y que no se comparte con la proteína problema. Por otra parte, no existen falsos negativos en la base de datos de perfiles, por lo tanto, podríamos estar haciendo una suposición falsa. Sin embargo podemos utilizar este tipo de herramienta para extraer la información precisamente porque no tiene los problemas asociados con el uso de las curvas ROC.

En todo caso, los métodos basados en perfiles, que tienen en cuenta las regiones más grandes de la proteína que los métodos basados en patrones, parecen funcionar mejor, sobre todo si las secuencias utilizadas para construir el perfíl no se han refinado en exceso. Este refinamiento elimina gran parte de la diversidad necesaria para encontrar los miembros remotos de la familia funcional de las proteínas.

Como se ha comentado anteriormente, el objetivo de este trabajo no es tanto diseñar nuevas herramientas para la predicción de las funciones moonlighting, sino más bien explorar si las herramientas existentes se pueden utilizar para identificar las proteínas multifuncionales. Algunos de los métodos utilizados son difíciles de sistematizar o no son fáciles de implementar como métodos automáticos, ya que requieren la interpretación de un lenguaje muy ambiguo y lleno de sinónimos. Nuestro objetivo es explorar si los métodos que utilizan los sistemas de anotación jerárquicos, más apropiados para estos fines, funcionan o no. Por otra parte, la evaluación del acierto de la función encontrada con la verdadera función moonlighting sigue siendo algo muy subjetivo, porque a pesar de que hace que sea posible obtener resultados similares en búsquedas sucesivas, si buscamos la misma proteína problema ejecutando la búsqueda en contra de la misma base de datos de secuencias, el procedimiento que calcula las similitudes no está exenta de un cierto grado de subjetividad. Esto es inherente a la necesidad de la simplificación requerida para parametrizar los descriptores con el fin de compararlos por métodos bioinformáticos. Los diferentes métodos para calcular estas similitudes nos dan una idea global de la similitud, pero sus resultados pueden variar en función del método utilizado para anotar las predicciones correctas y erróneas. Incluso si trabajamos con categorías predefinidas, tales como los términos de GO, a menudo lo que hacen es imponer limitaciones en la definición de la función debido al número limitado de términos GO predefinidos (9500 para *Molecular function*). Esto provoca la pérdida de algunos aspectos importantes de la función de la proteína a pesar de que aumenta la posibilidad de una coincidencia entre el término predicho y los términos semánticos originales. Los términos de los “troncos” en los árboles de GO son más inespecíficos, y estos términos serán cada vez más específicos, siempre y cuando se encuentren cerca de las hojas del árbol de GO, por lo que la identificación de las funciones de los padres tendrá poca o ninguna relevancia en algunos casos (Figura 40, panel C).

En los términos de GO, no toda la información de la función de la proteína se condensa. Por ejemplo, en el caso descrito a continuación, la de la proteína PIAS1, la segunda función está

relacionada con la activación de la proteína p53. Cuando observamos la lista de resultados obtenida, el subconjunto preseleccionado de funciones (que se detallan en el siguiente párrafo) tomado de los descriptores de las proteínas incluidas en el grupo (*seed of PfamB*) de PfamB, vemos una proteína con la misma función que la proteína canónica, otra con una función desconocida y la proteína Q54H95_DICDI que, al parecer, por las funciones descritas en GO (IR-0005622 intracelular; 0005085 actividad del factor de intercambio de nucleótidos de guanilo; 0051056 regulación de la transducción de señales mediada por GTPasa), mantiene una relación muy remota para la activación de la función de la actividad de p53, como la función más similar. Esta proteína podría estar relacionada con la regulación de algunas vías a través de un proceso de transducción. Sin embargo, nada más se puede concluir de estos datos que esté relacionado con la función moonlighting (Figura 40, panel C).

Sin embargo, si en lugar de utilizar el término GO expandimos la búsqueda, podemos ver que esta proteína tiene un patrón interno común en una familia de proteínas (CDC24 y Cdc25), que a su vez pueden tener interacciones con la proteína p53. Ahora la función que buscábamos está más cerca de la función moonlighting esperada y, al mismo tiempo, esta proteína también está implicada en la regulación del ciclo celular. Y podemos identificar la interacción correcta entre 25.000 proteínas del proteoma humano, y también está reconocida como una proteína reguladora, por lo que prácticamente se podría dar un 100% de coincidencia. Hay que tener en cuenta que con el fin de encontrar la función correcta, era necesario leer la documentación complementaria asociada con el perfíl. La identificación de la función real no es posible sólo mediante la lectura del descriptor principal de la función asociada con el código GO. Es por ello que otros investigadores podrían no encontrar el descriptor correcto, ya que pueden haber abortado la búsqueda en cualquier momento antes de la localización de la referencia correcta en la lista de referencias del artículo (Figura 40, panel C).

En resumen, podemos decir que la naturaleza del fenómeno moonlighting es tan variada que será difícil crear una única herramienta para hacer frente a todos los problemas de la búsqueda de las funciones moonlighting.

Dos enfoques adicionales pueden ayudarnos a predecir una proteína como verdadera moonlighting. Uno de estos enfoques es la alineación de la secuencia problema con

estructuras 3D conocidas del PDB, lo que además ayuda a mapar ambas funciones en la estructura de la proteína. Esta aproximación la hemos realizado mediante el programa PISITE, pero es objeto de otra tesis del grupo. En todo caso puede encontrase la descripción del procedimiento y los resultados en dos recientes publicaciones del grupo (Hernández et al., 2014b, 2015). El otro procedimiento es el análisis de la correlación de mutaciones de aminoácidos, lo que puede sugerir pistas para la evolución de la multifuncionalidad cuando se comparan con ejemplos mono-funcionales en la familia. En nuestro caso lo hemos aplicado, mediante el programa MISTIC, al caso de la enolasa (como se ha comentado en Métodos se requiere una amplia familia de proteínas para incorporarlas al programa como un multialineamiento). Como se ha descrito en el Apartado IV.B.5., un interesante resultado es que se necesitan tan sólo entre 5 y 8 mutaciones simultáneas en la enolasa para que la adaptación a la función moonlighting (unión a plasminógeno en este caso) pueda ocurrir. Y también se ha comentado en Resultados que esta proteína nos permite probar si la adquisición de una nueva función es un fenómeno que ocurre con frecuencia o no. Si esto se produce sólo muy de vez en cuando, la repetición de la misma función estará vinculada a la similitud entre las diferentes proteínas, lo que indica que esta función ha surgido a partir de un ancestro común de los microorganismos que contienen estas enolasas. En el caso contrario, si ninguna de las proteínas comparte esta función extra con cualquier secuencia estrechamente relacionada, podría significar que la multifuncionalidad es un evento frecuente en la evolución. El resultado de estos análisis es más consistente con la segunda hipótesis. Este no es un resultado concluyente, pero es una pista interesante en el sentido de que la lista actual de proteínas multifuncionales es sólo una representación mínima de lo que podemos esperar.

