Universitat Autònoma de Barcelona

PROTEÍNAS “MOONLIGHTING”: IDENTIFICACIÓN Y RELACIÓN CON LA INFECCIÓN POR MICROORGANISMOS PATÓGENOS Y CLÍNICA HUMANA

Luis Franco Serrano

PROTEÍNAS “MOONLIGHTING”: IDENTIFICACIÓN Y RELACIÓN CON LA INFECCIÓN POR MICROORGANISMOS PATÓGENOS Y CLÍNICA HUMANA

Luis Franco Serrano
UNIVERSITAT AUTÒNOMA DE BARCELONA
2018

PROTEÍNAS “MOONLIGHTING”: IDENTIFICACIÓN Y RELACIÓN CON LA INFECCIÓN POR MICROORGANISMOS PATÓGENOS Y CLÍNICA HUMANA

Tesis doctoral presentada por Luis Franco Serrano, Licenciado en Biología, especialidad en Biología Sanitaria, para optar al título de Doctor en Inmunología Avanzada por la Universitat Autònoma de Barcelona, tutorizada por el doctor Iñaki Àlvarez

Este trabajo ha sido realizado en el Institut de Biotecnologia i Biomedicina de la Universitat Autònoma de Barcelona bajo la dirección de los doctores Enrique Querol Murillo y Isaac Amela Abellan

2018

Luis Franco Serrano

Dr. Enrique Querol **Dr. Isaac Amela**

“The important thing in science is not so much to obtain new data, but to discover new ways of thinking about them.”

William Lawrence Bragg

“Science is built up of facts, as a house is with stones. But a collection of facts is no more a science than a heap of stones is a house.”

Henri Poincare

AGRADECIMIENTOS

A Enrique e Isaac por su paciencia durante estos años de tesis, por la sabiduría que me han transmitido, por la alegría con la que lo han hecho y por ayudarme a avanzar en mi carrera profesional pese a las dificultades. A Juan Cedano por sus ideas y consejos y a Óscar Conchillo por mantener vivo a Wallace.

A mi madre y mi hermano, por haber mantenido la familia incluso en las circunstancias más desfavorables además de por comprender mi falta de tiempo libre.

A mis amigos y amigas más cercanos, por ser mi punto de apoyo en los momentos difíciles, por haberme animado a continuar todos estos años y por haberme hecho vivir momentos increíbles, ya sea viajando, en la montaña, sentados en una terraza o jugando a algun que otro juego de mesa. Siempre os deveré a todos una sonrisa.

Si algo he aprendido de vosotros en todo este tiempo es que lo primero son personas y después todo lo demás.

ÍNDICE

INFORMACIÓN SUPLEMENTARIA

S1: Listado completo de proteínas moonlighting y toda la información contenida en la base de datos MultitaskProtDB-II.

S2: Prediccion de proteínas moonlighting utilizando PPIs + PsiBlast.

S3: Listado completo de folds presentes en las proteínas moonlighting y estadísticas de estos.

S4: Alineamientos de todas las proteínas encontradas por PiSite que encajan con las funciones moonlighting de las proteínas presentes en MultitaskProtDB.

S5: Relación de proteínas moonlighting y las enfermedades humanas en que están implicadas.

S6: Relación de proteínas moonlighting que son dianas farmacológicas y de los fármacos implicados.

S7: Proteínas moonlighting implicadas en la virulencia de microorganismos.

S8: Alineamientos de proteínas moonlighting implicadas en la virulencia de microorganismos patógenos frente a su homóloga humana con los epítopos B señalados para ambas.

S9: Listado de motifs presentes en enolasas de microorganismos patógenos que no estan en no patógenos.

S10: Microorganismos en los que estan presentes los motifs analizados de enolasas de microorganismos patógenos.

RESUMEN

Moonlighting (multitasking, multifunctional) es la capacidad de algunas proteínas de ejecutar dos o más funciones bioquímicas. Normalmente, las proteínas moonlighting son descubiertas por "serendipia". La multifuncionalidad de las proteínas adquiere una mayor importancia a partir de los recientes descubrimientos acerca del proteoma humano (y de animales modelo como el ratón). El mecanismo de “splicing” no da lugar a un gran número de proteínas como se pensaba hasta ahora, sinó que los genes humanos mayoritariamente expresan la denominada isoforma principal, a nivel de proteína. Por ello el moonlighting puede contribuir, o ser la clave, a la complejidad de la célula. Por esta razón, sería de utilidad que la bioinformática pudiera predecir la multifuncionalidad, especialmente debido a la gran cantidad de nuevas secuencias provenientes de los proyectos genómicos. Durante el presente trabajo se han analizado y descrito diferentes aproximaciones que utilizan secuencias, estructuras, interactómica, algoritmos bioinformáticos e información contenida en las bases de datos como UniProt, OMIM, etc, para tratar de contribuir a la identificación de las proteínas multifuncionales. Un primer objetivo de este trabajo ha sido la actualización de la base de datos de proteínas moonlighting, MultitaskProtDB-II (http://wallace.uab.es/multitaskII), que ha servido como banco de trabajo para todos los análisis realizados. Por ejemplo, hemos intentado mapar los dominios funcionales de cada función dentro de la estructura de algunas proteínas importantes de nuestra base de datos. Esto ha permitido, en un cierto número de casos, identificar los dominios relacionados con una o más enfermedades humanas basadas en la proteína multifuncional analizada. De MultitaskProtDB-II, se ha encontrado que un porcentaje significativo de proteínas moonlighting (78%) están relacionadas con enfermedades humanas y que un 48% de ellas son dianas para fármacos actuales. Además, un 25% de proteínas moonlighting de la base de datos actualizada están implicadas en la virulencia de microorganismos patógenos, y hemos propuesto una explicación basada en el hecho de que las proteínas moonlighting, al ser evolutivamente muy conservadas, el huésped evitaría desarrollar una respuesta inmune que podría desencadenar una enfermedad autoinmune. Esto tiene gran importancia en la selección de proteínas candidatas

a ser inmunogénicas protectivas en estudios de vacunología reversa. Desde el punto de vista de la evolución de las proteínas moonlighting y a partir del análisis de identificadores GO y de la ortología de interactomas, se sugiere que la segunda función estaría conservada filogenéticamente y que habría muchas más proteínas multifuncionales de lo que se pensaba anteriormente. Finalmente se ha analizado la posible relación evolutiva entre la multifuncionalidad y el desplazamiento de gen no ortólogo.

RESUM

Moonlighting (multitasking, multifunctional) és la capacitat d'algunes proteïnes d'executar dos o més funcions bioquímiques. Normalment, les proteïnes moonlighting són descobertes per casualitat. La multifuncionalitat de les proteïnes adquireix una major importància a partir dels recents descobriments sobre el proteoma humà (i d'animals model com el ratolí). El mecanisme de "splicing" no dóna lloc a un gran nombre de proteïnes com es pensava fins ara, sinó que els gens humans majoritàriament expressen l'anomenada isoforma principal, a nivell de proteïna. Per això el moonlighting pot contribuir, o ser la clau, a la complexitat de la cèl·lula. Per aquesta raó, seria d'utilitat que la bioinformàtica pogués predir la multifuncionalitat, especialment a causa de la gran quantitat de noves seqüències provinents dels projectes genòmics. Durant el present treball s'han analitzat i descrit diferents aproximacions que utilitzen seqüències, estructures, interactòmica, algoritmes bioinformàtics i informació continguda en les bases de dades com UniProt, OMIM, etc, per tractar de contribuir a la identificació de les proteïnes multifuncionals. Un primer objectiu d'aquest treball ha estat l'actualització de la base de dades de proteïnes moonlighting, MultitaskProtDB-II (http://wallace.uab.es/multitaskII), que ha servit com a banc de treball per a totes les anàlisis realitzades. Per exemple, hem intentat mapar els dominis funcionals de cada funció dins de l'estructura d'algunes proteïnes importants de la nostra base de dades. Això ha permès, en un cert nombre de casos, identificar els dominis relacionats amb una o més malalties humanes basades en la proteïna multifuncional analitzada. De MultitaskProtDB-II, s'ha trobat que un percentatge significatiu de proteïnes moonlighting (78%) estan relacionades amb malalties humanes i que un 48% d'elles són dianes per a fàrmacs actuals. A més, un 25% de proteïnes moonlighting de la base de dades actualitzada estan implicades en la virulència de microorganismes patògens, i hem proposat una explicació basada en el fet que les proteïnes moonlighting, en ser evolutivament molt conservades, l'hoste evitaria desenvolupar una resposta immune que podria desencadenar una malaltia autoimmune. Això té gran importància en la selecció de proteïnes candidates a ser immunogèniques protectives en estudis de vacunologia reversa. Des del punt de vista de l'evolució de les proteïnes moonlighting i a partir

de l'anàlisi d'identificadors GO i de la ortologia de interactomes, es suggereix que la segona funció estaria conservada filogenèticament i que hi hauria moltes més proteïnes multifuncionals del que es pensava anteriorment. Finalment s'ha analitzat la possible relació evolutiva entre la multifuncionalitat i el desplaçament de gen no ortòleg.

ABSTRACT

Moonlighting (multitasking, multifunctional) is the ability of some proteins to perform two or more biochemical functions. Normally, moonlighting proteins are discovered by "serendipity". The multifunctionality of proteins acquires greater importance from recent discoveries about the human proteome (and animal models such as mice). The splicing mechanism does not give rise to a large number of proteins as was thought until now, but human genes mainly express the so-called main isoform, at the protein level. Therefore, moonlighting can contribute, or be the key, to the complexity of the cell. For this reason, it would be useful to predict multifunctionality from a bioinformatics point of view, especially due to the large number of new sequences coming from the genomic projects. During this work we have analyzed and described different approaches that use sequences, structures, interactomics, bioinformatic algorithms and the information contained in databases such as UniProt, OMIM, etc., to try to contribute to the identification of multifunctional proteins. A first objective of this work was to update the moonlighting protein database, MultitaskProtDB-II (http://wallace.uab.es/multitaskII), which has been used as a work platform for all the analyses. For example, we mapped the functional domains of each function within the structure of the protein and this allowed, in a certain number of cases, the identification of the domains related to one or more human diseases caused by this multifunctional protein. From MultitaskProtDB-II, it has been found that a significant percentage of moonlighting proteins (78%) are related to human diseases and that 48% of them are targets for current drugs. In addition, 25% of moonlighting proteins from the updated database are involved in the virulence of pathogenic microorganisms, and we have proposed an explanation based on the fact that the moonlighting proteins, being evolutionarily very conserved, the host would avoid developing a response immune that could trigger an autoimmune disease. This is of great importance in the selection of candidate proteins in order to be immunogenic and protective in reverse vaccinology studies. From the point of view of the evolution of moonlighting proteins and from the analysis of GO identifiers and the orthology of interactomes, it has been suggested that the second function would be phylogenetically conserved and that there would be many more multifunctional proteins than previously thought. Finally, the possible

evolutionary relationship between multifunctionality and non-orthologous gene displacement has been analysed.

ABREVIATURAS

Blast = Basic Local Alignment Search Tool

EBI = European Bioinformatics Institute

ECM = Extracellular matrix

ESG = Extended Similarity Group

GAPDH = Glyceraldehyde-3-phosphate dehydrogenase

GO = Gene Ontology

GOA = Gene Ontology Annotation

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

MISTIC = Mutual Information Server to Infer Coevolution

NCBI = National Center for Biotechnology Information

NMR = Nuclear Magnetic Ressonance

NOGD = Non-Ortohologous Gene Displacement

OMIM = Base de datos Human Mendelian Inheritance in Man

PDB = Protein Data Bank

Pfam = Protein Families

PFP = Protein Function Prediction

Phyre = Protein Homology/Analogy Recognition Engine

PHP = Hypertext Preprocessor

PiSite = Protein Interaction Sites

PPI = Protein-Protein Interaction

ProDom = Protein Domains

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

ProtLoc = Protein Localization

PSI-Blast = Position Specific Iterated BLAST

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

RMSD = Root Mean Square Deviation

SQL = Structured Query Language

I. INTRODUCCIÓN

I. A. PROTEÍNAS MOONLIGHTING O MULTITASKING

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. Este orden funcional es simplemente un orden histórico, que hace referencia a la fecha de identificación de las funciones, y no implica una relevancia funcional especial. El hecho de adquirir una segunda función no implica una pérdida de la función canónica. El par funcional más común corresponde a ser una enzima o proteína implicada en el metabolismo como función canónica y una función más compleja y de adquisición evolutivamente posterior como función moonlighting. En el caso particular de proteína Aldehyde dehydrogenase queda claro este hecho, ya que incluso los organismos más simples presentan enzimas, pero hasta la aparición en organismos superiores con ojo, esta proteína no adquirió su función moonlighting). Las funciones moonlighting no estan necesariamente separadas en dos dominios proteicos distintos. Cuando se analizan las proteínas multifuncionales, se genera una pregunta fundamental ¿cómo definimos y detectamos una función biológica? La ingeniería genética sugiere que la función proteica es una propiedad absoluta y transferible a cualquier otra célula, organismo o sistema, y esto no es cierto. Porque, la función de una proteína depende del contexto en el que se encuentra (Kriston et al., 2010).

Las proteínas moonlighting son de reciente descubrimiento. Cronológicamente, la primera proteína moonlighting descubierta y publicada fue la Aldehyde dehydrogenase identificada como una proteína estructural en la lente del ojo además de su función enzimática, por Wistow y Piatigorsky en 1987. Posteriormente, la neuroleucina, una citoquina identificada por Chaput et al., en 1988 resultó ser la Phosphoglucose isomerase, reportada con anterioridad. Las proteínas moonlighting suelen descubrirse por serendipia (por casualidad) cuando se identifica un gen/proteína relacionado con alguna función que resulta ser una proteína conocida con otra función, generalmente ancestral y más

básica. Es bastante común que correspondan a una función del metabolismo primario. Las proteínas moonlighting se citan por primera vez en un texto de Bioquímica General, en la última versión de Lehninger (Capítulo 16, página 624) y han sido incorrectamente traducidas al español como "enzimas del claro de luna" ...

Según los últimos datos disponibles, el genoma humano contiene solo unos 20.000 genes. Mecanismos como el splicing alternativo pueden dar lugar a muchas más proteínas, pero la multifuncionalidad agrega más capacidad funcional sin aumentar el número de proteínas. El fenómeno moonlighting también es presente en Procariotas en los cuáles no hay splicing alternativo. Por otra parte, estudios recientes ponen en duda la existencia de los numerosos polipéptidos a los que daría lugar el mecanismo del splicing alternativo como base de la complejidad de la célula (para una revisión ver Tress et al., 2017). Algunos artículos sobre proteínas moonlightings son: Wool, 1996; Jeffery, 1999, 2003, 2004 y 2009; Piatigorsky 2007; Gancedo y Flores, 2008; Nobeli y otros, 2009; Huberts y van der Kiel, 2010; Copley, 2012; Jeffery, 2013, 2014; Henderson y Martin, 2014.

Los siguientes aspectos suelen asociarse a la multifuncionalidad (Figura 1):

- **Localización celular**: Algunas proteínas pueden localizarse en más de un compartimiento celular. Como se comentará en la Sección IV.E., la presencia de una proteína en una región celular en la que no está realizando su función canónica, puede sugerir que es una proteína moonlighting. 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. Otro ejemplo es la Enolasa de *Plasmodium falciparum,* esta se encuentra también en el núcleo o asociada al citoesqueleto, aunque se desconoce su función en esas regiones, su presencia puede sugerir que está desarrollando otra función.