Como se ha mencionado en la Introducción (Apartado I.I.) el fenómeno de la multifuncionalidad conduce a un cierto número de preguntas importantes tanto para la función como para la evolución de proteínas que todavía no tienen respuesta. Por ejemplo: ¿qué ventaja evolutiva representa el que una proteína tenga más de una función, a veces ambas funciones indispensables, en vez de utilizar la duplicación de un gen y evolución independiente del parálogo, lo cual evita el denominado “conflicto adaptativo”? o ¿Cuál es el mecanismo que conduce a la aparición de la segunda función? o ¿hay conservación filogenética de la multifuncionalidad? Aunque en el presente trabajo no se puede dar

respuesta a las mismas sí merece dedicarles algunos comentarios. Vamos a ampliar lo descrito en la Introducción y sugerir algunas hipótesis adicionales.

Respecto a la primera pregunta ¿qué ventaja evolutiva representa el que una proteína tenga más de una función? Ya se ha mencionado que existe una doble respuesta: (a) que con menos genes el organismo puede llevar a cabo más funciones y (b) la Vida más que diseñar polipéptidos enteramente nuevos suele reutilizar lo previamente existente, pues ya ha sido optimizado para dar lugar a un plegamiento y conformación estables y es más fácil añadir una función a una estructura previa que en muchos casos, como ya se ha mencionado en el apartado I.E., tan sólo requieren unos pocos cambios en aminoácidos (Jeffery, 2015). Aunque aparentemente "no falte" DNA, especialmente en organismos superiores, la síntesis proteica representa un alto consumo de energía. Se ha establecido que en *E. coli* la biosíntesis de aminoácidos representa el 40% del gasto energético de la bacteria (sólo el triptófano representa el 1.25%). Por otra parte, utilizar una proteína para más de una función podría generar el denominado conflicto adaptativo (dificultad de optimizar ambas funciones en el contexto de una sola proteína). Normalmente este conflicto se resuelve duplicando el gen y evolucionando por separado y este proceso ha dado lugar a superfamilias de enzimas, transportadores, etc. La pregunta es: ¿por qué las proteínas moonlighting no han separado las dos, o más, funciones en diferentes genes/proteínas? ¿O representan un estadio intermedio en vías de su separación? Esto último es poco probable dado el alto nivel de conservación de secuencia que suelen mostrar la mayoría de ellas. Por otra parte ya se ha descrito (por ejemplo para la chaperona GroEL) que en muchos casos el número de mutaciones que conlleva la segunda función es mínimo y no interfieren con la función canónica. De hecho, los resultados de la ingeniería de proteínas demuestran que la mayor parte de las mutaciones resultan ser neutras (como ya sugirió Kimura en 1985), y además existen las mutaciones compensatorias, que incluso nos han sugerido un método de predecir las regiones responsables de la multifuncionalidad, como se ha mostrado en el Apartado IV.B.5. sobre correlación mutacional de aminoácidos. Finalmente, también se ha mencionado anteriormente que la multifuncionalidad puede ser una manera de conectar diferentes rutas metabólicas a través de una sola proteína lo cual favorecería la conservación de la misma.

Respecto a como aparece la segunda función no se sabe pero ya se ha mencionado anteriormente el caso de la chaperona GroEL en la que mutando 4 aminoácidos dan lugar a

una segunda función toxina de insectos. También tendrían lugar casos de Non Orthologous Gene Displacement o de reclutamiento enzimático (aspecto que está siendo analizando por otro doctorando del grupo, L. Franco). Y en otros casos las segundas funciones provienen de la fusión de genes o dominios de proteínas preexistentes, dando lugar a una proteína multifuncional (Gancedo and Flores, 2008). En nuestro caso nos planteamos mapar las 2 o más funciones en la secuencia/estructura de las proteínas de la base de datos. Dado que en muy pocos casos los autores lo especificaban en sus publicaciones hicimos un mailing masivo preguntándolo pero tan sólo unos pocos respondieron diciendo que lo habían determinado. Por ello otro doctorando (L. Franco) está mapeando las funciones mediante programas de superposición de estructuras 3D de proteínas (p.e., mediante PISITE). Al no ser objeto de la presente tesis no se detallan los resultados, pero pueden encontarse en dos publicaciones recientes (Hernández et al, 2014 y 2015).

Respecto a la conservación filogenética de la multifuncionalidad no se conoce apenas dado que habría que realizar experimentación en cada organismo que se sospeche. En el apartado IV.E., y a partir de multialineamientos, etc, hemos propuesto que hay bastante conservación, en algunos casos altísima (más del 90% de identidad) como se ha mostrado en las Figuras 36 y 37. Por otra parte, en los casos de proteínas de la glicólisis o del ciclo de Krebs relacionadas con virulencia hay numerosos casos conocidos que presentan ambas funciones en diferentes microorganismos patógenos. Por ejemplo la enolasa es moonlight en 44 especies de microorganismos y la GADPH en 33 especies (Tabla 1 y Material suplementario S1 y S2). Asimismo en la Figura 37 puede verse como la proteína ribosomal S10 de *E. coli* presenta conservados los dos motifs relacionados con las dos funciones en otras especies en las que no se ha determinado si existe la bifuncionalidad; y el más conservado es el de la función moonlighting (unión a NusB). Todo ello sugiere que la multifuncionalidad se conservaría filogenéticamente, aunque la demostración definitiva requeriría su determinación experimental.

Ya se ha descrito en la Introducción que probablemente las proteínas moonlighting abunden más de lo que se cree. Nosotros creemos que como dice C. Jeffery (2004) “Current moonlighting appear to be only the tip of the iceberg”. Recientemente a partir de la topología de los grafos de interactómica de proteínas el grupo de Brun (Chapple et al., 2015a) ha predicho 340 proteínas moonlighting humanas. Una pregunta que cabe plantearse es la de que si las proteínas multifuncionales abundan ¿por qué son tan difíciles de encontrar? Ya se

ha mencionado anteriormente que la función canónica de las proteínas multifuncionales en muchos casos suele corresponder con una actividad del metabolismo primario. Pero la función moonlighting no, o por lo menos es más compleja y evolutivamente posterior, de hecho suelen ser "nuevas" e "inesperadas". Si se miran los pares de clases funcionales de la Tabla 2 se puede observar que en 12 de los 26 pares (marcados con un asterisco) la función moonlighting corresponde a funciones incluso de nivel superior al descriptor *GO: Biological Process*. Son funciones muchas de ellas relacionadas con sistema, fisiología, organismo, etc que corresponden a mecanismos complejos, con muchos componentes, o en el caso de ser factores de transcripción intervienen en procesos regulatorios complejos y evolutivamente avanzados. O por ejemplo, en el caso de microorganismos, mecanismos de formación de biofilms, supervivencia en condiciones anormales del medio, infección, etc, que no se encuentran en los medios de cultivo "normales". O sea, en muchos casos las funciones moonlighting corresponden a funciones de fácil identificación funcional. En este sentido cabe mencionar que los microorganismos modelo de célula/genoma mínimos, los micoplasmas, presentan, dependiendo de la especie, entre un 25-42% de sus genes sin función conocida (huérfanos). Y actualmente TreEMBL o UniProt contienen unos 47 millones de secuencias de proteínas para buscar por alineamiento, anotación transitiva, etc. O sea, determinar la función o funciones de una proteína no es fácil. Finalmente, un análisis reciente realizado para otra tesis del grupo (L. Franco) ha encontrado que el 76% de las proteínas moonlighting humanas están involucradas en enfermedades, de acuerdo con la base de datos de las mismas OMIM (Hamosh et al., 2005). Y que el 47% de los fármacos existentes tienen como diana una proteína moonlighting. Estos datos sugieren fuertemente que las proteínas moonlighting no son una excepción sino que deben de ser muy abundantes.

Finalmente describimos nuestra sugerencia de **Reglas para predecir bioinformáticamente si una proteína puede ser moonlighting:**

1.- Ejecutar el algoritmo PsiBlast en una base de datos no redundante, por defecto con 5 iteraciones. Comprobar cuidadosamente al menos las 20-30 primeras posiciones de la lista de resultados. Alternativamente o adicionalmente a PSIBLAST ejecutar el programa HMMER.

2.- Buscar proteínas asociadas a la proteína candidata en las bases de datos de interactómica no curadas. Aquellas proteínas que muestran similares partners de

interactómica en diferentes especies se refuerzan. Realizar un enriquecimiento mediante programas como GOstat (o alternativamente, el que se realiza desde el servidor GO).

3.- La superposición de los resultados de PSIBLAST y de interactómica (en la actualidad se tiene que realizar de forma manual) refuerzan la predicción y minimizan los falsos positivos. La inspección manual es más precisa porque el investigador puede identificar mejor relaciones ocultas para un sistema automatizado.

4.- Ejecutar una búsqueda de motivos y dominios funcionales (PROSITE, PFAM A y B, ProDom, etc.)., por ejemplo, mediante la base de datos InterPro. También mediante la base de datos Blocks.

5.- La localización celular -experimental o por predicción- pueden ayudar a predecir la posibilidad de ser una proteína moonlighting. La presencia de hélices transmembrana y otras señales de tráfico celular son muy informativas. La correspondencia entre los motifs/dominios y las localizaciones celulares predichas pueden reforzar la identificación de la multifuncionalidad.

6.- Aunque no es uno de los objetivos de la presente tesis, en el caso de que la secuencia de la candidata a proteína moonlighting presente homología de secuencia con proteínas para las que existan estructura 3D, utilizar programas de análisis de estructura y modelado como Pisite y Phyre. Si además hay un gran número de secuencias alineables por programas de alineamiento múliple se pueden utilizar programas de correlación de mutaciones como Mistic.

7.- Otros datos procedentes de diferentes fuentes experimentales son de gran ayuda. Por ejemplo, los resultados inesperados de los experimentos de noqueo, efectos secundarios de medicamentos, análisis de co-expresión génica, etc..

Y finalmente volver a recordar la frase de Koonin ya indicada en la Introducción acerca de la predicción bioinformática de la función de una proteína “Annotation is not a routine activity. On the contrary, this is exciting research, somewhat akin to detective work, which has the potential of teasing out deep mysteries of life from genome sequences” (Koonin & Galperin, 2003).

CONCLUSIONES

1.- Se ha creado la primera Base de Datos, MultitaskProtDB, de proteínas moonlighting o multifuncionales. El análisis de las características de las proteínas de la base de datos indica que el par de clases funcionales (canónica/moonlighting) más numerosos es enzima/factor de transcripción.

2.- Se ha comprobado que las bases de datos de interactómica (PPI) contienen información, en muchos casos "partners" previamente considerados como falsos positivos, que permite identificar proteínas multifuncionales.

3.- Se ha determinado que los programas de homología remota como PsiBlast identifican proteínas multifuncionales, especialmente aquellas en que la función está localizada en dominios estructurales diferentes.

4.- Desde el punto de vista de la predicción bioinformática la major estrategia es combinar el análisis de homología remota con los resultados existentes en bases de datos de interactómica (PPI) poco "curadas". Un problema importante es que la anotación funcional existente en las bases de datos de secuencias contiene numerosas anotaciones de baja calidad, cuando no erróneas debido a la denominada "catástrofe de la anotación transitiva". El desarrollo de la anotación de acuerdo a reglas semánticas como Gene Ontology (GO) permitirá mejorar este tipo de problemas.

5.- Los programas de identificación de motifs o de dominios pueden ayudar a predecir las proteínas multifuncionales. Los que mejor los identifican son ProDom, Blocks y PfamB. Al igual que en el caso anterior la sucesiva "curación" de los motifs/dominios para mejorar la predicción de la función canónica ha ido en detrimento de la identificación de la función moonlighting. Por ello, programas menos "curados", como en el caso de PfamB, o por estar descontinuados, como en el caso de Blocks, pueden ser mejores predictores.

6.- Los programas que predicen localización celular con diversas puntuaciones de acuerdo con las diferentes sublocalizaciones celulares pueden, en algunos casos, ayudar en la predicción pero lo hacen mejor para la localización de la función canónica que para la moonlighting. Los programas de predicción de helices transmembrana son mucho mejores

pero en muchos casos las proteínas moonlighting asociadas a membrana o secretadas carecen de tales estructuras.

7.- La predicción de Proteínas (o Regiones) Intrinsecamente Desestructuradas (IDPs o IDRs) indica que las proteínas moonlighting no presentan una mayor tendencia a pertenecer a la clase IDP/IDR que el resto de las las proteínas. En los casos en que presentan regiones desestructuradas éstas suelen corresponder con los grandes lazos de las proteínas.

8.- Desde el punto de vista de la conservación de secuencia se ha determinado mediante alineamiento múltiple que las proteínas moonlighting están altamente conservadas entre diferentes especies, a veces alejadas filogenéticamente. Aunque en esto puede influir el hecho de que muchas funciones canónicas corresponden a funciones del metabolismo primario, que suelen estar conservadas filogenéticamente, el hecho sugiere que la función moonlighting también estará conservada. Sin embargo esto no puede demostrase bioinformáticamente y requiere la corroboración experimental.

9.- Las proteínas moonlighting está muy relacionadas con la patogenicidad de los microorganismos (el 20% de las de la base de datos) y con patologías humanas (el 76% de las proteínas moonlighting humanas corresponden a enfermedades genéticas descritas en la base de datos OMIM). De las dianas farmacéuticas conocidas, un 47% corresponden con proteínas moonlighting.

10.- Conclusión General: La Boinformática puede, hasta cierto punto, predecir las proteínas multifuncionales a partir de la secuencia de aminoácidos de las mismas. Sin embargo su principal papel es el de sugerir o contribuir a corroborar una hipótesis o explicar unos resultados inesperados. Pero la demostración de la multifuncionalidad requiere del análisis experimental complementario.