- **Intracelular/secretada**: Algunas proteínas cuya función es claramente intracelular, se pueden encontrar secretadas. Por ejemplo, la Phosphoglucose isomerase, es una enzima de la glucólisis en el citoplasma y también una neuroleuquina (un factor de crecimiento de células nerviosas) cuando es

secretada. Otro ejemplo es la Enolasa de microorganismos patógenos, que actúa como enzima de la glucólisis en el citoplasma y como factor de virulencia por unirse al Plasminógeno del huésped cuando es secretada.

- **Expresión diferencial**: Una misma proteína puede realizar funciones distintas en función del factor que la ha activado. Por ejemplo, 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**: Dependiendo de su estado de oligomerización una misma proteína puede tener distintas funciones. Por ejemplo, la Glyceraldehyde-3-phosphate dehydrogenase (GAPDH), como tetrámero y en el citoplasma, es una 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 la regulación de la transcripción vía unión al terminador de transcripción NusB.

- **Modificación postraduccional**: La Phosphoglucose isomerase fosforilada en la Ser185 no actúa como enzima sinó 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 función 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.

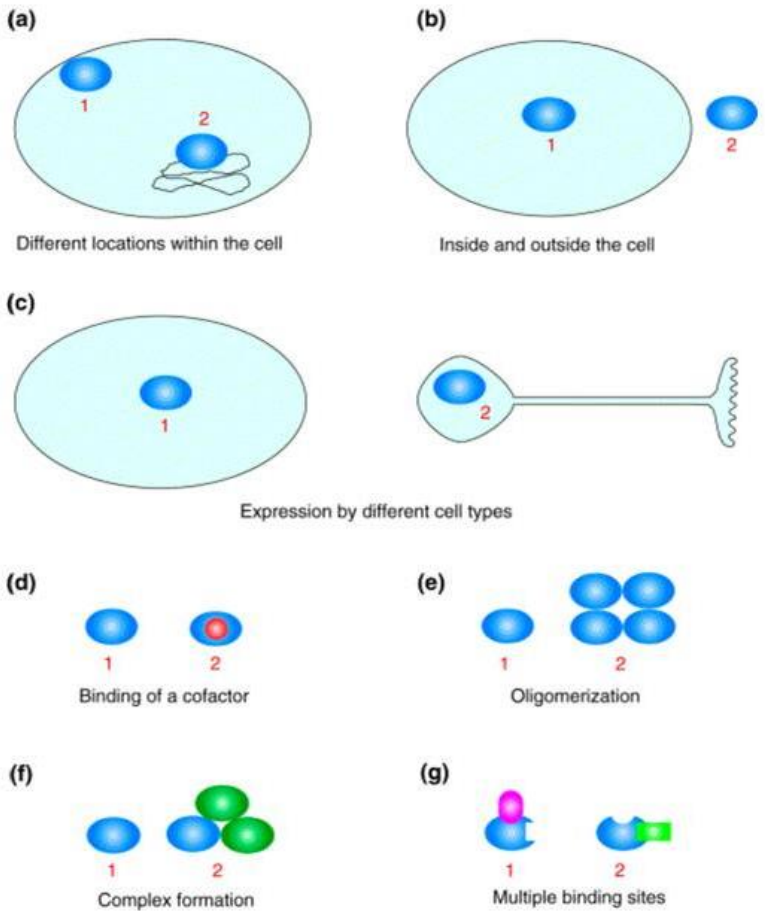


Figura 1: Características con las que se suele asociar el moonllighting.

No se puede universalizar la función moonlighting a más de una especie a no ser que experimentalmente se hayan demostrado las distintas funciones. Se hablará ampliamente sobre este punto en la Sección IV.E.

Históricamente las proteínas moonlighting has recibido diversos nombres. G. Piatigorsky, uno de sus descubridores, introdujo el concepto de "*Gene sharing*" pero debido a que en el splicing alternativo también se comparte un gen, no es un término apropiado. La principal diferencia entre splicing y moonlighting es que el primero da lugar a diferentes polipéptidos mientras que en el moonlighting se

realizan diferentes funciones con un mismo polipéptido. El término *Moonlighting* se debe a Constance Jeffery (Jeffery, 1999), pero ella lo utiliza de forma más restrictiva que el de *Multitasking*. Según Jeffery el término moonlighting estaría restringido a aquellos casos en que la segunda función no es derivada de una fusión génica. Tampoco incluiría dentro del término moonlighting, los casos con dominios claramente diferenciados por provenir probablemente de fusiones génicas. Otro aspecto que Jeffery excluye es presentar dos funciones moleculares 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. Además, desde un punto de vista evolutivo estos aspectos pueden ser muy difíciles de determinar dado que habría que 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 que puedan realizar funciones diferentes a partir del mismo polipéptido, y no tanto que provengan de fusiones de genes o dominios.

El término de *proteínas promiscuas* fue introducido por Noveli (Noveli et al., 2009). Este término se utiliza mucho en proteínas que mediante la misma función molecular intervienen en muchas funciones biológicas. Por ejemplo, una quinasa sería una proteína promiscua pero no sería una proteína moonlighting ni multifuncional puesto que realiza la misma función molecular, fosforilar una proteína diana, en diferentes dianas que a su vez estarán involucradas en diferentes rutas metabólicas. La proteína moonlighting ideal 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).

Los casos que **no se considerarían moonlighting** son los siguientes:

- Variantes de proteínas por el mecanismo del **splicing alternativo** o producto de una proteólisis que dé lugar a fragmentos con diferentes funciones biológicas.

- Aquellas enzimas que presentan una **amplia gama de substratos** (por ejemplo, proteínas como las pertenecientes a la superfamilia Aldehído deshidrogenasas, los citocromos P_{450}). Incluso enzimas muy específicos suelen presentar otras funciones secundarias, "adventicias", pero órdenes de magnitud más ineficientes.

- Los factores de transcripción que pueden **unirse a diferentes promotores**. De esta forma. En función de con que genes interaccionen pueden producir diferentes efectos, pero esencialmente estan desarrollando la misma función molecular.

- Una enzima que participa **en diferentes rutas metabólicas** mediante la misma función molecular (p.e., Ribulose-phosphate 3-epimerase actúa en la ruta de las pentosas fosfato y en el metabolismo del formaldehído).

- Únicamente **para algunos autores** (p.e., C. Jeffery) una **proteína producto de la fusión de 2 genes** (por ejemplo., la HisB de *E. coli* producto de la fusión de la proteína Imidazoleglycerol-phosphate dehydratase y la Histidinolphosphatase) no sería un verdadero caso de moonlighting dado que ambas funciones, aunque por separado, ya preexistían. Sin embargo, otros muchos autores defienden (p.e., J. Thornton, Ch. Brun) que sí lo sería dado que se trata de una multifuncionalidad.

- Estos mismos autores más restrictivos, opinan que las proteínas con dos actividades enzimáticas en un mismo centro activo no son verdaderos casos de moonlighting. Sin embargo, otros como J. Thornton defienden que sí se trata de multifuncionalidad. Esta última autora las considera ejemplo de proteína "promiscuas" (Nobeli et al., 2009). Hay que tener en cuenta que estas segundas funciones enzimáticas promiscuas suelen ser bastante ineficientes, por tanto, sólo afectan al fenotipo si se producen en gran cantidad (Copley et al., 2003). Jeffery ha propuesto que se utilice el término moonlighting en sentido estricto de acuerdo con su definición y multitasking (multitarea o multifuncional…) para el resto, incluyendo las provenientes de fusiones de genes o dominios. Pero muchos autores, como es nuestro caso, seguimos utilizando el término moonlighting para todos estos casos.

SI deberían considerarse verdaderos casos de moonligting

- Los casos en los que la multifincionalidad esté ligada a modificaciones post-traduccionales. Algunos ejemplos son, la GAPDH en la que algunas de sus funciones están relacionadas con nitrosilación, fosforilación y acetilación. La Citrate synthase, que fosforilada es una enzima y defosforilada es un componente estructural de los filamentos citoplasmáticos. Y por útlimo, p53 que fosforilada en los residuos S46+T55 se une al TFα y fosforilada en los residuos S15+S20 bloquea la interacción con la proteína MDM2.
- Las proteínas que ocasionalmente se anclan a la membrana uniéndose a un ácido graso y presentando una función alternativa, también serían verdaderos casos de moonlighting.

La ancestralidad de las enzimas del metabolismo primario (glicólisis, ciclo de Krebs, etc) está muy ligada a la moonlighticidad (Sriram et al., 2005). Estas son enzimas muy conservadas. Las Figuras 2 y 3 muestran que la mayoría de las enzimas de la glicólisis y del ciclo de Krebs son moonlighting, ligadas a una o más funciones. Hasta el momento la enzima Glyceraldehyde-3-P-dehydrogenase (GAPDH) es la proteína con más funciones identificadas, 18, (pero esas 18 funciones lo son en diferentes organismos, es obvio que en *E. coli* no puede estar en el cristalino). En la Figura 4 están representadas todas estas funciones.

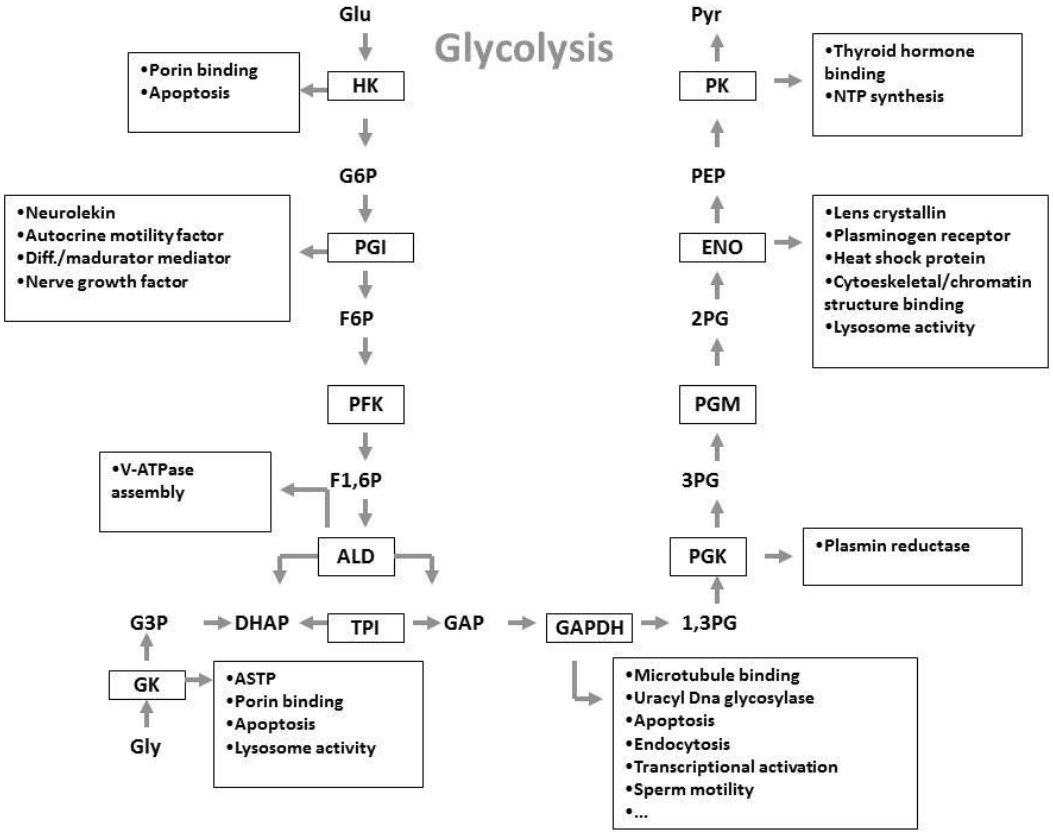


Figura 2: Casi todas las proteínas de la glucólisis son proteínas moonlighting. Esta figura muestra las funciones adicionales que tienen estas proteínas, además de su función canónica en el metabolismo primario.

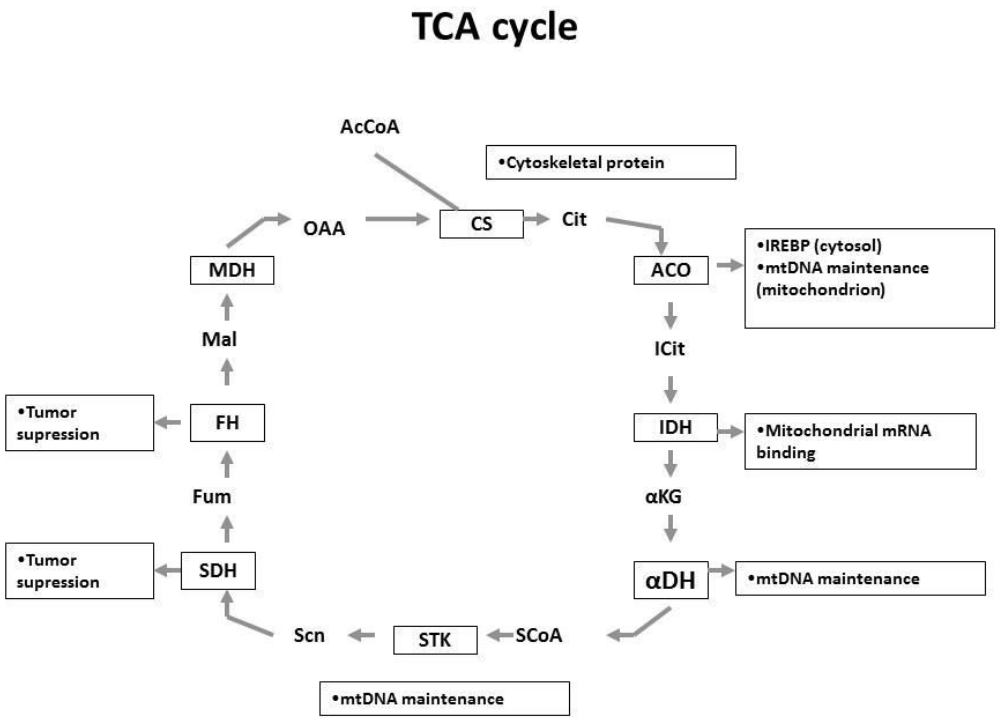


Figura 3: También hay proteínas moonlighting en el ciclo de TCA. Otro ejemplo de proteínas moonlighting en el metabolismo primario