BIBLIOGRAFÍA

- Aloy P., Cedano J., Oliva B., Avilés F.X. & Querol E. (1997). TransMem: a neural network implemented in Excel spreadsheets for predicting transmembrane domains in proteins. *Comput. Appl. Biosci.,* 13:231–234

- Altschul S.F., Madden T.L., Shaffer A.A., Zhang J., Zhang Z., Miller W. & Lipman D.J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* 25: 3389-3402.

- Amblee V. & Jeffery C. J. (2015). Physical features of intracellular proteins that moonlight on the cell surface. *PLOSOne*. Doi:10.1371/journal.pone.0130575

- Amitai G., Gupta R.D. & Tawfik D.S (2007) Latent evolutionary potentials under the neutral mutational drift o fan enzyme. HSFP J. 1: 67-78

- Ashburner M., Ball C., Blake J., Botstein D., Butler H., Cherry J., Davis A., Dolinski K., Dwight S., Eppig J., Harris M.A., Hill D.P., Issel-Tarver L., Kasarskis A., Lewis S., Matese J.C., Richardson J.E., Ringwald M., Rubin G.M. & Sherlock G. (2000). Gene ontology: tool for the unification of biology. *The Gene Ontology Consortium. Nat Genet* 2000, 25: 25–34

- Backert S., Feller S.M., & Wessler S. (2008). Emerging roles of Abl family tyrosine kinases in microbial pathogenesis. *Trends Biochem. Sci.* 33: 80–90

- Balasubramanian S., Kannan T. R., Hart P. J. & Baseman J. B. (2009). Amino acid changes in elongation factor tu of *Mycoplasma pneumoniae* and *Mycoplasma genitalium* influence fibronectin binding. *Infect. Immun.*, 77: 3533-3541

- Baltes N., Hennig-Pauka I., Jacobsen I., Achim D., Gruber A.D., Gerlach G.F. (2003). Identification of Dimethyl Sulfoxide Reductase in *Actinobacillus pleuropneumoniae* and Its Role in Infection. *Infect. Immun*. 73: 6784–6792

- Baltes N. & Gerlach G.F. (2004). Identification of genes transcribed by Actinobacillus pleuropneumoniae in necrotic porcine lung tissue by using selective capture of transcribed sequences. *Infect. Immun.* 72:6711–6716

- Becker E., Robisson B., Chapple Ch. E. Guénoche A. & Brun Ch. (2012). Multifunctional proteins revealed by overlapping clustering in protein interaction network. *Bioinformatics*, 28, 84-90

- Benton B.M., Zhang J.P., Bond S., Pope C.,Todd C., Lee L., Winterberg K.M., Schmid M.B. & Buysse J.M. (2004). Large-Scale Identification of Genes Required for Full Virulence of *Staphylococcus aureus*. *J. Bacteriol*. 186: 8478-8489

- Beissbarth T. & Speed T. P. (2004). GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics*: 20, 1464–1465.

- Bork P. & Koonin E.V. (1998). Predicting functions from protein sequences- where are the bottlenecks? *Nat. Genet.*, 18: 313-318

- Boucher D.J., Adler B. & Boyce J.D. (2005). The *Pasteurella multocida nrfE* gene is upregulated during infection and is essential for nitrite reduction but not for virulence. *J. Bacteriol.* 187: 2278-2285

- Butler G.S. & Overall C.M. (2009). Proteomic identification of multitasking proteins in unexpected locations complicates drug targeting. *Nat. Rev. Drug Discov.*, 8:935-948

- Cedano J., Aloy P., Pérez-Pons J. A. & Querol E. Relation between amino acid composition and cellular location of proteins.*J. Mol. Biol* ., 266, 594-600 (1997)

- Chapple Ch. E., Robisson B., Spinelli L., Guien C., Becker E. & Brun C. (2015a). Extreme multifunctional proteins identified from a human protein interaction network. *Nature Commun*. 6:7412. doi: 10.1038/ncomms8412

- Chapple Ch. Herrmann C. & Brun C. (2015b). PrOnto database: GO term functional dissimilarity inferred from biological data. *Frontiers Genet*. Doi: 10.3389/fgene.2015.00200

- Cooper D.N. & Krawczak M. (1998). The human gene mutation database. *Nucleic Acids Res* 26: 285-287

- Copley S.D. (2003). Moonlighting functions and catalytic promiscuity. *Curr. Opin. Chem. Biol*., 7: 265-272

- Copley S.D. (2012). Moonlighting is mainstream: Paradigm adjustment required. *Bioessays*. 34: 578-588

- Cvekl A. & Piatigorsky J. (1996). Lens development and crystallin gee expresión: many roles for Pax-6. *Bioassays* 18: 621-630

- Deslandes V., Nash J.H.E., Harel J., Coulton J.W. & Jacques M. (2007). Transcriptional profiling of *Actinobacillus pleuropneumoniae* under iron-restricted conditions. *BMC Genomics*, 13: 8-72

- Dinkel H., Van Roey K., Michael S., Davey N.E., Weatheritt R. J., Born D., Speck T., Kruüger D., Grebnev G., Kuban M., Strumillo M., Uyar B., Budd A., Altenberg B., Seiler M., Chemes L.B., Glavina J., Sánchez I.E., Diella F. & Gibson T.J. (2014). The eukaryotic linear motif resource ELM: 10 years and counting. *Nucleic Acids Res.*, 42: D259-D266

- Dosztanyi Z., Csizmok V., Tompa P. & Simon I. (2005). IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, 21:3433-3434

- Dyson HJ (2011) Expanding the proteome: disordered and alternatively folded proteins. Quaterly Rev Biophys 44: 467-518

- Edelstein P.H., Martha A. C. Edelstein, Futoshi Higa & Stanley Falkow. (1999). Discovery of virulence genes of Legionella pneumophila by using signature tagged mutagenesis in a guinea pig pneumonia model. *Proc. Natl. Acad. Sci. USA, Microbiology*, 96, 8190-8195

- Espadaler J, Eswar N, Querol E, Aviles FX, Sali A, Martı-Renom M, Oliva B (2008)

Prediction of enzyme function by combining sequence similarity and protein interactions. BMC Bioinformatics 9:249

- Finn R.D., Clements J & Eddy S. R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res*., 39: W29-W37

- Fuller T.E., Kennedy M.J., Lowery D.E. (2000a). Identification of Pasteurella multocida virulence genes in a septicemic mouse model using signature-tagged mutagenesis. *Microb. Pathogen.* 29: 25-38

- Fuller T.E., Martin S., Teel J.F., Alaniz G.R., Kennedy M.J. & Lowery D.E. (2000b). Identification of *Actinobacillus pleuropneumoniae* virulence genes using signature-tagged mutagenesis in a swine infection model. *Microbial pathogenesis*, 29: 39-51