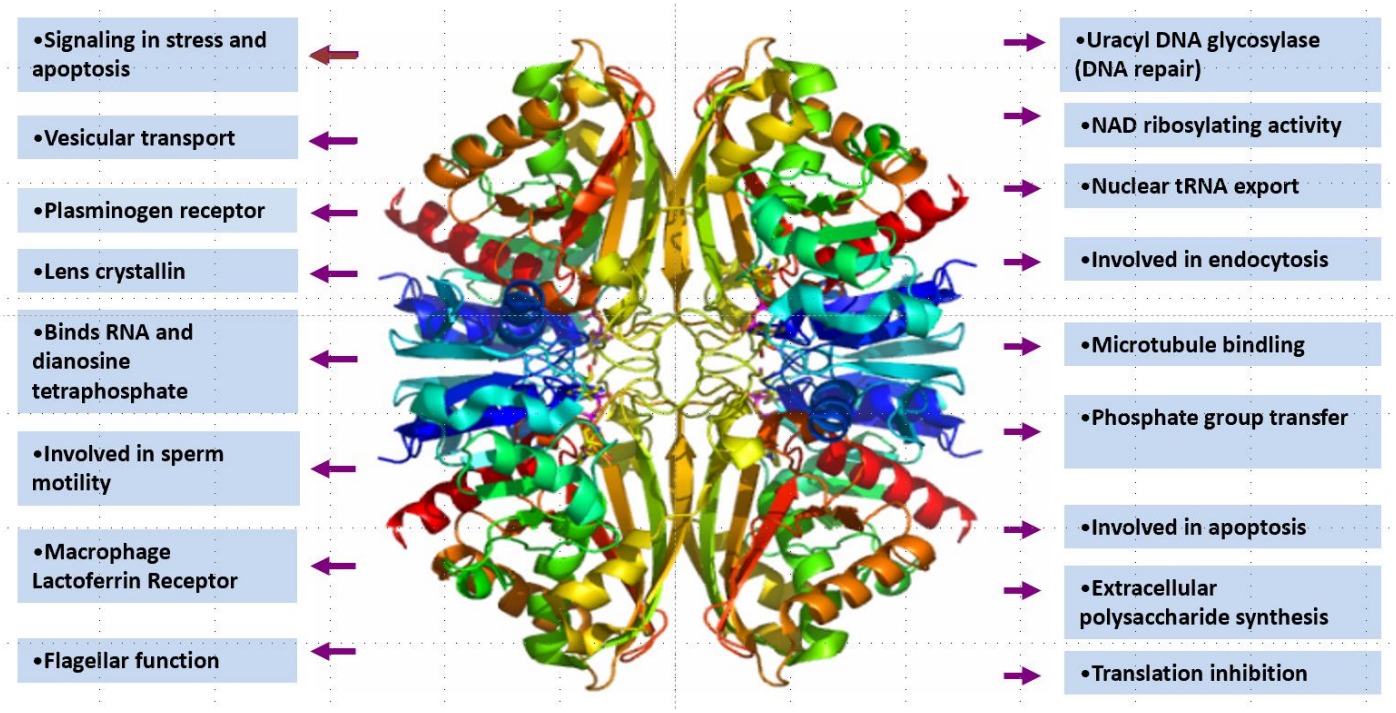


Figura 4: La Glyceraldehyde-3-Phosphate-Dehydrogenase (GAPDH) es la proteína moonlighting con más funciones conocidas en la actualidad, 18 en diferentes organismos.

Probablemente, en poco tiempo, la proteína p53 superará a la GAPDH. Ya que, no solo tiene diferentes funciones descritas, sino que tiene cientos de interacciones con proteínas reguladoras (Figura 5). Otras proteínas con muchas funciones son: Hsp90, Hsp70, HMGB1, Aconitasa, Enolasa, Cpn10, Dihidrolipolamida deshidrogenasa, Ubiquitina y EFTu.

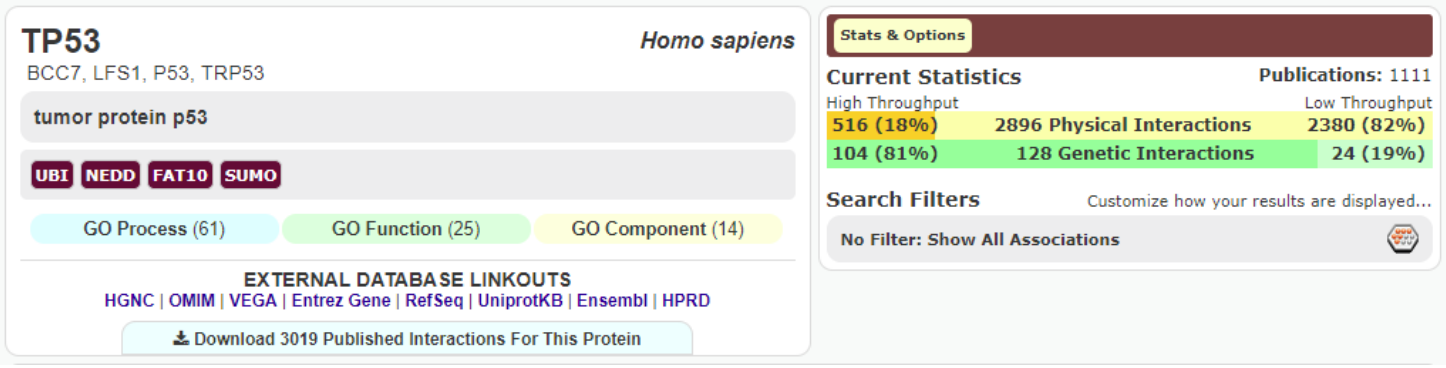


Figura 5: Interacciones proteína – proteína de p53 (según el servidor BioGrid).

En el caso de la proteína más abundante en la sangre, la Albúmina sérica, no solo es un transportador universal, sino que representa el 13% de las proteínas en la córnea, además, su núcleo hidrofóbico es el sitio activo para diferentes reacciones. Se ha propuesto una nueva clasificación de proteínas moonlighting: proteínas moonlighting de una sola función adicional (SAFMP) y proteínas moonlighting de función adicional múltiple (MAFMP) ya que tienen una o múltiples funciones adicionales.

Las proteínas moonlighting también estan presentes en muchos virus, especialmente aquellos con un genoma corto. La mayoría estan involucradas en la evasión del sistema inmune del huésped. Por ejemplo, el virus del papiloma humano tiene solo 8 proteínas. La proteína E5 interactúa con la proteína humana p53 y el complejo se degrada en el proteasoma. Esta proteína viral también está involucrada en la expresión de la proteína humana SET7 que metila p53, esta metilación aumenta la estabilidad de p53. La proteína E6 también interactúa con proteínas relacionadas con la proliferación celular, la apoptosis, la adhesión, la estabilidad cromosómica, el reconocimiento del sistema inmunitario, la polarización celular y la estructura del epitelio (Tungteakkun y Duerksen-Hughes, 2008). Todas estas funciones están relacionadas con las consecuencias de la infección del virus del papiloma humano.

Un aspecto sorprendente es que, diferentes proteínas en diferentes especies, incluso genéticamente no cercanas pueden estar desarrollando la misma función moonlighting. Es el caso, por ejemplo, de las proteínas del cristalino (Piatigorsky, 2007) y algunas enzimas citosólicas en microorganismos patógenos como: GAPDH, Enolasa, Fosfoglucomutasa, Fosfoglicerato quinasa, DnaK, Peroxirredoxina y el factor de elongación Tuf. Todas estas proteínas están involucradas en la adhesión entre el microorganismo patógeno y el huésped (Henderson y Martin, 2011; 2013).

I.B. RELEVANCIA DE LAS PROTEÍNAS MOONLIGHTING EN LA BIOQUÍMICA DE PROTEÍNAS

El fenómeno del moonlighting es de vital importancia en investigación debido fundamentalmente a las siguientes consecuencias:

- **Anotación genómica funcional:** causa imprecisiones o descripción insuficiente.

- **Interpretaciones de experimentos de Knock-outs y Knock-downs**: debido a la moonlighticidad, al bloquear un gen, si este codifica para una proteína moonlighting los efectos pueden deberse a diferentes funciones y no a una sola.

- El mecanismo de **reclutamiento de enzimas** y el desplazamiento de genes **no ortólogos**.

- **Análisis metabólico** y predicción de vías metabólicas, etc.

- **Análisis de redes de interatómica**: en las proteínas moonlighting es más compleja la interpretación de redes de interatómica. Por ejemplo, algunas interacciones relevantes para segundas funciones aún por descubrir podrían determinarse como falsos positivos al no encajar con la función descrita.

- **Acción farmacológica y su mecanismo**: selección de fármacos, efectos secundarios y toxicidad, polifarmacología, etc.

- La existencia de **enfermedades como consecuencia de que una proteína adquiera una función adicional** (denominada "neomorphic moonlighting"). Esta puede ser debida a la hiperactividad de una proteína, por ejemplo, la GAPDH también tiene actividad apoptótica, y la función estaría involucrada en Alzheimer, Parkinson y la isquemia cerebral.

- La **resistencia bacteriana a antibióticos**: por ejemplo, la Glutamate racemasa de *M. tuberculosis* presenta funciones moonlighting de resistencia al ciprofloxacino.

- El uso de proteínas moonlighting, en especial de aquellas del metabolismo primario, por parte de microorganismos patógenos como proteínas de **virulencia**

y adhesión al huésped (por ejemplo, Enolasa y otras proteínas de la glicólisis ...). Estas, ubicadas en la membrana, tienen una segunda función como factor de virulencia para facilitar la unión a las células del huésped.

- Además, el moonlighting llevaría al llamado **"conflicto adaptativo"** (las mutaciones beneficiosas pueden ser favorables para una función y perjudiciales para la otra). En general, los conflictos adaptativos se resuelven mediante la duplicación de genes y la evolución independiente de los parálogos. Esto ha resultado en superfamilias de proteínas (portadores, reguladores de la transcripción ...).

- Finalmente, y como se discutirá en la Sección IV.C., a lo largo del presente trabajo se encontró que las proteínas moonlighting están **relacionadas con enfermedades humanas y son dianas de medicamentos** actuales.

Todo lo anterior hace que el análisis y la predicción bioinformática de las proteínas moonlighting sea un objetivo importante.

I.C. CLASES FUNCIONALES DE LAS PROTEÍNAS MOONLIGHTING

Las proteínas moonlighting se puede agrupar en función de los pares de funciones que presentan. En la base de datos que hemos diseñado (Franco-Serrano et al., 2018), y como se describe en el capítulo IV.B., la combinación más abundante es ser una "enzima" y una "proteína de unión a ácido nucleico" (por ejemplo, actuando como factor de transcripción). La segunda combinación más abundante es ser una "enzima" y estar implicada en la "adhesión celular".

Las proteínas moonlighting presentan otras muchas categorías funcionales, como, por ejemplo, receptores, proteínas del citoesqueleto, chaperonas, etc. En células Eucariotas frecuentemente estas funciones son desarrolladas en diferentes compartimentos subcelulares. También pueden realizar las diferentes funciones al mismo tiempo o en diferentes momentos del ciclo celular.

Sin embargo, pocas proteínas moonlighting son proteínas de membrana, tal vez debido a complicaciones en el proceso de plegamiento cuando se presentan

algunos dominios transmembrana. Sin embargo, en microorganismos patógenos, las proteínas citosólicas o nucleares se transportan a la membrana celular sin péptido señal para participar en la adhesión del microorganismo al huésped. Muchos microorganismos usan la Enolasa para unirse al Plasminógeno del huésped (Henderson y Martin, 2011; 2013). Otro ejemplo es la Histona H1, que actúa como un receptor de Tiroglobulina (Brix et al., 1998). Obviamente, no están utilizando la vía normal de secreción de RE-Golgi a la membrana, pero todavía no se conoce el mecanismo que utiliza. En las células Eucariotas, pueden usar interacciones con parejas específicas o usar modificaciones posteriores a la traducción, como es el caso de GAPDH (Tristan et al., 2011).

Por otro lado, las proteínas multifuncionales pueden contribuir a coordinar diferentes actividades celulares mediante los siguientes mecanismos: facilitando el "cambio" o las conexiones entre diferentes vías metabólicas, proporcionando mecanismos reguladores como la retroalimentación, etc. y también debido a que corresponden a centros clave en las redes metabólicas y de interacción. Ya se ha mencionado anteriormente que muchas enzimas del metabolismo primario presentan actividad moonlighting y, por otro lado, analizando los datos de interactómica se sabe que las proteínas con el mayor número de conexiones son las del metabolismo energético y los mecanismos de traducción.

Analizando los datos publicados en “Protein atlas” (Uhlén et al., 2015) accesible en: http://www.proteinatlas.org/humanproteome/housekeeping se puede observar que un gran número de proteínas moonlighting tienen regulación constitutiva (housekeeping genes), probablemente debido a que muchas son proteínas del metabolismo primario.

Debido al hecho de que la función canónica acostumbra a ser una función más primaria y vital para la célula y además se descubrió antes que las funciones moonlighting se suele considerar, erróneamente, que la función canónica es la principal y la moonlighting es secundaria o accesoria.

I.D. IDENTIFICACIÓN DE LAS PROTEÍNAS MOONLIGHTING

Identificar la función de una proteína es habitualmente difícil y lo es mucho más, identificar una multifuncionalidad. Sin embargo, los siguientes indicadores nos pueden sugerir que una proteína es moonlighting:

- Encontrar una enzima en **otro compartimento** que no sea el correcto para esa proteína. Por ejemplo, la Lactato deshidrogenasa, Fosfoglicerato sintasa, Aldolasa y GAPDH se han encontrado en el núcleo, y se sabe que actúan como factor de transcripción. La Lactato deshidrogenasa se encuentra además en el cristalino. Otra posibilidad, muy común en las proteínas moonlighting es que la proteína se exporte a la membrana o se secrete aún careciendo de los “motifs” apropiados para la secreción, este es el caso de la Enolasa en algunos microorganismos patógenos.
- Hay una **mayor cantidad de la necesaria** para realizar su función canónica (p. Ej., Lactato deshidrogenasa LDHB4 representa el 5% de las proteínas de los oocitos de ratón, y la Albúmina sérica el 13% de las proteínas de la córnea).
- Al noquear un gen, se encuentra un **fenotipo inesperado**.
- Por **interactómíca**: se descubre que la proteína interacciona con partners inesperados (consideramos en adelante “partner” como una proteína con la que interacciona la proteína analizada) y no es una proteína promiscua o “sticky” (sticky, en la terminología de interactómica son aquellas proteínas que presentan muchas interacciones, pero son biológicamente irrelevantes).
- **Bioinformáticamente:** por ejemplo, con la combinación del análisis de dominio/motivos + programas de predicción de homología, especialmente remota, PSI-Blast/ByPass + a partir de la información contenida en las bases de datos de interactómica (PPI), etc. Es parte de los OBJETIVOS de este trabajo.

I.E. BASE ESTRUCTURAL Y EVOLUTIVA DE LAS PROTEÍNAS MOONLIGHTING

Únicamente 243 de un total de 694 proteínas de nuestra base de datos, tienen publicada su estructura 3D. Además, de estas 243, en muy pocas se han decrito las regiones funcionales para las funciones moonlighting. De hecho, 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. Además, como se ha comentado anteriormente, las funciones no tienen porque estar en regiones diferentes de la proteína. En algunos casos, una mínima región de la estructura es la responsable de la multifuncionalidad como se ha descrito en las proteínas: GroEL (Yoshida et al., 2001), la Cpn60 (Henderson et al., 2013) o la GAPDH (Sirover et al., 2014). Aunque en el caso de la GAPDH, que tan sólo tiene 37kDa, las diferentes funciones estan relacionadas con modificaciones post-traduccionales y con el estado de oligomerizacion. En algunos casos, la multifuncionalidad depende de muy pocos aminoácidos. Por ejemplo, la EFTu de *Mycoplasma pneumoniae* y *Mycoplasma genitalium* se diferencian por su interacción con la Fibronectina, pero presentan un 96% de identidad de secuencia. Esto indica que, esa interacción depende de unos pocos aminoácidos (Balasubramanian et al., 2009).