- Galperin M Y, Walker D. R. & Koonin E. V. (1998). Analogous enzymes: Independent inventions in enzyme evolution. *Genome Res.* 8: 779-790

- Gancedo C. & Flores C.-L. M. (2008). Moonlighting proteins in yeast. *Microbiol. Mol. Biol. Reviews*, 72, 197-210

- Gilsdorf J.R., Marrs C.F. & Foxman B. (2004). *Haemophilus influenzae*: Genetic Variability and Natural Selection To Identify Virulence Factors. *Infection and Immunity*, 72, 2457-2461

- Gómez A., Domedel N., Cedano J., Piñol J. & Querol E. (2003). Do current sequence analysis algorithms disclose multifunctional (moonlighting) proteins?. *Bioinformatics*, 19, 895-896

- Gómez A., Cedano J., Espadaler J., Hermoso A., Piñol J. & Querol E. (2008). Prediction of protein function improving sequence remote alignment search by a fuzzy logic algorithm. *Protein J.*, 27, 130-139

- Gómez A., Hernández S., Amela I., Piñol J. Cedano J. & Querol E. (2011). Do proteína-protein interaction databases identify moonlighting proteins? *Mol. BioSystems*. 7: 2379-2382

- Gotz, S., Garcia-Gomez, J. M., Terol, J., Williams, T. D., Nagaraj S. H., Nueda, M. J., Roblews M., Talon M., Dopazo J. & Conesa A. (2008). High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res.* 36, 3420–3435

- Hamosh A., Scott A. F.,Amberger. S., Bocchini C. A. & McKusick V. A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*., 33: 514-517

- Hara M.R., Agrawal N., Kim S.F., Cascio M.B., Fujimuro M., Ozeki Y., Takahashi M., Cheah J.H., Tankou S.K., Hester L.D., Ferris C.D., Hayward S.D., Snyder S.H. & Sawa A. (2005). S-nitrosylated GAPDH initiates apoptotic cell death by nuclear translocation following Siah1 binding. *Nat. Cell. Biol.* 7: 665–674

- Harper M., Boyce J.D., Wilkie I.W. and Adler B. (2003). Signature-Tagged Mutagenesis of *Pasteurella multocida* Identifies Mutants Displaying Differential Virulence Characteristics in Mice and Chickens., *Infect. and Immun.,* 71, 5440–5446

- Hedegaard J., Skovgaard K., Mortensen S., Sørensen P., Jensen T.K., Hornshøj H., Bendixen C. & Heegaard P.M.H. (2007). Molecular characterisation of the early response in pigs to experimental infection with *Actinobacillus pleuropneumoniae* using cDNA microarrays. *Acta veterinaria scandinavica*, 49:11

- Heger, A., Wilton, C. A., Sivakumar, A., & Holm, L. (2005). ADDA: a domaindatabase with global coverage of the protein universe. *Nucleic Acids Res.* 33: D188–D191

- Henderson B. & Martin A. (2011). Bacterial virulence in the moonlight: Multitasking bacterial moonlighting proteins are virulence determinants in infectious disease. *Infect. and Immun.,*79: 3476-3491

- Henderson B. & Martin A. Bacterial Moonlighting Proteins and Bacterial Virulence. (2013a). *Curr. Top Microbiol. Immunol.*, 358:155-213

- Henderson B., Fares M-A. & Lund P.A. (2013b). Chaperonin 60: a paradoxical, evolutionary conserved protein family with multiple moonlighting functions. *Biol. Rev. Camb. Philos. Soc*., 88: 955-987

- Henderson B. & Martin A. (2014). Protein moonlighting: a new factor in biology and medicine. *Biochem. Soc. Trans*., 42: 1671-1678

- Henikoff S., Henikoff J. G. & Pietrokoski S. (1999). Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations. *Bioinformatics*, 15: 471-479

- Hensel M., Shea J.E., Gleeson C., Jones M.D., Dalton E & Holden D.W. (1995). Simultaneous identification of bacterial virulence genes by negative selection. *Science* 269:400–403

- Herbert M. A., Hayes S., Deadman M. E., Tang C. M., Hood D. W. & Moxon E. R. (2002). Signature Tagged Mutagenesis of *Haemophilus influenzae* identifies genes required for in vivo survival. *Microbial Pathogenesis*, 33, 211-223

- Hernández S., Amela I., Cedano J., Piñol J., Perez-Pons J.A., Mozo-Villarias A. & Querol E. (2012). Do moonlighting proteins belong to the intrinsic disordered proteins class? *J. Proteom. Bioinf.,* 5: 262-264

- Hernández S., Ferragut G., Amela I., Cedano J., Perez-Pons J.A., Piñol J., Mozo-Villarias A. J. Cedano & Querol E. (2014a). MultitaskProtDB: a database of multitasking proteins. *Nucleic Acids Res.,* D517-D520

- Hernández S., Calvo A., Ferragut G., Franco L., Amela I., Gómez A., Querol E. & Cedano J. (2014b). Can Bioinformatics help in the identification of moonlighting proteins? *Biochem. Soc. Trans.*, 42, 1692-1697

- Hernández S., Calvo A., Ferragut G., Franco L., Amela I., Gómez A., Querol E. & Cedano J. (2015). Bioinformatics and moonlighting proteins (2015). *Frontiers Bioengineer. Biotechnol*. doi: 10.3389/fbioe.2015.00090

- Higurashi M., Ishida T. & Kinoshita K. (2009). PiSite: a database of protein interaction sites

using multiple binding states in the PDB. *Nucleic Acids Res.* 37: D360-D364

- Hodgetts A., Bossé J.T., Simon Kroll J. and Langford P.R. (2004). Analysis of differential protein expression in *Actinobacillus pleuropneumoniae* by Surface Enhanced Laser Desorption Ionisation—ProteinChipTM (SELDI) technology. *Veterinary Microbiology* 99: 215–225

- Hot D. Antoine R., Renauld-Mongenie, Caro V., Hennuy B., Levillain E., Huot L., Wittmann G., Poncet D., Jacob-Dubuisson F., Guyard C., Rimlinger F., Aujame L., Godfroid E., Guiso N., Quentin-Millet, M.J., Y. Lemoine Y. & Locht C. (2003). Differential modulation of *Bordetella pertussis* virulence genes as evidenced by DNA microarray análisis. *Mol. Gen. Genomics*, 269, 475–486

- Hu S., Xie Z., Onishi A., Yu X., Jiang L., Lin J., Rho H. ., Woodard C., Wang H., Jeong J. S. **et al** (2009). Profiling the human protein-DNA interactome reveals EKR2 as a transcripcional repressor of interferon signaling. *Cell*, 139: 610-622

- Huberts D. & van der Kiel I. (2010). Moonlighting proteins: An intriguing mode of multitasking. *Biochim. Biophys. Acta*, 1803: 520-525

- Hunt M.L., Boucher D.J., Boyce J.D. & Adler B. (2001). In vivo-expreseed genes of *Pasteurella multocida*. *Infection and immunity*,69 , 3004-3012