La proteína ribosomal S10, es un ejemplo de que no siempre las funciones moonlighting estan localizadas en regiones distintas de la proteína, en este caso, la función “transcription regulation” prácticamente solapa con la de “rRNA binding” (Figura 6) (Luo et al., 2008).

Los cambios conformacionales también pueden generar casos de moonlighticidad. Por ejemplo, la proteína reguladora de la transcripción p53 que presenta varios lazos involucrados en diferentes interacciones con las proteínas Sirtuin, Cyclin A, CBP y S100bb, y de esta forma presenta funciones diferentes (Olfield et al., 2005). Otro ejemplo es la Linfoactina humana, que en función del estado de oligomerización realiza una función distinta: como monómero y con un fold tipo chemokine realiza la función canónica y como dímero con un fold distinto, se une a glucosaminoglicanos (Tuinstra et al., 2008).

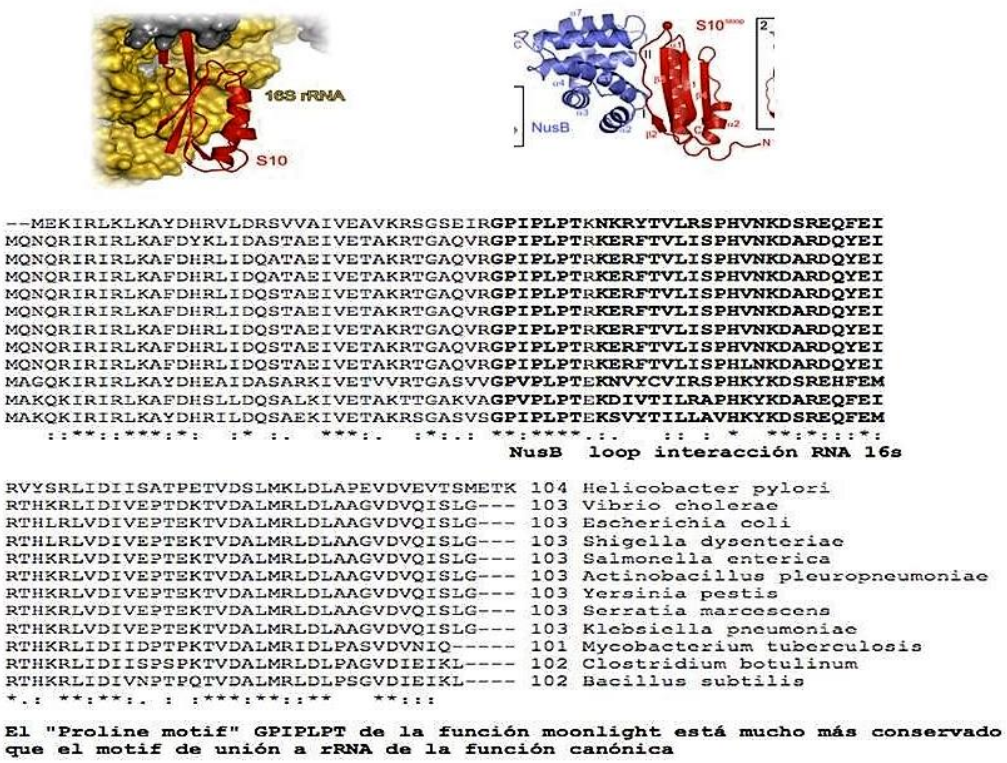


Figura 6: Estructura 3D de la proteína ribosomal S10 de *E. coli* que muestra los dos motifs responsables de sus dos funciones (unión a RNAR 16s y al factor de transcripción NusB). También se muestra el alineamiento con las secuencias ortólogas de diversas especies. Puede verse que el motif de la función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

Algunos autores sugieren que las proteínas moonlighting son proteínas conocidas como IDPs (Intrisically Disordered Proteins), estas proteínas al interaccionar con muchos partners distintos, darían lugar a múltiples funciones (Tompa et al., 2005). Sin embargo, se ha demostrado que en la mayoría de casos las proteínas multifuncionales no pertenecen a la clase de proteínas IDPs (Hernández et al., 2012; Chapple et al., 2015).

Las proteínas moonlighting nos demuestran que la naturaleza reutiliza secuencias y estructuras de proteínas, de esta manera no ha de diseñarlas de nuevo. Además, no se requieren largos periodos de tiempo, para adquirir una segunda función. En muchos casos, por ejemplo, adquirir una segunda función no conlleva más que un cambio en el patrón de expresión, sin la necesidad de cambios en la secuencia, por ejemplo, las proteínas del cristalino del ojo (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, en la la chaperona GroEL de *Enterobacter aerogenes* únicamente se necesitan 4 aminoácidos para que presente una segunda actividad como toxina para insectos. Únicamente se diferencia de la misma proteína en *E. coli* en 11 aminoácidos, de los cuales únicamente 4 son clave para la segunda función (Yoshida et al., 2001). Además, nosotros sugerimos que el Non Orthologous Gene Displacement o el 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 en especies cercanas. Por ejemplo, la Pyruvate carboxylase de algunas especies presentan diferentes funciones: la de *Saccharomyces cerevisiae* no es moonlighting pero presenta un 80% de identidad de secuencia con la de *H. polymorpha* que sí lo es y participa en la translocación del peroxisoma. La de *P. pastoris* también presenta un 80% de identidad de secuencia con *H. polymorpha*, y en este caso sí conserva la función moonlighting (Ozimek et al., 2006).

Una proteína puede ser multifuncional en especies diferentes, pero sin tener la misma función moonlighting. Por ejemplo, la función canónica de la Aconitasa es catalizar la conversión de citrato a isocitrato, pero tiene diferentes funciones moonlighting según la especie, como “Iron-Responsive Element”, mantenimiento y reparación del DNA, etc. Las proteínas del cristalino o las Cpn60 presentan numerosos ejemplos similares. Como describiremos en el apartado de Resultados, creemos que la conservación de secuencia, especialmente de “motifs” funcionales, sugiere que habrá conservación de la multifuncionalidad, pese a que la demostración de la función moonlighting haya de ser experimental.

La interpretación de las redes interactómicas se complica con la existencia de la multifuncionalidad. Esa complejidad dependerá de cuantas proteínas moonlighting haya, pero se han hecho muy pocos estudios sobre esto. En un trabajo se analizaron cuantas proteínas humanas podían ser de unión 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 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 algunos casos. Por ello era necesaria la creación de un base de datos de proteínas multifucionales (ver Apartado IV.A.). El análisis de la base de datos que hemos construido muestra que la multifuncionalidad es un fenómeno presente en todos los reinos de la vida en mayor número del esperado inicialmente (ver Apartado IV.E.).

I.F. INTENTOS PREVIOS DE PREDICCIÓN DE PROTEÍNAS MOONLIGHTING

La mayoría de los avances realizados en este aspecto, a nivel internacional, los ha realizado nuestro grupo de investigación y, posteriormente, los de Kihara y de Brun. 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 moonlighting. 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 PSI-Blast 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 moonlighting conocidas en ese momento, unas 30 y fue publicado (Gómez et al., 2003) (Figura 7). Estos análisis han sido realizados por nuestro grupo para las 288 proteínas de nuestra base de datos inicial (Hernández et al., 2014).

MOONLIGHTING PROTEIN		PSI-BLAST	SAM	PROSITE	BLOCKS	EMOTIF	PRODOM	SMART	PFAM	P-SORT	PROTLOCK	TRANSMEM	TRANSCOUT
a.PtsH protease (*T.thermophilus*) BAA96085.1	+	0.0	+		+	+	+			+(cyt)	+ (ic)	-	-
b.Chaperone activity	+	0.0	+	+	+	+	+	+	+		+ (ic)		
a.Uracyl-DNA-glycosilase (*H.s.*) CAA37794.1	+	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_00836	+	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: Los primeros análisis de homología remota (PSI-Blast y SAM) y de buscadores de motivos y dominios funcionales realizados por nuestro grupo a partir del escaso número de proteínas moonlighting conocidas en aquel momento.

El grupo de investigación de Kihara ha utilizado también el algoritmo PSI-Blast 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 la base de datos Gene Ontology (GO) www.geneontology.org. Por su parte el grupo de investigación de Brun describe 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 de GO restringe mucho las posibles anotaciones comparadas puesto que la

base de datos GO tan sólo tiene incluidas en este momento 44.947 anotaciones funcionales (29.621 “Biological process”, 11.132 “Molecular functions”, 4.194 “Celular components”) y existen muchísimas más funciones. Una dificultad añadida es que las bases de datos de (NCBI, EBI…) no siguen criterios semánticos como GO y la mayoría estan llenas de anotaciones sin ningún tipo de orden. Existen anotaciones tan imprecisas como “17 kilodalton protein”. Esto dificulta la comparación automática de los resultados de programas de homología remota y de interactómica.

Otra aproximación la ha realizado nuestro grupo mediante la información contenida en las bases de datos de interactómica de proteínas (PPIs). En estas bases de datos podemos encontrar partners de interacción que no encajen con la función canónica y nos sugieran nuevas funciones de la proteína Figura 8 (Gómez et al., 2011). Además, combinando los partners de interacción con homología remota hemos visto que aumenta la predicción bioinformática de la multifuncionalidad de una proteína.

Tabla 1: Los primeros análisis de bases de datos de interactómica, realizados por nuestro grupo, con objeto de determinar si en muchos de los partners considerados falsos positivos eran identificables las funciones moonlighting de proteínas multifuncionales conocidas

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

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

Inicialmente, en la investigación llevada a cabo por nuestro grupo, se han utilizado como ejemplos las proteínas moonlighting descritas en la bibliografia (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). En estas revisiones se mencionó que se conocían 80-100 proteínas moonlighting, pero se presentaron tablas con solo 20 o 30. Luego, diseñamos la primera base de datos de proteínas moonlighting con 288 proteínas (Hernandez et al., 2014). En el presente trabajo, hemos recopilado más proteínas moonlighting de la bibliografía, aproximadamente 700, y creado una base de datos actualizada de ellas (http://wallace.uab.es/multitaskII/), (Franco-Serrano et al., 2018), que corresponde al Objetivo 1 de esta tesis. Esta base de datos actualizada también contiene más información para cada proteína, como la predicción estructural, la relación con enfermedades humanas o si son dianas de fármacos actuales, este punto se analizará ampliamente en la Sección IV.C. Desde que tuvimos esta base de datos, se han hecho mejores aproximaciones en la predicción, evolución, etc., de las proteínas moonlighting. Después de la publicación de nuestra base de datos, el grupo de Jeffery ha publicado la suya (Jeffery et al., 2018), que es más incompleta y más complicada de usar (por ejemplo, hay menos enlaces o no hay enlaces a las secuencias, UniProt, etc.)

I.H. RELACIÓN ENTRE PROTEÍNAS MOONLIGHTING Y VIRULENCIA DE MICROORGANISMOS PATÓGENOS.

Como se ha descrito anteriormente, muchas proteínas del metabolismo primario (Enolasa, GAPHD ...) de diversos microorganismos patógenos se utilizan como factores de virulencia, por ejemplo, para la adhesión al huésped (Henderson y Martin, 2011; 2013). Pero, además, existe el mecanismo de moonlighting forzoso ("forced moonlighting") o el secuestro de proteínas del huésped para forzarlas a una segunda función que facilita la actividad del patógeno. Por ejemplo, *E. coli* recluta la Actina del huésped para facilitar la unión a su epitelio intestinal (Backert et al., 2008). Además, los virus habitualmente utilizan proteínas del huésped para realizar funciones que por ellos mismos no pueden hacer, en contra del propio

huésped. Ser capaz de predecir bioinformáticamente si una proteína es multifuncional sería de gran ayuda para crear vacunas e identificar dianas farmacológicas, lo cual es de gran interés para el grupo ya que una parte importante de su investigación se relaciona con la obtención de vacunas.

Ya se ha descrito que las proteínas del metabolismo primario son muy propensas a presentar funciones moonlighting, probablemente debido a su ancestralidad. Pero otro hecho relacionado con la patogenicidad y la virulencia de los microorganismos es que las diferentes proteínas del metabolismo primario tienen la misma función moonlighting en relación con la virulencia. Esto implica que estas enzimas que no tienen similitud de secuencia comparten alguna característica conformacional, o algún motif no identificado. Entre los objetivos de esta tesis está identificar qué motifs o dominios podrían estar involucrados, y conservados, en esta interacción con una proteína del huésped, esto se discutirá en la Sección IV.D.

Muchos autores, incluido nosotros, hemos intentado identificar regiones comunes en estas proteínas que permitan interactuar con el Plasminógeno humano u otras proteínas del huésped. Muchos grupos lo han hecho también investigando como las células cancerígenas invaden los tejidos para hacer metástasis. Durante este tiempo, incluso hemos contactado personalmente con algunos investigadores que trabajan en este tema, pero sin obtener una conclusión clara.

Se han descrito algunos motifs en *Streptococcus* que participan en la unión al Plasminógeno humano. Estos consisten en dos Lys en los residuos N-terminales de la proteína o dos Lys separadas por otros dos aminoácidos hidrofóbicos (Bergmann et al., 2005; Derbise et al., 2004). Pero ninguno de estos motifs es exclusivo para microorganismos patógenos y muchos microorganismos no patógenos, incluso sin una actividad probada de unión a Plasminógeno, los tienen. Además, la mutación de estas regiones reduce la actividad de unión al Plasminógeno, pero no suprime toda la actividad. Otro hecho a tener en cuenta es que, estos motifs no estan compartidos con otros microorganismos patógenos o con otras proteínas de virulencia como GAPDH, Phosphoglucose mutase o kinase. En la Sección IV.D se discutirá ampliamente este tema. Durante mucho tiempo, se asumió que los residuos de lisina localizados en el extremo C de las

proteínas bacterianas constituyen los únicos sitios de unión al Plasminógeno u otras proteínas del huésped. También se ha identificado que aminoácidos cargados positivamente son potenciales sitios de unión a la Enolasa o la Endopeptidasa neumocócica. Curiosamente, una arginina e histidina, pero no una lisina, representa lo esencial en cuanto a aminoácidos para mediar la interacción proteína-proteína. A pesar de las diferencias entre los motifs de unión al Plasminógeno en distintas proteínas y microorganismos, la presencia de aminoácidos cargados positivamente en un entorno hidrofóbico parece ser el principal requisito para la unión al Plasminógeno del huésped. (Fulde et al., 2013).

Otra cuestión importante es que estas proteínas moonlighting, que de forma natural son citosólicas, se transportan a la membrana celular sin señal de exportación. Por ejemplo, en *Mycobacterium tuberculosis*, un importante sistema de exportación para los factores de virulencia es el complejo Sec. Como se muestra en la Figura 7, el complejo Sec transporta proteínas con una señal de exportación N-terminal, pero también proteínas sin esta señal. Queda por conciliar cómo SecA2 está involucrado en la exportación de proteínas que carecen de péptidos señal en *M. tuberculosis* y proteínas con péptidos señal en *M. smegmatis*. Una posibilidad es que las proteínas exportadas por SecA2 no necesiten el reconocimiento del péptido señal por parte del receptor y tanto las proteínas que contienen péptido señal como las que carecen de este puedan ser reconocidas y exportadas por SecA2. Otra alternativa es que el papel de SecA2 en la secreción de proteínas que carecen de péptido señal sea indirecto. SecA2 puede exportar una proteína actualmente desconocida que contiene un péptido señal. Esta proteína desconocida podría ser parte de un aparato de secreción especializado a través del cual se secretan proteínas como SodA (Feltcher et al., 2010; Costa et al., 2015).