- Ishida T. & Kinoshita K. (2007). PrDOS: prediction of disordered protein regions from amino acid sequence. *Nucleic Acids Res.*, 35: W460-W464

- Jacobsen I., Hennig-Pauka I., Baltes N.,Trost M. & Gerlach G-F. ((2005a).) Enzymes involved in anaerobic respiration appear to play a role in *Actinobacillus pleuropneumoniae* virulence. *Infection and immunity*, 73 , 226-234

- Jacobsen I., Gerstenberg J., Gruber A.D., Bossé J.T., Langford P.R., Hennig-Pauka I., Meens J. & Gerlach G-F. (2005b). Deletion of the ferric uptake regulator fur impairs the in vitro growth and virulence of *Actinobacillus pleuropneumoniae*. *Infection and immunity*, 73 , 3740-3744

- Jacobsen I., Meens J., Baltes N., & Gerlach G-F. (2005c). Differential expression of non-cytoplasmic Actinobacillus pleuropneumoniae proteins induced by addition of bronchoalveolar lavage fluid. *Veterinary microbiology*, 109: 245-256

- Jeffery, C. J. (1999). Moonlighting proteins.*Trends Biochem. Sci.*, 24: 8-11.

- Jeffery, C. J. (2003). Moonlighting proteins: old proteins learning new tricks.*Trends Genet.*, 19: 415-417.

- Jeffery, C. J. (2004). Molecular mechanisms for multitasking: recent crystal structures of moonlighting proteins.*Curr. Opin Struct. Biol.*, 14: 663-668.

- Jeffery, C. J. (2009). Moonlighting proteins- an update. *Mol . BioSyst.,* 5, 345-350

- Jeffery, C. J. (2011). Proteins with neomorphic moonlighting functions in disease. IUBMB Life, 63:489-494.

- Jeffery, C. J. (2013). New ideas on protein moonlighting. In: Moonlighting cell stress proteins in microbial infections. Edited B. Henderson. *Springer, London*. Mol . pp. 51-66.

- Jeffery, C.J. (2014). An introduction to protein moonlighting. *Biochem. Soc. Trans.,* 42**,** 1679-1683

- Jeffery, C.J. (2015). Protein species and moonlighting proteins: very small changes in a protein's covalent structure can change its biochemical function. *J Proteomics*, doi.org/10.10167j.jprot.2015.10.003

- Jensen L.J. & Bork P. (2008). Biochemistry. Not compatible but complementary. *Science*, 322: 56-57

- Jenner R.G. & Young R.A. (2005). Insights into host responses against pathogens from transcriptional profiling. *Nat. Rev. Microb.*, 3, 281-294

- Karlyshev A.V., Oyston P. C. F., Williams K., Clark G. C., Titball R. W., Winzeler E. A. & Wren B. W. (2001). Application of High-Density Array-Based Signature-Tagged Mutagenesis To Discover Novel Yersinia Virulence-Associated Genes. *Infection and Immunity*, 69, 7810–7819

- Karplus P.A. & Schulz GE (1985). Prediction of chain flexibility in proteins. *Naturwissenchaften*, 72:212–213

- Kelley, L.A., & Sternberg, M.J. (2009). Protein structure prediction on the Web: a case study using the Phyre server. *Nat. Protoc.,* 4**,** 363-371

- Keskin O. & Nussinov R. (2007). Similar binding sites and different partners: Implications to shared proteins in cellular pathways. *Structure*, 15: 341-354

- Khan I.K., Chitale M., Rayon C. & Kihara D. (2012). Evaluation of function predictions by PFP, ESG, and PSI-BLAST for moonlighting proteins. *BMC Proceedings*, 6(Suppl 7):S5

- Khan, I., Chen, Y., Dong, T., Hong, X., Tekeuchi, R., Mori, H. & Kihara, D. (2014a). Genome-scale identification and characterization of moonlighting proteins. *Biol. Direct.,* 9, 30

- Khan, I. & Kihara, D. (2014b). Computational characterization of moonlighting proteins. *Biochem. Soc. Trans.,* 42, 1780-1785

- Kim M.S., Pinto S.M., Getnet D., Nirujogi R.S., Manda S.S., Chaerkady R., Madugundu A.K., Kelkar D.S., Isserlin R., Jain S., Thomas J.K., Muthusamy B., Leal-Rojas P., Kumar P., Sahasrabuddhe N.A., Balakrishnan L., Advani J., George B., Renuse S., Selvan L.D., Patil A.H., Nanjappa V., Radhakrishnan A., Prasad S., Subbannayya T., Raju R., Kumar M., Sreenivasamurthy S.K., Marimuthu A., Sathe G.J., Chavan S., Datta K.K., Subbannayya Y., Sahu A., Yelamanchi S.D., Jayaram S., Rajagopalan P., Sharma J., Murthy K.R., Syed N., Goel R., Khan A.A., Ahmad S., Dey G., Mudgal K., Chatterjee A., Huang T.C., Zhong J., Wu X., Shaw P.G., Freed D., Zahari M.S., Mukherjee K.K., Shankar S., Mahadevan A., Lam H., Mitchell C.J., Shankar S.K., Satishchandra P., Schroeder J.T., Sirdeshmukh R., Maitra A., Leach S.D., Drake C.G., Halushka M.K., Prasad T.S., Hruban R.H., Kerr C.L., Bader G.D., Iacobuzio-Donahue C.A., Gowda H. (2014). A draft map of the human proteome. *Nature*, 509: 575-581

- Kimura M. (1983). The neutral theory of molecular evolution. *Cambridge University Press*, Cambridge.

- Koonin E.V. and Galperin M.V. (2003). Sequence-Evolution-Function: Computational Approaches in Comparative Genomics. *Ed. Kluwer*, pp. 225

- Kriston L. McGary, Tae Joo Park, John O. Woods, Hye Ji Cha, John B. Wallingford & Edward M. Marcotte (2010). Systematic discovery of nonobvious human disease models through orthologous phenotypes. *Proc. Natl. Acad. Sci. USA*, 103: 6544-6549

- Kuhner, S., V. van Noort, M. J. Betts, A. Leo-Macias, C. Batisse, M. Rode, T. Yamada, T. Maier, S. Bader, P. Beltran-Alvarez, D. Castano-Diez, W. H. Chen, D. Devos, M. Guell, T. Norambuena, I. Racke, V. Rybin, A. Schmidt, E. Yus, R. Aebersold, R. Herrmann, B. Bottcher, A. S. Frangakis, R. B. Russell, L. Serrano, P. Bork & A. C. Gavin, (2009). Proteome Organization in a Genome-Reduced Bacterium. *Science,* 326: 1235-1240

- Kyte J & Doolittle RF (1982). A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* 157:105–132

- Lehoux D.E., Sanschagrin F. & Levesque R.C. (2002). Identification of in vivo essential genes from *Pseudomonas aeruginosa* by PCR-based signature-tagged mutagenesis. *FEMS Microbiology Letters*, 210: 73-80

- Linding R., Jensen L.J., Diella F., Bork P., Gibson T.J. & Russell R.B . (2003). Protein disorder prediction: implications for structural proteomics. *Structure*,11, 1453-59

- Luo X., Hsiao H.H., Bubunenko M., Weber G., Court D.L., Gottesman M.E., Urlaub H., Wahl M.C. (2008). Structural and functional analysis of the E. coli NusB-S10 transcription antitermination complex. Mol Cell 32: 791-˘802

- Ma B, Kumar S, Tsai CJ, Nussinov R (1999) Folding funnels and binding mechanisms. Protein Eng 12: 713-˘720

- Mahan M.J., Slauch J.M. & Mekalanos J.J. (1993). Selection of bacterial virulence genes that are specifically induced in host tisues. *Science* 259:686–68857

- Mahony J.B. (2002). *Chlamydiae* host cell interactions revealed using DNA microarrays. *Annals New York Academy of Sciences*, 975: 192-201

- Mani M., Chen, C., Amblee V., Liu H., Mathur T., Zwicke G., Zabad S., Patel B., Thakkar J. & Jeffery C.J.(2015). MoonProt: a database for proteins that are known to moonlight. *Nucleic Acids Res*., 43, D277-282

- Marra A., Asundi J., Bartilson M., Lawson S., Fang F., Christine J., Wiesner C., Brigham D., Schneider W.P., & Hromockyj A.E. (2002). Differential Fluorescence Induction Analysis of *Streptococcus pneumoniae* Identifies Genes Involved in Pathogenesis. *Infection and Immunity*, 70, 1422-1433

- Mi T., Merlin J.C., Deverasetty S., Gryk M.R., Bill T.J., Brooks A.W., Lee L.Y., Rathnayake V., Ross Ch. A., Sargeant D.P., Strong Ch. L., Watts P., Rajasekaran S. & Schiller M. R.

(2012). Minimotif Miner 3.0: database expansion and significantly improved reduction of false-positive predictions from consensus sequences. *Nucleic Acids Res.,* 40: D252-D260

- Mitchell A., Chang H.Y., Daugherty L., Fraser M., Hunter S., Lopez R., McAnulla C., McMenamin C., Nuka G., Pesseat S., Sangrador-Vegas A., Scheremetjew M., Rato C., Yong S.Y., Bateman A., Punta M., Attwood T.K., Sigrist C.J., Redaschi N., Rivoire C., Xenarios I., Kahn D., Guyot D., Bork P., Letunic I, Gough J, Oates M, Haft D, Huang H, Natale DA, Wu CH, Orengo C, Sillitoe I., Mi H., Thomas P.D. & Finn R.D. (2015). The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Research,* 43:D213-221

- Moser R.J., Reverter A., Kerr C.A., Beh K. J. & Lehnert S. A. (2004). A mixed-model approach for the analysis of cDNA microarray gene expression data from extreme-performing pigs after infection with *Actinobacillus pleuropneumoniae*. *American Society of Animal Science*, 82:1261–1271

- Nakai, K., and Horton, P. (1999). PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. Trends Biochem Sci 24, 34-36

- Nobeli I., Favia A.D., Thornton J.M. (2009). Protein promiscuity and its implications for biotechnology. *Nat. Biotechnol*. 27:157-167

- Oldfield C.J., Cheng Y., Cortese M.S., Brown C.J., Uversky V.N. & Dunker A.K. (2005). Comparing and combining predictors of mostly disordered proteins. *Biochemistry*, 44: 1989-2000

- Omelchenko M.V, Galperin M.Y. Wolf Y.I. & Koonin E.V. (2010). Non-homologous isofunctional enzymes. A systematic analysis of alternative Solutions in enzye evolution. *Biol. Direct.,* 5: 31-51

- Orihuela C.J., Radin J.N., Sublett J.E., Gao G., Kaushal D. & Tuomanen E.I. (2004). Microarray analysis of pneumococcal gene expression during invasive disease. *Infection and immunity*, 72, 5582-5596

- Ozimek, P. Kotter, M. Veenhuis & I.J. van der Klei (2006). Hansenula polymorpha and Saccharomyces cerevisiae Pex5p's recognize different, independent peroxisomal targeting signals in alcohol oxidase, *FEBS Lett.,* 580: 46–50

- Palmqvist N., Foster T., Tarkowski A. & Josefsson E. (2002). Protein A is a virulence factor in *Staphylococcus aureus* arthritis and septic death. Microbial *Pathogenesis*; 33, 239-249

- Paustian M.L., May B.J., Cao D., Boley D. and Kapur V. (2002). Transcriptional Response of *Pasteurella multocida* to Defined Iron Sources. J. *Bacteriol.,* 183, 6714–6720

- Piatigorsky Joran (2007). Gene Sharing and Evolution. The Diversity of Protein Function. *Harward University Press, London*

- Polesky A.H., Julianna T. D. Ross, Stanley Falkow, & Lucy S. Tompkins. (2001). Identification of Legionella pneumophila Genes Important for Infection of Amoebas by Signature-Tagged Mutagenesis. *Infection and Immunity*, 69: 977-987

- Prieto C. & de la Rivas J. (2006). APID: Agile Protein Interaction DataAnalizer. *Nucleic Acids Res.*, 2006, 34, W298-302

- Qin C., Zhang C., Zhu F., Xu F., Chen S.Y., Zhang P., Li Y.H., Yang S.Y., Wei Y.Q., Tao L., Chen Y.Z., (2014). Therapeutic target database update 2014: a resource for targeted therapeutics. Nucleic Acids Res., 42: D1118-1123

- Sheehan B. J., Bosse J.T., Beddek A. J., Rycroft A. N., Kroll J. S. & Paul R. Langford P. R. (2003). Identification of *Actinobacillus pleuropneumoniae* Genes Important for Survival during Infection in Its Natural Host. *Infection and Immunity,* 71, 3960–397

- Simonetti F.L., Teppa E., Chernomoretz A., Nielsen M. & Marino Buslje C. (2013). MISTIC: mutual information server to infer coevolution. *Nucleic Acids Res*., 41, W8–W14

- Sirover M.A. (2014). Structural analysis of glyceraldehyde-3-P-deshydrogenase functional diversity. *Internat. J. Biochem & Cell Biol.,* 57: 20-26

- Spears P.A., Temple L.M., Miyamoto D.M., Maskell D.J. & Orndorff P.E. (2003). Unexpected similarities between *Bordetella avium* and other pathogenic Bordetellae. *Infection and immunity*, 71, 2591-2597

- Sriram G., Martinez JA, McCabe ER, Liao J.C. & Dipple KM. (2005). Single-gene disorders: what role could moonlighting enzymes play?. *A. J. Human Genet.* 76: 911-924