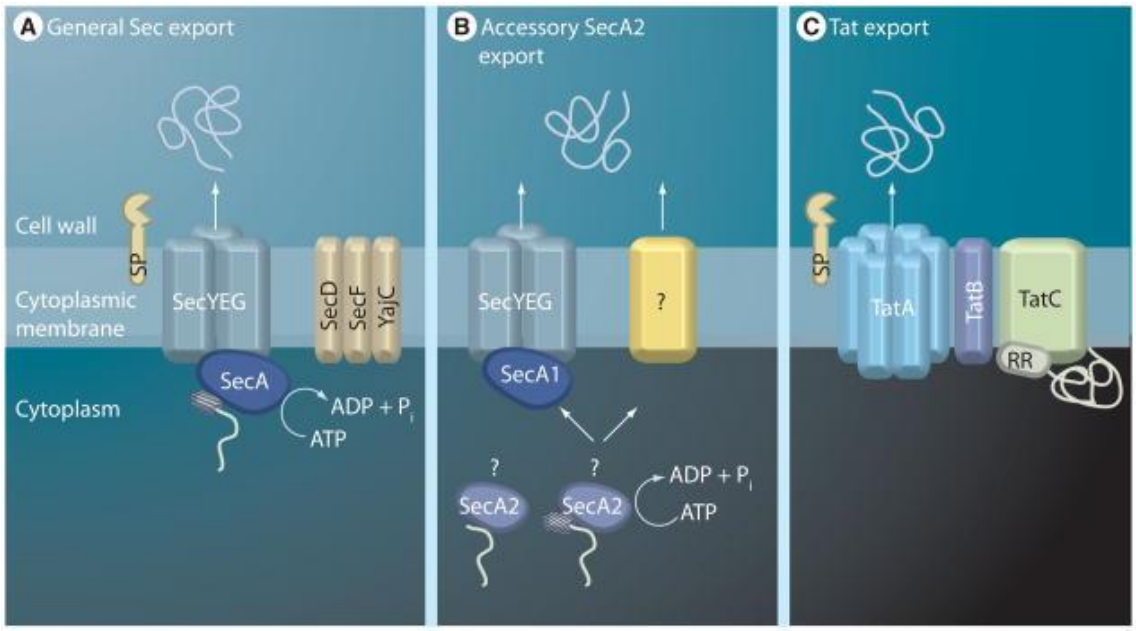


Figure 7: Sistema de exportación de proteínas Sec, SecA2 y Tat en micobacterias.

Otra pregunta importante es: ¿Por qué el Plasminógeno? La Enolasa y otras proteínas moonlighting relacionadas con virulencia interaccionan con el Plasminógeno humano. También con otras proteínas como la Fibronectina, pero principalmente con Plasminógeno. ¿El Plasminógeno está involucrado en esta interacción? Un hecho importante es que la Enolasa humana también interactúa con el Plasminógeno. Entonces, ¿qué característica secuencial o estructural tiene el Plasminógeno? Estas preguntas permanecen sin respuestas. Algunos investigadores dicen que los motifs de interacción del Plasminógeno son "pockets" con una región central hidrofóbica con aminoácidos cargados positivamente alrededor. Pero este es un motif común en muchas enzimas, no es exclusivo del Plasminógeno.

Por otra parte, las proteínas moonlighting de virulencia no solo se unen al Plasminógeno, sinó que lo activan, lo que facilita la degradación del tejido y la invasión del microorganismo patógeno. El Plasminógeno es una serin proteasa, estas son enzimas que escinden enlaces peptídicos en proteínas en las que la serina sirve como el aminoácido nucleofílico en el sitio activo (Figura 8). En humanos, las funciones del Plasminógeno son: digestión, respuesta inmune, coagulación sanguínea, disolución de coágulos de sangre con fibrina. También se sabe que el Plasminógeno participa en la regeneración y reorganización tisular. Por eso, el Plasminógeno también está relacionado con el cáncer, en

especial en los procesos de metástasis. Gran parte de los estudios sobre los activadores del Plasminógeno y el cáncer se guiaron por la hipótesis de que la proteólisis de los componentes de la matriz extracelular, iniciada por la liberación del activador del Plasminógeno de las células cancerosas, desempeña un papel decisivo en la degradación del tejido normal y, por lo tanto, un crecimiento invasivo y consecuentemente, metástasis (Dano et al., 1985). Este mecanismo puede ser el mismo que usan los microorganismos patógenos para degradar los tejidos e invadirlos.

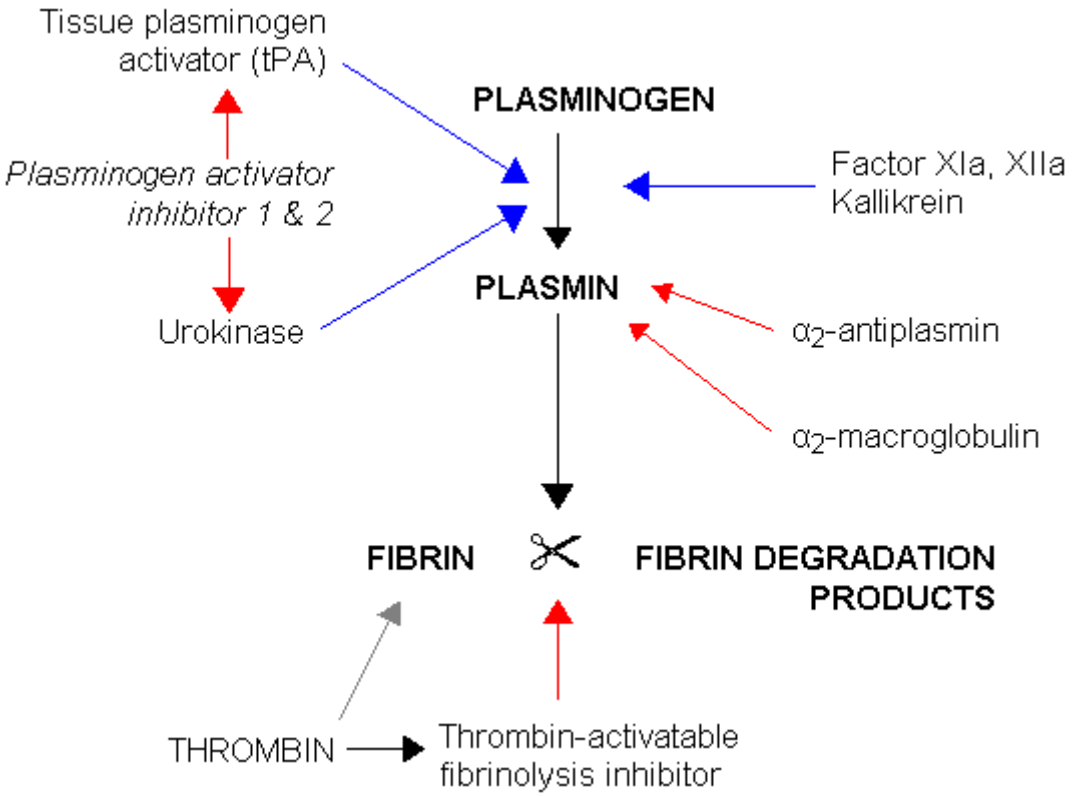


Figura 8: Sistema de activación del Plasminógeno a Plasmina en el que se muestran los principales reguladores, tPA y uPA. El más importante para la regeneración de tejidos es tPA y por tanto el más relacionado con cáncer y virulencia de microorganismos patógenos.

I.I. PROTEÍNAS MOONLIGHTING Y ENFERMEDADES HUMANAS

Las proteínas moonlighting también pueden estar relacionadas con patologías humanas, generalmente por alguna mutación. Por ejemplo, GAPDH participa en la neurodegeneración y el Alzheimer, la IGP en la anemia hemolítica, etc. (Sriram et al., 2005). Por otro lado, hay numerosos ejemplos de proteínas en las que la función moonlighting no es una función "normal" y está relacionada con diferentes patologías por una ganancia de función tóxica. Jeffery ha acuñado el término “Neomorphic Function” (Función Neomórfica) (Jeffery, 2011). Cabe mencionar que a partir del análisis de las proteínas en nuestra base de datos hemos encontrado que el 78% de las proteínas moonlighting humanas en MultitaskProtDB-II están actualmente involucradas en patologías conocidas, según la base de datos OMIM (Hamosh et al., 2005) http://www.omim.org y la base de datos de mutaciones genéticas humanas, http://www.hgmd.cf.ac.uk/ (Cooper y Krawczak, 1998). Además el 48% de las proteínas moonlighting humanas de nuestra base de datos corresponden a dianas de fármacos existentes, según las bases de datos de dianas farmacológicas Therapeutic Target Database (Qin et al., 2014), xin.cz3.nus. edu.sg/group/ttd/ttd/asp, y DrugBank (Wishart et al., 2008), http://www.drugbank.ca. Todo esto complica el análisis de las dianas farmacológicas y la toxicidad de los fármacos, pero a su vez en algunos casos pueden representar una ventaja farmacocinética. Por ejemplo, si el fármaco (caso de un anticuerpo monoclonal) no penetra en la célula y solo afecta la actividad moonlighting si esta es extracelular, o si el medicamento es específico para bloquear la interacción con un “partner” en particular. En este segundo caso, el fármaco tendría menos efectos secundarios (Butler y Overall, 2009). Por ejemplo, THC346, un derivado del deprenyl es neuroprotector por prevenir la S-nitrosilación y, por lo tanto, la interacción de GAPHD con Siah1 (Hara et al., 2005).

I.J. DESPLAZAMIENTO DE GEN NO ORTÓLOGO (NOGD)

El Non-Ortohologous Gene Displacement, o NOGD, describe una forma variante de un sistema o vía en la que un componente esperado se reemplaza por un equivalente funcional que difiere en su origen evolutivo. El NOGD puede ser el ejemplo más familiar de un bioinfotropo. Un desplazamiento génico no ortólogo a menudo se descubre como la explicación de un hueco funcional en una determinada ruta metabólica. El término NOGD fue introducido por Koonin (Koonin et al., 1996).

Una manera de sugerir posibles NOGD es cuando se anota un genoma por alineación de secuencia en bases de datos, algunas secuencias de proteínas previamente bien definidas no coinciden con ninguna del genoma de consulta, creando un "agujero" en la ruta correspondiente. Aunque no se puede descartar la existencia de diferencias en la ruta, es más probable que la función que falta sea realizada por otra proteína con una secuencia de aminoácidos diferente. Si esta proteína mantiene su función anterior (canónica) se convertirá en un verdadero caso de moonlighting.

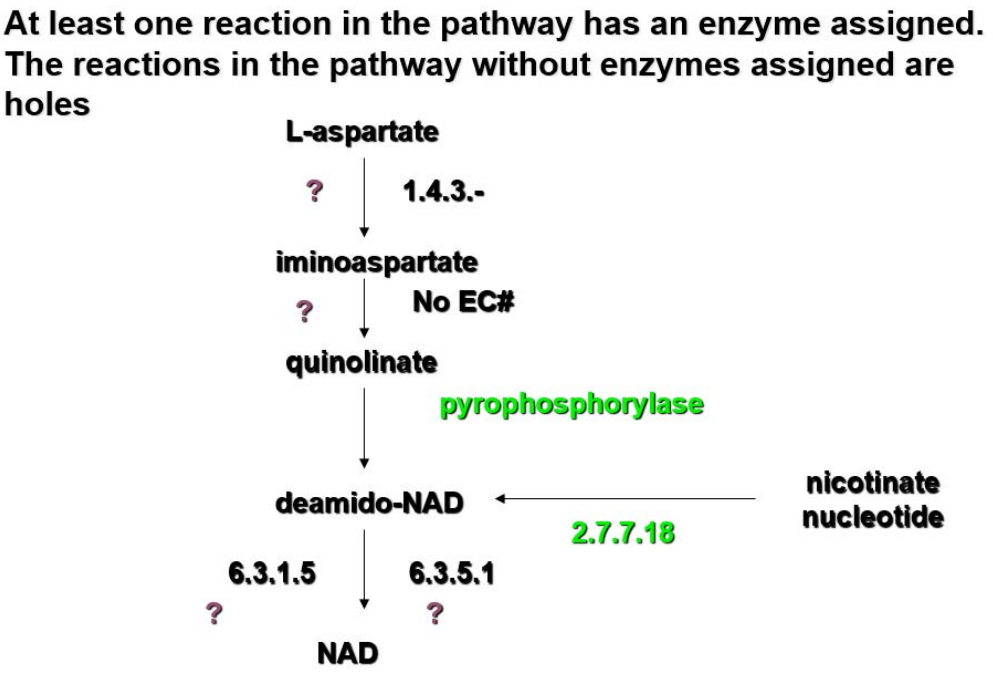


Figura 9: “Pathway holes” o huecos funcionales. Una proteína/gen falta en una ruta metabólica pero la función está siendo realizada por otra proteína.

Galperin publicó una base de datos de enzimas que realizan las mismas funciones en diferentes organismos, pero con una estructura proteica completamente diferente (Galperin et al., 1998) y está disponible en:

https://www.ncbi.nlm.nih.gov/Complete_Genomes/AnalEnzymes.html

Estas proteínas pueden tener relación con las proteínas moonlighting debido a su origen. Como se discutirá en la Sección de Resultados, durante este trabajo, hemos encontrado que 26 proteínas presentes en esta base de datos de proteínas NOGD también están presentes en nuestra primera base de datos de proteínas moonlighting.

Tabla 2: Número de enzimas análogas esperadas en diferentes especies

Organism	Proteins encoded in the complete genome	EC nodes with two analogous enzyme forms in the same genome
Bacteria		
Escherichia coli	4289	35
Haemophilus influenzae	1717	8
Helicobacter pylori	1566	4
Synechocystissp.	3169	18
Borrelia burgdorferi	850	2
Bacillus subtilis	4100	30
Mycoplasma genitalium	467	0
Mycoplasma pneumoniae	677	0
Archaea		
Methanococcus jannaschii	1715	1
Methanobacterium	1869	5
thermoautotrophicum		
Archaeoglobus fulgidus	2407	3
Eukaryotes		
Saccharomyces cerevisiae	5932	22
Caenorhabditis elegans	12178	17

Si dos proteínas son estructuralmente diferentes, aunque estén realizando la misma función, son claramente una proteína diferente con diferente origen evolutivo. Creemos que una "línea de proteína" puede conservar la multifuncionalidad y la otra no. En la Sección Resultados, se puede ver una lista de todas las proteínas NOGD que también son moonlighting.

I.K. ALGUNAS PREGUNTAS RELEVANTES SOBRE LAS PROTEÍNAS MOONLIGHTING

El fenómeno de la multifuncionalidad conduce a una serie de preguntas importantes tanto para la función como para la evolución de las proteínas que aún no se han respondido. En esta tesis intentaremos obtener algunas respuestas o estimar o especular sobre otras. Algunas preguntas son:

a) ¿Qué ventaja evolutiva representa para una proteína que tiene más de una función, a veces ambas funciones indispensables, el hecho de utilizar la duplicación de un gen y la evolución independiente del parálogo? Este hecho evitaría el llamado "conflicto adaptativo". El conflicto adaptativo se refiere a que una mutación que puede 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) ¿Existe una conservación filogenética de la multifuncionalidad?

e) ¿Cuántas proteínas moonlighting hay?

f) ¿Cuál es la relación entre las proteínas moonlighting y el "desplazamiento de genes no ortólogos"?

g) ¿Por qué hay tantos casos de moonlighting en proteínas de adhesión y virulencia de microorganismos patógenos y muchas de ellas son enzimas de la glucólisis?

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

i) ¿Cuántas proteínas moonlighting son posibles dianas terapéuticas y en qué medida son responsables de la toxicidad y los efectos secundarios de muchos fármacos?

j) Finalmente, y es el objetivo fundamental de esta tesis, ¿la multifuncionalidad es predecible bioinformáticamente?

Todas estas preguntas intentan ser respuestas en el apartado de Discusión General.

II. OBJETIVOS

Los objetivos del presente trabajo son:

El **OBJETIVO GENERAL** es el de estudiar el fenómeno de la multifuncionalidad de las proteínas y la relación con su estructura y su evolución.

Como **OBJETIVOS ESPECÍFICOS**:

A. Actualizar la base de datos de proteínas moonlighting publicada previamente por nuestro grupo en 2014, tanto en el número de proteínas como en la información contenida para cada proteína.
B. Contribuir a la predicción de las proteínas moonlighting y a la localización de la función “moon” en la secuencia y estructura de la proteína.
C. Analizar la implicación de las proteínas moonlighting en las enfermedades humanas y como dianas farmacológicas.
D. Profundizar en el papel que juega la multifuncionalidad en la patogenicidad y virulencia y en el mecanismo de infección por microorganismos patógenos y en su posible uso para el diseño de vacunas.
E. Estudiar algunos aspectos evolutivos de las proteínas moonlighting (conservación filogenética, relación con el desplazamiento de genes no ortólogos, abundancia etc.).

Cada uno de estos objetivos corresponde a un capítulo en la Sección Resultados del presente trabajo, y se explica ampliamente en las secciones correspondientes.

III. MÉTODOS

III.A. BASES DE DATOS Y SERVIDORES UTILIZADOS

Bases de datos de proteínas moonlighting

El grupo diseñó la primera base de datos de proteínas moonlighting, con 288 proteínas (Hernández et al., 2014). Durante el desarrollo de este trabajo, se ha publicado una actualización de esta base de datos (http://wallace.uab.es/multitaskII/) (Franco-Serrano et al., 2018) con 694 proteínas moonlighting y con más información respecto a la anterior como se describirá en la Sección IV. A. Hay que tener en cuenta que todos los análisis con respecto a las proteínas moonlighting detallados en la tesis se han realizado utilizando la base de datos actualizada.

Actualmente hay otras dos bases de datos de proteínas moonlighting, MoonProt (http://moonlightingproteins.org/) (Chen et al., 2018), y MoonDB (http://tagc.univ-mrs.fr/MoonDB/) (Chapple et al., 2015), pero contienen menos información que la de nuestro grupo.

Bases de datos de interactómica (PPIs)

Los partners de interacción de las proteínas multifuncionales de la base de datos MultitaskProtDB-II se identificaron en el servidor APID (Alonso-López et al., 2016) (http://cicblade.dep.usal.es:8080/APID/init.action). La nueva versión de APID contiene la mayoría de información de las bases de datos de interacción de proteínas como MINT, DIP, BIOGRID, IntAct, HPRD, BIND, etc. También presenta las proteínas de acuerdo con la anotación GO (Gene Ontology) (www.geneontology.org) (The Gene Ontology Consortium, 2017). Hemos considerado que los datos de interactómica revelan la segunda función de una proteína moonlighting si la base de datos de PPI identifica como “partner” una función molecular, o en algunos casos un proceso biológico (según la anotación GO), que coincide con la función moonlighting esperada. Para filtrar los éxitos y mejorar la precisión, es aconsejable realizar un análisis de enriquecimiento ontológico de genes, por ejemplo, en nuestro caso, utilizando el paquete GOStat de R (Beissbarth y Speed, 2004) como se describió anteriormente (Gómez et al., 2011). Cabe señalar que, en el caso de los datos de interactómica, es

conveniente utilizar bases de datos "sin curar" (por ejemplo, DIP, MINT y el servidor APID) porque en las curadas, muchas interacciones consideradas como falsos positivos se han eliminado y realmente pueden ocultar una segunda función identificable por los partners. Las direcciones web de las principales bases de datos y servidores utilizados son las siguientes:

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
STRING: http://string-db.org

Servidor Expasy/SwissProt (www.expasy.org/swissprot)

Es una base de datos y servidor para el análisis de proteínas creado en 1986 y mantenido en colaboración entre el Swiss Institute of Bioformatics (SIB) y el European Institute of Bioinformatics (EBI). Proporciona un alto nivel de anotación, un nivel mínimo de redundancia de secuencia, un alto nivel de integración con otras bases de datos biomoleculares y una extensa documentación externa. Además de numerosos programas para el análisis de la estructura, función y evolución de las proteínas.

UniProt (www.UniProt.org)

Es una excelente base de datos con información sobre la estructura, secuencia, funciones y bibliografía sobre proteínas (algo más de 111 millones de ellas). En algunos casos, también describe las funciones moonlighting de una proteína. No todas las entradas de proteínas se revisan, por lo que proporciona un número relativamente alto de proteínas redundantes y no del todo identificadas. La información más “fiable” es la que lleva el término “reviewed” y es la que hemos utilizado en el presente trabajo (algo más de 500 mil). Está diseñado y actualizado por el UniProt Consortium. (The Uniprot Consortium, 2017)

PDB (www.pdb.org)

Protein Data Bank (PDB) es una base de datos que contiene estructuras tridimensionales de proteínas obtenidas principalmente por rayos X o RMN. Contiene unas 140.000 estructuras de proteínas la mayoría obtenidas principalmente con rayos X. *Homo sapiens* es el organismo con más estructuras de proteínas en PDB, seguido de *E. Coli* y *Mus Musculus* (Berman et al., 2000).

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

InterPro es un servidor para la identificación de familias de proteínas, dominios, sitios funcionales y "motifs" (secuencias de aminoácidos relacionados con una función, modificación post-traduccional, etc.) (Finn et al., 2017). 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 (2017) contiene 21.165 familias; 9.137 dominios y 912 sitios funcionales.

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 actualizada, ("curada"), para identificar preferentemente y presentar unicamente el motif de mejor puntuación, permite que se puedan indentificar motifs adicionales, 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) que permite identificar familias y dominios de proteínas de forma más o menos restrictiva (PfamA es más "curado" y restrictivo mientras que PfamB lo es 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.

GO (http://www.geneontology.org/)

Una ontología consiste en desarrollar un sistema jerárquico de vocabulario controlado y estructurado para describir con precisión los conceptos y sus relaciones. En Biología Molecular consiste en la descripción de productos genéticos usando términos controlados. Esto evita la anarquía que ha existido durante mucho tiempo en los descriptores de función de las proteínas (anotación funcional). La ontología ha sido desarrollada por un consorcio (GOC) (The Gene Ontology Consortium, 2017) que la actualiza mensualmente. La anotación según GO se llama GOA. GO representa un descriptor triple: (a) *Biological process*: Proceso biológico en el que participa el producto del gen; (b) *Molecular function*: Función molecular o actividad bioquímica; (c) *Cellular component*: lugar de la célula en la que el producto génico está activo. Recientemente, dentro del (a) se ha añadido *Biological phase*, que hace referencia al período o etapa en un proceso o ciclo biológico. También hay una serie de códigos de evidencia adicionales sobre el 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

Se están desarrollando y publicando varios servidores y aplicaciones que ayudan a tratar la información contenida en GO. Uno de ellos, utilizado en algunos análisis durante esta tesis, es QuickGO (www.ebi.ac.uk/QuickGO/).

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

La base de datos SCOP es una clasificación jerárquica, de estructuras de proteínas conocidas, organizadas de acuerdo con sus relaciones evolutivas y estructurales. La base de datos se divide en cuatro niveles jerárquicos: Class, Fold, Superfamily y Family. Una clase compartiría una arquitectura de dominio. Un fold presenta una similitud estructural importante. Una superfamilia de proteínas presenta un probable origen evolutivo común. Una familia de proteínas presenta una clara relación evolutiva (Andreeva et al., 2014).

OMIM (www.omim.org)

OMIM es un compendio exhaustivo y autorizado de genes humanos y fenotipos genéticos relacionados con enfermedades humanas, que está disponible y actualizado diariamente. Las entradas completas en OMIM contienen información sobre todos los trastornos mendelianos conocidos y más de 15,000 genes. (Hamosh et al., 2005)

The human gene mutation database (HGMD) (http://www.hgmd.cf.ac.uk)

Esta base de datos recoge información sobre las alteraciones genéticas conocidas responsables de enfermedades hereditarias humanas. En la versión más reciente incluye 224.642 mutaciones diferentes relacionadas con enfermedades, las más comunes son del tipo “nonsense”, seguidas de splicing. (Cooper y Krawczak, 1998)

Drug Bank (www.drugbank.ca)

La base de datos de DrugBank es un recurso bioinformático y de quimioinformática que combina datos detallados de medicamentos con información integral sobre los mismos (Wishart et al., 2018). La última versión de

DrugBank (versión 5.0.11, del 2017) contiene 11.002 entradas de fármacos, incluidos 2.504 fármacos de molécula pequeña aprobados, 943 fármacos biotec (proteína/péptido o ácido nucleico) aprobados, 109 nutraceúticos, además de 5.110 fármacos experimentales. Cada entrada de DrugBank contiene más de 200 campos de datos con la mitad de la información dedicada a datos farmacológicos/químicos y la otra mitad dedicada a datos farmacológicos o de proteínas.

VIOLINET (http://www.violinet.org)

Vaccine Investigation and Online Information Network (VIOLINET) es una base de datos de todas las vacunas, humanas o no, existentes en el Mercado o en fase de desarrollo, en la que además se indican los microorganismos patógenos y los genes involucrados. (He et al., 2014)

The human protein atlas (www.proteinatlas.org)

The Human Protein Atlas es una base de datos que recoge todas las proteínas humanas en células, tejidos y órganos mediante la integración de varias tecnologías ómicas, que incluyen imágenes basadas en anticuerpos, proteómica basada en espectrometría de masas, transcriptómica y biología de sistemas. Todos los datos son de acceso abierto. (Uhlén et al., 2015)

Non-Orthologous Gene Displacement (NOGD)

(www.ncbi.nlm.nih.gov/Complete_Genomes/AnalEnzymes.html)

Se trata de una base de datos de proteínas NOGD. En ella se pueden encontrar aquellas proteínas que realizan la misma función molecular (mismo E.C. number y por lo tanto misma actividad enzimática) pero tienen una estructura distinta (diferentes folds estructurales). (Galperin et al., 1998)

III. B. DISEÑO DE UNA ACTUALIZACIÓN DE LA BASE DE DATOS DE PROTEÍNAS MOONLIGHTING (MultitaskProtDB-II)

Como se ha mencionado anteriormente, el grupo diseñó y publicó la primera base de datos de proteínas moonlighting (Hernández et al., 2014). Durante el desarrollo de este trabajo, se ha actualizado y rediseñado esta base de datos con más proteínas y con más información relevante para cada proteína, como se discutirá en la Sección IV.A.

La información sobre proteínas multifuncionales se ha recopilado de la literatura a través del servidor NCBI PubMed (http://www.ncbi.nlm.nih.gov/pubmed) utilizando los siguientes términos y palabras clave: *moonlighting proteins; moonlight proteins; multitask proteins; multitasking proteins; moonlight enzymes; moonlighting enzymes; gene sharing*. Además, algunas características importantes de la proteína fueron recopiladas utilitzando el servidor UniProt

Con el fin de identificar qué proteínas de nuestra base de datos están involucradas en enfermedades humanas, la información presente en las bases de datos Online Mendelian Inheritance in Man (OMIM) y Human Gene Mutation Database (HGMD), han sido inspeccionadas. Además, para verificar qué proteínas de nuestra base de datos son dianas farmacológicas, se han utilizado las bases de datos Therapeutic Targets Database (TTD) (http://bidd.nus.edu.sg/group/cjttd/) y DrugBank (www.drugbank.ca). La estructura tridimensional de esas proteínas sin estructura 3D previamente resuelta presente en PDB se modeló aplicando los servidores I-Tasser (http://zhanglab.ccmb.med.umich.edu/I-TASSER) y Phyre2 (http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index), ver Sección III.D. Ambos métodos usan modelos de estructura terciaria basados en patrones.

La base de datos MultitaskProtDB-II ha sido diseñada utilizando MySQL. El servidor web ha sido diseñado con PHP y asistido por PHPRunner, una aplicación que ayuda a generar código PHP y crea archivos, informes, listas y formularios que facilitan el desarrollo de partes importantes de la web. El sistema tiene un motor de búsqueda avanzado para permitir una búsqueda más precisa o más restringida. Este tipo de procedimiento sirve para limitar la búsqueda al subconjunto de proteínas en el que realmente se quiere centrar el estudio.

III.C. ALINEAMIENTOS DE SEQUENCIAS

III.C.1. ALINEAMIENTOS USANDO BLAST Y PSI-BLAST

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 general se asume que, en el caso de las proteínas, por encima de 25-30% de identidad en fracciones o tramos de por lo menos 80 aminoácidos, probablemente presentan la misma función, o una función bioquímica y estructura tridimensional similar. Aunque en sentido estricto para que dos secuencias de dos organismos diferentes 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 PSI-Blast, 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 moonlighting 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 se reordena el listado y envia a posiciones superiores dianas que, a pesar de ser verdaderos positivos, han acabado en posiciones alejadas.

III.C.2. ALINEAMIENTOS MÚLTIPLES DE SECUENCIAS

Los alineamientos múltiples son útiles para predecir estructuras de proteínas, motifs y dominios funcionales, así como identificar aminoácidos clave para la función de proteínas y esenciales para el análisis filogenético. En el presente trabajo se ha utilizado el alineamiento múltiple, por ejemplo, para sugerir que la conservación de secuencias entre diferentes especies, especialmente de regiones o motivos implicados en cada una de las funciones, implicaría que la característica de multifuncionalidad se conservaría evolutivamente. También para la comparación de secuencias de proteínas humanas con las ortólogas de microorganismos patógenos con objeto de determinar si comparten epítopos lineales. Se ha utilizado el programa Clustal-Omega en el servidor del EBI (European Bioinformatic Institute), http://www.ebi.ac.uk/Tools/msa/clustalo/ (Li et al., 2015).