- The UniProt Consortium. (2013). Update on activities at the universal protein resource (UniProt) in 2013. *Nucleic Acids Res.,* 41, D43–D47

- Tompa P., Szasz C & Buday L. (2005). Structural disorder throws new light on moonlighting. *Trends Biochem. Sci*., 30, 484-489

- Tristan C., Shahani N., Sedlak T.W., Sawa A. (2011). The diverse functions of GAPDH: views from different subcellular compartments. *Cell Signal,* 23: 317–23

- Tsai CJ, Ma B, Nussinov R (1999) Folding and binding cascades: shifts in energy landscapes. Proc Natl Acad Sci USA 96: 9970-9972

- Tsai CJ, Ma B, Sham YY, Kumar S, Nussinov R (2001) Structured disorder and conformational selection. Proteins 44: 418-427

- Tsai C. J., Ma B. & R. Nussinov R. (2009). Protein-protein interaction networks: how can a hub protein bind so many different partners? *Trends Biochem. Scien.*, 34, 594–600

- Tuinstra R.L., Peterson F.C., Kutlesa S., Elgin E.S., Kron M.A. & Volkman BF (2008). Interconversion between two unrelated protein folds in the lymphotactin native state. *Proc. Natl. Acad. Sci. USA*, 105: 5057–62

- Tungteakkhun S.S. & Duerksen-Hughes P.J. (2008). Cellular binding partners of the human papillomavirus E6 protein. *Arch. Virol.* 153: 397–408

- Uhlén M., Fagerberg L., Hallström B.M., Lindskog C., Oksvold P., Mardinoglu A., Sivertsson Å., Kampf C., Sjöstedt E., Asplund A., Olsson I., Edlund K., Lundberg E., Navani S., Szigyarto C.A., Odeberg J., Djureinovic D., Takanen J.O., Hober S., Alm T., Edqvist P.H., Berling H., Tegel H., Mulder J., Rockberg J., Nilsson P., Schwenk J.M., Hamsten M., von Feilitzen K., Forsberg M., Persson L., Johansson F., Zwahlen M., von Heijne G., Nielsen J., Pontén F. (2015). Tissue-based map of the human proteome. *Science*, 347: (6220):1260419. doi: 10.1126/*science*.1260419

- Valdivia R.H. & Falkow S. (1997). Fluorescence-based isolation of bacterial genes expressed within host cells. *Science,* 277:2007-2011

- Wagner T.K. & Mulks M.H. (2007). Identification of the *Actinobacillus pleuropneumoniae* Leucine-Responsive Regulatory Protein and Its Involvement in the Regulation of In Vivo-Induced Genes. *Infection and Immunity*, 75, 91–103

- Wang J., Mushegiant A., Loryt S. & Jin S. (1996). Large-scale isolation of candidate virulence genes of *Pseudomonas aeruginosa* by in vivo selection. *Proc. Natl. Acad. Sci. USA*, 93, 10434-10439

- Ward J.J., Sodhi J.S., McGuffin L.J., Buxton B.F. & Jones D.T. (2004). Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.*, 337, 635-645

- Wilhelm M., Schlegl J., Hahne H., Moghaddas Gholami A., Lieberenz M., Savitski M.M., Ziegler E., Butzmann L., Gessulat S., Marx H., Mathieson T., Lemeer S., Schnatbaum K., Reimer U., Wenschuh H., Mollenhauer M., Slotta-Huspenina J., Boese J.H., Bantscheff M., Gerstmair A., Faerber F., Kuster B. (2014). Mass-spectrometry-based draft of the human proteome. *Nature*, 509: 582-587

- Wishart D.S., Knox C., Guo A.C., Cheng D., Shrivastava S., Tzur D., Gautam B & Hassanali M. (2008). DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res*., 36:D901-916

- Wistow G.J. & Piatigorsky J. (1987). Recruitment of enzymes as lens structural proteins. *Science* 236: 1554-1556

- Woods D. E. (2004). Comparative genomic analysis of *Pseudomonas aeruginosa* virulence. *Trends Microbiol.,*12, 437-439

- Wool, I. G. (1996). Extraribosomal functions of ribosomal proteins. *Trends Biochem*., 21: 164-165.

- Yoshida N., K. Oeda, E. Watanabe, T. Mikami, Y. Fukita, K. Nishimura, K. Komai & K. Matsuda (2001). Protein function. Chaperonin turned insect toxin, *Nature*, 411 44

PUBLICACIONES A QUE HA DADO LUGAR EL PRESENTE TRABAJO

- **Hernández** S., Gómez A., Cedano J. & Querol E. (2009). Bioinformatics annotation of the hypothetical proteins found by omics techniques in genomes of respiratory pathogens can help to disclose additional virulence factors. *Current Microbiol*. 59, 451-456

- **Hernández** S., Gómez A., Cedano J. & Querol E. (2009). Omics techniques and identification of pathogen virulence genes. Application to the analysis of respiratory pathogens. *J. Comput. Sci & Sys. Biol*., 2, 124-132

- Gómez A., **Hernández** S., Amela I., Piñol J. Cedano J. & Querol E. (2011). Do protein-protein interaction databases identify moonlighting proteins? *Mol. BioSystems.*, 7, 2379-2382

- **Hernández** S., Amela I., Cedano J., Piñol J., Perez-Pons J.A., Mozo-Villarias A. and Querol E. (2012). Do moonlighting proteins belong to the intrinsic disordered proteins class? *J. Proteom. & Bioinf.,* 5, 262-264

- **Hernández** S., Ferragut G., Amela I., Pérez-Pons J.A., Piñol J., Mozo-Villarias A., Cedano J. & Querol E. (2014). MULTITASKPROTDB: A Database of multitasking proteins. *Nucleic Acids Res*., 42, D517-520

- **Hernández** S., Calvo A., Ferragut G., Franco L., Amela I., Gómez A., Querol E. & Cedano J. (2014). Can Bioinformatics help in the identification of moonlighting proteins? *Biochem. Soc. Transact.*, 42, 1692-1697

- **Hernández** S., Calvo A., Ferragut G., Franco L., Amela I., Gómez A., Querol E. & Cedano J. (2015). Bioinformatics and moonlighting proteins. *Frontiers Bioengineer. Biotechnol*. 3: doi:10.3389/fbioe.2015.00090

- Franco L., **Hernández** S., Cedano J., Pérez-Pons J.A., Piñol J., Mozo-Villarias A.Amela I. & Querol E. Moonlighting proteins: involvement in human diseases and targets of current drugs. Enviado

AGRADECIMIENTOS

Desearía agradecer la paciencia y ayuda de Enrique Querol y de todo el grupo (Juan Cedano, Isaac Amela, Antonio Gómez) todos estos años de intentar combinar el desarrollo de una tesis y un trabajo.

Agradezco a mis padres también su apoyo todos estos años.

Gracias también a mis amigos y compañeros de trabajo que han aguantado con paciencia el tiempo invertido en esta tesis doctoral y que siempre me han dado un gran apoyo y comprensión.