III.D. ANÁLISIS Y MODELADO DE LA ESTRUCTURA TRIDIMENSIONAL DE PROTEÍNAS

Los programas para identificar sitios funcionales pueden ayudar a revelar funciones adicionales de la proteína si se conoce su estructura tridimensional (Aloy et al., 2001). Para comprobar si los principales sitios estructurales/funcionales para ambas funciones se pueden descubrir a partir de la secuencia de la proteína, hemos utilizado PiSite (Higurashi et al., 2009), un programa de modelado estructural que alinea trozos de secuencia de la proteína problema con secuencias de proteínas cuya estructura 3D esté en la base de datos PDB. Este procedimiento de identificar patrones tiene una especial relevancia para nuestros objetivos porqué permite mapar dominios funcionales independientes, y por ello multifuncionalidad. Este programa está disponible en la web http://PiSite.hgc.jp/. Phyre2 (Kelley y Sternberg, 2009) es otro programa

de modelado estructural que hemos utilizado, y que es accesible en http://www.imperial.ac.uk/phyre. Este programa tiene la ventaja de que proporciona una puntuación ("score") de calidad de la estructura modelada. El programa también puede revelar residuos de aminoácidos funcionales clave, que pueden ayudar a identificar sitios funcionales adicionales. Finalmente, otro programa de modelado estructural, más complejo de utilizar, pero superior en la calidad de la estructura final (mejor RMSD) es I-Tasser (Wang et al., 2017). Este programa se ha utilizado para los modelados 3D de las proteínas moonlighting de la base de datos para las cuales no había una estructura tridimensional en el PDB. Otros programas complementarios que ayudan a trabajar y visualizar las estructuras de proteínas son:

- Chimera (www.cgl.ucsf.edu/chimera/)

- Swiss PDB viewer (https://spdbv.vital-it.ch/)

- RasMol (http://www.openrasmol.org/)

Swiss PDB viewer, permite alinear o "ajustar" diferentes proteínas minimizando el RMSD (Root-Mean-Square Deviation) de las posiciones atómicas, incluso permite calcularlo. El RMSD es la medida cuantitativa más utilizada de la similitud entre dos coordenadas atómicas superpuestas. Los valores RMSD se presentan en Å y se calculan mediante:

$$RMSD = \sqrt{\frac{1}{n}\sum_{i=1}^{n} d_i^2}$$

Donde el promedio se realiza sobre los n pares de átomos equivalentes y d_i es la distancia entre los dos átomos en el i-th par. RMSD se puede calcular para cualquier tipo y subconjunto de átomos; por ejemplo, átomos Cα de la proteína completa, átomos Cα de todos los residuos en un subconjunto específico (por ejemplo, las hélices transmembrana, sitios de unión o un loop), todos los átomos de un subconjunto específico de residuos o todos los átomos de un pequeño ligando.

Con el objetivo de encontrar si dentro del grupo de proteínas moonlighting, abunda algún fold estructural en concreto, se recogieron datos de los folds

presentes en todas las proteínas de nuestra base de datos de proteínas moonlighting del servidor SCOPe (http://scop.berkeley.edu). En la Sección IV.A.2. se pueden consultar los folds más abundantes entre las proteínas moonlighting.

III.E. OTROS PROGRAMAS DE ANÁLISIS DE SEQUENCIAS PARA IDENTIFICAR MOTIFS Y DOMINIOS FUNCIONALES

Como ya se ha descrito en la Introducción y también se desarrollará posteriormente en Resultados, un 25% de las proteínas moonlighting de la base de datos MultitaskProtDB-II son factores de virulencia de microorganismos patógenos. La mayoría son proteínas del metabolismo primario y, tras ser secretadas, se unen al Plasminógeno y otras proteínas de la matriz extracelular. Nos preguntamos si tales proteínas, secuencial y estructuralmente diferentes, comparten dominios o motifs que permitan la interacción con las mismas dianas. Con este fin, se analizaron las proteínas de MultitaskProtDB-II con servidores de búsqueda de patrones, especialmente InterPro, que ya se ha comentado anteriormente que incorpora 11 subprogramas. El resultado no mostró la existencia de motifs relacionables con la unión a Plasminógeno o la virulencia del mismo. Para ello se rastrearon motifs de menor tamaño utilizando el programa MinimotifMiner (Mi et al., 2012)

http://cse-mnm.engr.uconn.edu:8080/MNM/SMSSearchServlet

Este servidor, busca motifs cortos, en general relacionados con sitios de interacción, modificaciones post-traduccionales, etc. Se ha diseñado un programa que rastrea automáticamente con MinimotifMiner aquellos motifs presentes en las proteínas moonlighting de los microrganismos patógenos y ausentes en los no patógenos, como se describe en el esquema de la Figura 10.

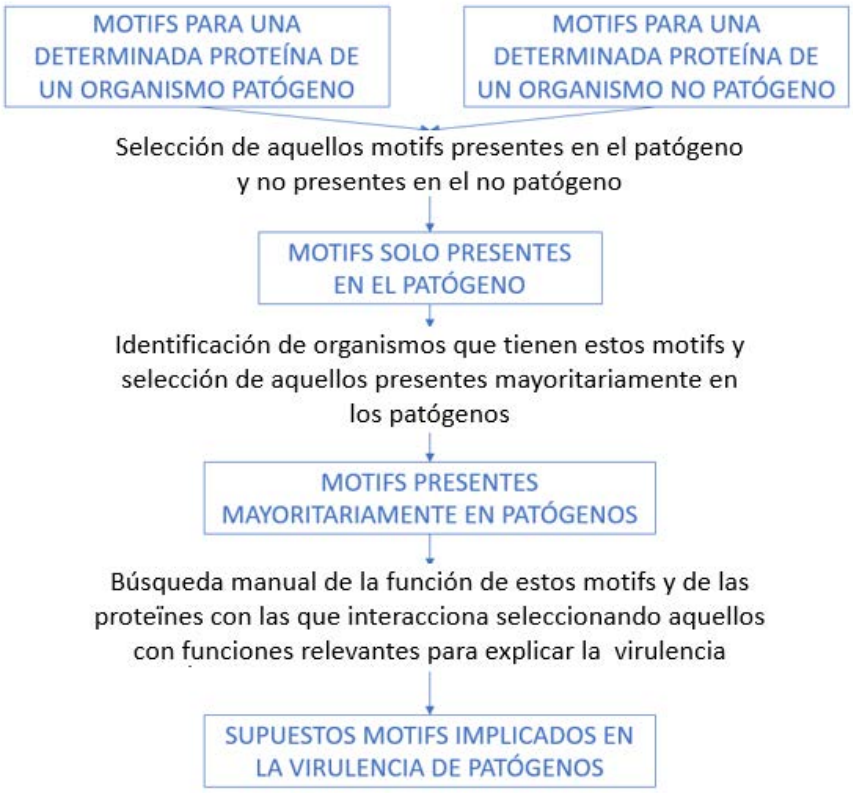


Figura 10: Proceso seguido para la identificación de los motifs relacionados con la virulencia de microorganismos patógenos descritos en la Sección IV.D.

III.F. IDENTIFICACIÓN DE NUEVAS PROTEÍNAS MOONLIGHTING Y SU CONSERVACIÓN FILOGENÉTICA

III.F.1. MEDIANTE INTERACTÓMICA

Ya en un trabajo anterior habíamos propuesto que en las bases de datos de interactómica (PPI databases) hay mucha información que permitiría identificar proteínas multifuncionales (Gomez et al., 2011). En muchos casos se trataba de partners de interactómica descartados como Falsos Positivos por los investigadores experimentales. También hemos propuesto que combinar el análisis de similitud de secuencia (por Blast o PSI-Blast) 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 "culpable 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 estas identifican una "*Molecular Function"* o, en algunos casos, un "*Biological Process*" de acuerdo con la anotación del GO y además está de acuerdo con la función moonlighting de la proteína de nuestra base de datos MultitaskProtDB-II. A continuación, y para filtrar e incrementar la precisión de la predicción, es aconsejable realizar un análisis de enriquecimiento GO. Para ello, para cada proteína moonlighting incluida en APID se capturaron los términos GO de los partners de interacción y se calculó el "GO term enrichment" mediante el programa "GOStat R package" (Beissbarth and Speed et al., 2004). Esta función calcula p-values 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 moonlighting aquellos terminos GO con un p-value menor que 0.05, lo que nos permite eliminar bastantes descriptores GO inespecíficos (Gomez et al., 2011).

Con el objetivo de determinar si las funciones moonlighting se conservan entre organismos filogenéticamente cercanos, se ha realizado un análisis del interactoma de las proteínas moonlighting de MultitaskProtDB-II y comparado con el interactoma de proteínas equivalentes de especies cercanas (lo que se puede denominar como ortología de interactomas). Para ello se ha utilizado la base de datos de PPI, APID. El proceso seguido para este análisis es el siguiente: (a) se analizan los partners de interacción de las proteínas moonlighting utilizando APID, identificando aquellos partners cuyas funciones coincidan con las funciones moonlighting, (b) se seleccionan proteínas equivalentes a la anterior en otras especies, en las que no se ha descrito que sean moonlighting, (c) se analizan los partners de interacción de estas proteínas para detectar aquellos que puedan relacionarse con la función moonlighting que podría estar conservada en esa especie. Los resultados se pueden observar en la Sección IV.E.2.b.

III.F.2. A PARTIR DE LA INFORMACIÓN EXISTENTE EN LA BASE DE DATOS UNIPROT

Las bases de datos UniProt y GO contienen información relativa a las funciones de las proteínas. UniProt describe las funciones con texto y GO las codifica con una serie de GO numbers identificadores, aunque en ocasiones redundantes. Es importante mencionar que, UniProt proporciona una gran cantidad de información de la proteína tales como: aminoácidos claves para las funciones, bibliografía, enlaces a las estructuras tridimensionales, enfermedades y las mutaciones relacionadas, etc.

UniProt no permite una fácil automatización del análisis de las funciones de las proteínas ya que su descripción consiste en un texto largo sin ningún patrón repetido ni ninguna clasificación funcional, pero contiene mucha información para cada proteína. Hay una gran diferencia entre aquellas proteínas “reviewed”, es decir, aquellas que miembros de UniProt han revisado y las “no-reviewed”. Las primeras contienen mucha más información funcional y estructural, además han eliminado las redundancias con otras proteínas, ya que UniProt contiene entradas tales como “unknown protein” en muchos casos mostrando únicamente la secuencia de la proteína.

UniProt a su vez, añade algunos códigos GO relevantes, aunque no todos para cada proteína. GO en cambio, únicamente contiene información funcional. Esto permite una automatización del análisis y estamos trabajando en un programa que permitirá en un futuro analizar y proponer qué proteínas son moonlighting utilizando la clasificación GO. De momento los resultados son prometedores ya que el programa permite predecir más de un 90% de nuestra base de datos de proteínas moonlighting. El problema principal es establecer los listados de identificadores que realmente vayan ligados de forma unívoca a las funciones biológicas, pues en GO hay una enorme cantidad de identificadores redundantes o poco significativos.

Pero sin diseñar y utilizar ningún programa también es posible encontrar proteínas moonlighting, únicamente utilizando el buscador avanzado de UniProt. Con este buscador podemos pedirle que nos muestre aquellas proteínas que presenten diversos identificadores GO concretos, por ejemplo, algunos

asociados a las funciones de enzima y factor de transcripción, o bien que nos muestre aquellas proteínas presentes en dos compartimentos celulares distintos. Se han identificado un buen número de posibles proteínas moonlighting utilizando el buscador avanzado de UniProt, los pares de funciones más abundantes los describimos en las Tablas 15,16 y 17 de la Sección IV.E.

III.F.3. A PARTIR DE LA INFORMACIÓN EXISTENTE EN LA BASE DE DATOS OMIM

Con el objetivo de cruzar la información entre las bases de datos UniProt y OMIM, se ha seguido el procedimiento secuencial descrito en la Figura 11: (a) hacer una lista de proteínas UniProt relacionadas con enfermedades según la base de datos OMIM y la literatura (se recogieron 3600 proteínas inicialmente); (b) eliminar las entradas de enfermedades causadas por más de una proteína; (c) seleccionar solo las proteínas que causan más de una enfermedad; (d) revisar manualmente las enfermedades causadas por cada una de estas proteínas y: (d1) seleccionar aquellas proteínas en las que las enfermedades no están relacionadas entre sí, o (d2) seleccionar aquellas proteínas en las que las enfermedades no parecen estar relacionadas (esto significa una base molecular diferente) a la función canónica de la supuesta proteína moonlighting. Esta lista final debe contener entradas que puedan caracterizarse como supuestas proteínas moonlighting (Figura 11).

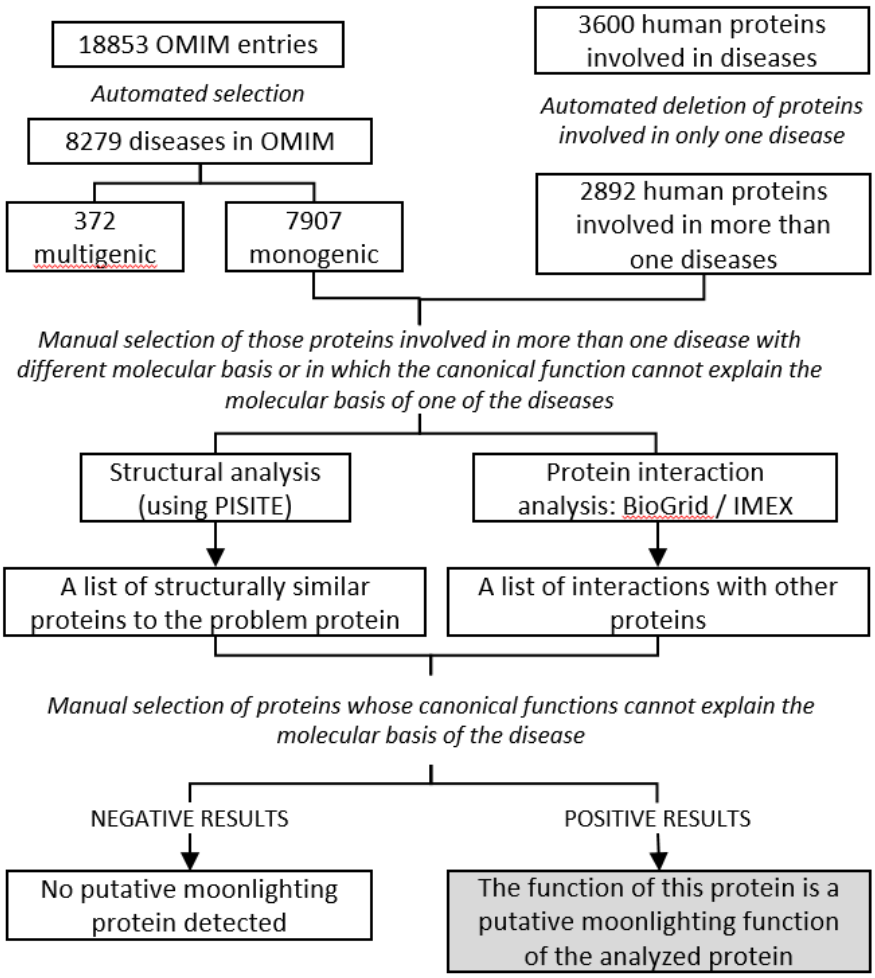


Figura 11: Proceso seguido para predecir proteínas moonlighting a partir de la base de datos OMIM, análisis estructural y de interacción proteína – proteína.

III.G. RELACIÓN ENTRE PROTEÍNAS MOONLIGHTING, ENFERMEDADES HUMANAS Y DIANAS FARMACOLÓGICAS.

Bases de datos

Utilizamos las proteínas moonlighting de la base de datos MultitaskProtDB-II (Franco-Serrano et al., 2018) (http://wallace.uab.es/multitask). La información presente en la base de datos Human Mendelian Inheritance in Man (OMIM, www.omim.org) (Hamosh et al., 2005) y la base Human Gene Mutation Database (HGMD, www.hgmd.cf.ac.uk) (Cooper et al., 1998), cada una de las proteínas ha sido cuidadosamente inspeccionada. Con esta estrategia, se identificaron las proteínas moonlighting que están involucradas en enfermedades humanas. Además, se ha rastreado la base de datos de dianas terapéuticas (TTD,

http://bidd.nus.edu.sg/group/cjttd) (Quin et al., 2014) y la base de datos de DrugBank (www.drugbank.ca) (Wishart et al., 2008) en busca de información relevante para ver si cada una de las proteínas moonlighting es una diana farmacológica. Algunas características importantes de la proteína han sido obtenidas de UniProt (www.UniProt.org). Si está disponible, la estructura tridimensional de la proteína se ha obtenido del PDB (www.rcsb.org).

La significación estadística de los datos obtenidos a partir de estas fuentes se analizó mediante un ODD ratio test. El intervalo de confianza (IC) utilizado fue del 95%. El ODD ratio test es un sistema estándar para medir el grado de asociación entre las variables categóricas de dos estados (en nuestro caso, enfermedad/no enfermedad frente al moonlighting/no-moonlight y druggable/no-druggable versus moonlighting/no-moonlighting). Para establecer la significación estadística de las diferencias observadas, se ha calculado la Fisher's exact test utilizando R.

Mapado y vinculación de enfermedades a la función canónica o a la moonlighting

Se realizó un análisis exhaustivo de la literatura relacionada con las enfermedades asociadas con las proteínas multifuncionales humanas de la base de datos MultitaskProtDB-II. Para cada uno de estos casos, se han estudiado las características de la patología con objeto de relacionar a nivel molecular, la enfermedad con la función canónica, moonlighting o ambas funciones. En algunos de ellos, no se encontraron suficientes datos en la literatura para relacionar la patología con una de las funciones. Cuando existen datos relevantes y para demostrar esta condición y tratar de mapar las funciones canónicas y moonlighting en la estructura de la proteína, se ha propuesto el uso de una combinación de diferentes métodos. Si la estructura tridimensional de la proteína está disponible, se puede utilizar el programa PiSite (Higurashi et al., 2009), según lo descrito en el apartado III.D. (Hernández et al., 2014). Este algoritmo busca proteínas con una estructura tridimensional similar a la proteína de consulta. Los resultados nuevamente deberían revisarse manualmente y usarse para buscar estructuras de proteínas que puedan explicar la enfermedad que no es causada por la función canónica. En estos casos, la función que realiza esta nueva proteína podría ser la función moonlighting de la proteína original.

También se pueden usar bases de datos de interacción de proteínas, como APID (Alonso-Lopez et al., 2016), BioGRID (Chatr-Aryamontri et al., 2015) o IMEX (Orchard et al., 2012). En este caso, podemos buscar proteínas que interaccionen con nuestra proteína y que puedan sugerir una explicación de la enfermedad no relacionada con la función canónica de la proteína de consulta. Además, esta proteína "partner" de interacción podría tener una estructura tridimensional ya resuelta y, por lo tanto, el método PiSite mencionado anteriormente podria aplicarse para predecir regiones funcionales de la proteína, utilizando las anotaciones estructurales que pueden obtenerse de la literatura y de UniProt. Esta información contiene las regiones funcionales clave y los aminoácidos de la proteína que deben usarse para relacionar estas características con el efecto patológico de la enfermedad. Finalmente, estas regiones importantes deberían estar localizadas en la estructura de la proteína, verificando que están en diferentes zonas para mapar las funciones canónicas y moonlighting.

III.H. RELACIÓN ENTRE PROTEÍNAS MOONLIGHTING Y VIRULENCIA DE MICROORGANISMOS PATÓGENOS

¿Por qué muchas proteínas del metabolismo primario estan relacionadas con la virulencia de microorganismos patógenos?

En un trabajo previo (Amela et al., 2007) el grupo sugirió que el sistema inmune evita desarrollar respuesta contra proteínas de microorganismos patógenos que comparten epítopos con alguna proteína del huésped. Basándonos en esto y en que la gran mayoría de proteínas moonlighting de patógenos corresponden a funciones del metabolismo primario, es decir, evolutivamente conservadas, se procedió a predecir sus correspondientes epítopos B con objeto de ver si los compartían con las ortólogas humanas o animales. Asimismo, como control se procedió a comprobar si las proteínas que han dado lugar a una vacuna por subunidades presentaban homología secuencial y por tanto de epítopos, con alguna proteína humana o animal.

Las proteínas de virulencia de los patógenos que son moonlighting se obtuvieron de MultitaskProtDB-II (Franco-Serrano et al., 2018). Las proteínas candidatas a

ser vacunales se obtuvieron de la base de datos Violinet (He et al., 2014), que contiene 800 proteínas que se han probado como vacunas de subunidades (recombinantes o aisladas) y luego purificadas. Estas vacunas pueden estar en diferentes estados: ya comercializadas; licenciadas y en estado de investigación.

Los epítopos lineales de células B de las proteínas ortólogas humanas de las proteínas de virulencia de patógeno previamente mencionadas (por ejemplo, Enolasas) se predijeron usando el algoritmo BepiPred (Larsen et al., 2006). Los alineamientos de secuencia se realizaron con BLASTP del servidor NCBI (Altschul et al., 1997), y los alineamientos múltiples se realizaron con Clustal-Omega del servidor EBI (Li et al., 2015). Ambos análisis se realizaron bajo los parámetros predeterminados del servidor. Los resultados obtenidos pueden verse en la Sección IV.D de este trabajo.

En cuanto a la relación de las proteínas moonlighting y la virulencia de microorganismos patógenos se han seguidos dos líneas de investigación, ambas basadas en la abundancia de proteínas del metabolismo primario con función de virulencia. Por un lado, se han intentado localizar motifs o regiones comunes en estas proteínas que puedan implicar una interacción con las proteínas del huésped a las que se unen (Plasminógeno, Fibronectina, etc.) y por otro lado se ha intentado explicar porque las proteínas del metabolismo primario relacionadas con virulencia no producen respuesta inmune eficaz en el huésped y no son buenas candidatas como dianas vacunales. Estos resultados han sido publicados durante la realización de esta tesis. (Franco-Serrano et al., 2018b)

Identificación de motifs o regiones comunes relacionadas con la virulencia

En este sentido, tal y como se describe en el apartado III.E., se ha utilizado el servidor MinimotifMinner, además de programas de creación propia y análisis manual para la identificación de motifs que probablemente estén relacionados con virulencia. Esto es porque interaccionan con proteínas que regulan el sistema inmunitario y participan en diversas funciones relacionadas con la sangre y el Plasminógeno. Este análisis también se ha realizado para determinar que los motifs de unión a Plasminógeno descritos actualmente en la bibliografía estan presentes tanto en microorganismo patógenos como no patógenos.

IV. RESULTADOS

IV.A. BASE DE DATOS MULTITASKPROTDB-II Y ALGUNAS CONSIDERACIONES ACERCA DE SU CONTENIDO

IV.A.1. ACTUALIZACIÓN DE LA BASE DE DATOS

El grupo diseñó y publicó la primera base de datos de proteínas moonlighting accesible en: http://wallace.uab.es/multitask/ (Hernández et al., 2014). En el presente trabajo se ha rediseñado, actualizado y publicado una base de datos renovada con importantes mejoras tanto en el número de proteínas como en la información para cada proteína (Franco-Serrano et al. 2018) y está disponible en: http://wallace.uab.es/multitaskII (Figura 12). La nueva base de datos contiene 694 proteínas moonlighting.

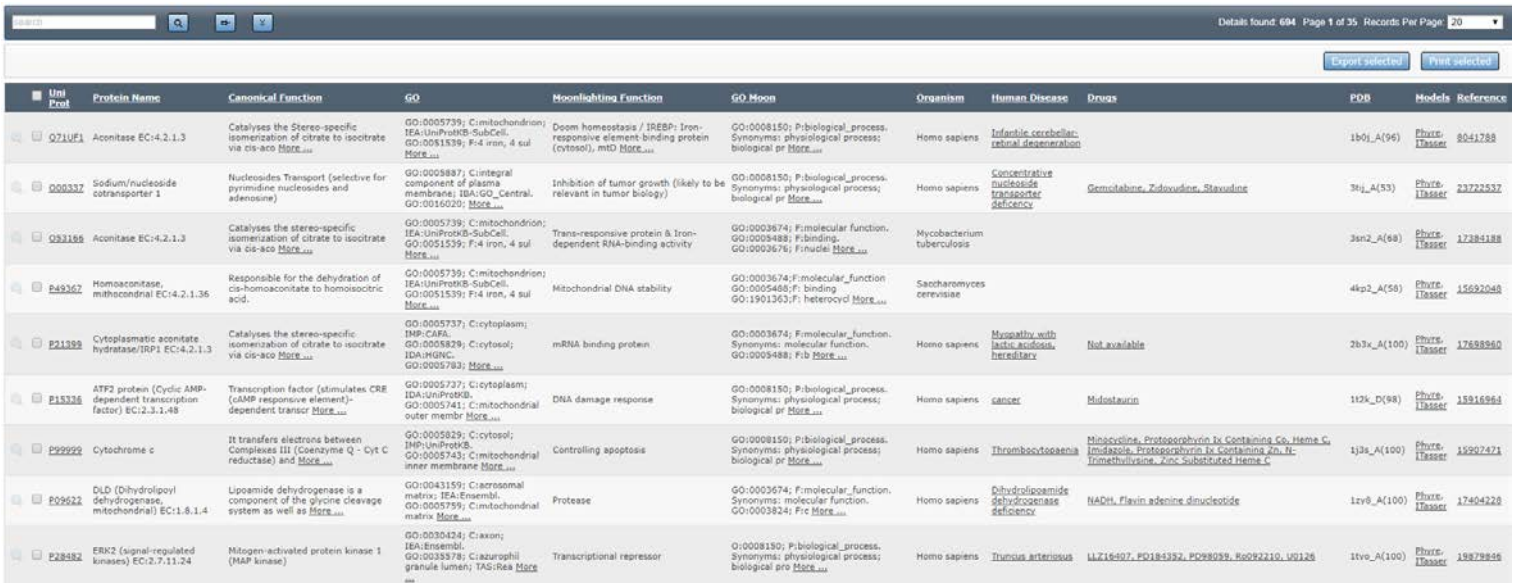


Figura 12: Una captura de pantalla de MultitaskProtDB-II.

La información disponible para cada proteína, en orden de columnas de izquierda a derecha, es la siguiente. Las columnas están enlazadas a la información correspondiente.

- La columna 1 es un botón que permite visualizar las características de una proteína moonlighting concreta.
- La columna 2 es un botón que permite seleccionar una proteína.
- La columna 3 (*UniProt*) muestra el número de acceso de UniProt y permite acceder a esta base de datos.
- La columna 4 (*Protein Name*) muestra el nombre de la proteína.
- La columna 5 (*Canonical Function*) contiene una descripción detallada de la función canónica de la proteína.
- Las columnas 6 y 8 (*GO y GO Moon*) muestran los números GO relacionados con las funciones canónica y moonlighting de la proteína.
- La columna 7 (*Moonlighting Function*) contiene una descripción detallada de las funciones moonlighting.
- La columna 9 (*Organism*) indica el organismo en el que la proteína actúa como una proteína moonlighting de acuerdo con la bibliografía.
- La columna 10 (*Human Disease*) indica las enfermedades asociadas, en el caso de las proteínas moonlighting humanas y están enlazadas a la base de datos OMIM.
- La columna 11 (*Drugs*) indica si la proteína es una diana conocida de algún medicamento actual y está enlazada con la información correspondiente en la base de datos de DrugBank.
- La columna 12 (*PDB*) indica la estructura tridimensional en la base de datos PDB correspondiente a la proteína resuelta experimentalmente, si se resolvió experimentalmente. La cifra entre paréntesis indica el porcentaje de identidad con una proteína de PDB en el caso de que se haya modelado utilizando esta estructura homóloga de PDB como plantilla. Cuando la estructura está resuelta experimentalmente la cifra es 100.
- La columna 13 (*Models*) proporciona el modelo de estructura 3D de la proteína moonlighting de acuerdo con los servidores I-Tasser o Phyre2.

- La columna 14 (*References*) proporciona un enlace a la referencia bibliográfica en PubMed.

Además, la página web proporciona algunos recursos, como botones de visualización, impresión o búsqueda. También incluye, un fácil proceso de exportación de toda la base de datos o de algunas entradas seleccionadas. El tipo de archivo de datos obtenido a través de la opción de exportación se puede seleccionar según el tipo de archivo de datos requerido por el usuario (Excel, Word, CSV o XML). Además, se puede ordenar alfabéticamente cada columna haciendo clic en el título de la columna, y esto permite ordenar, por ejemplo, por organismos. En la Información Suplementaria S1, se puede encontrar un listado completo de las proteínas moonlighting en MultitaskProtDB-II.

IV.A.2. ALGUNAS INFERENCIAS A PARTIR DE LA BASE DE DATOS MULTITASKPROTDB-II

Folds que presentan las proteínas moonlighting

Una de las posibles preguntas que se pueden hacer sobre las proteínas moonlighting es si éstas muestran preferencias por ciertas arquitecturas o folds estructurales. Para ello, utilizando el código SCOP asociado a la estructura PDB con la que se alineó la secuencia de la proteína moonlighting, se realizó una tabla de frecuencias de los principales folds observados (Tabla 3).

Tabla 3: Principales folds observados en las proteínas moonlighting

SCOPe ID	FOLDS	FREQUENCY
c1	TIM beta/alpha-barrel	9
c2	NAD(P)-binding Rossmann-fold domains	9
b1	Immunoglobulin-like beta-sandwich	6
c47	Thioredoxin fold	6
c37	P-loop containing nucleoside triphosphate hydrolases	5
c55	Ribonuclease H-like motif	4
c57	Molybdenum cofactor biosynthesis proteins	4
d144	Protein kinase-like (PK-like)	4
d54	Enolase N-terminal domain-like	4
i1	Ribosome and ribosomal fragments	4
a118	alpha-alpha superhelix	3
c8	The "swivelling" beta/beta/alpha domain	3
d15	beta-Grasp (ubiquitin-like)	3
d162	LDH C-terminal domain-like	3
d58	Ferredoxin-like	3
a127	L-aspartase-like	2
a45	GST C-terminal domain-like	2
b26	SMAD/FHA domain	2
b29	Concanavalin A-like lectins/glucanases	2
b35	GroES-like	2
b42	beta-Trefoil	2
b69	7-bladed beta-propeller	2
b85	beta-clip	2
c23	Flavodoxin-like	2
c26	Adenine nucleotide alpha hydrolase-like	2
c42	Arginase/deacetylase	2
c58	Aminoacid dehydrogenase-like, N-terminal domain	2
c67	PLP-dependent transferase-like	2
	OTHER	74

Se puede ver que, los dos folds más abundantes son el TIM beta/alpha-barrel y el NAD(P)-binding Rossmann-fold. Ambos folds son muy abundantes y se corresponden con el par de funciones prevalentes entre las proteínas moonlighting, es decir, ser una enzima y un factor de transcripción.

- **TIM beta/alpha-barrel** es un motif estructural que contiene un barril paralelo de hojas beta. El TIM barrel es un fold proteico que consta de ocho hélices α y ocho hojas β paralelas que se alternan a lo largo de la cadena principal del péptido. La estructura lleva el nombre de la triosafosfato isomerasa, una enzima metabólica conservada. Los TIM barrel son uno de los folds de proteína más comunes. También es común en las enzimas, que son el tipo de proteína más abundante entre las proteínas moonlighting. Quizás esas son las razones por las que el TIM barrel también es el más común entre las proteínas moonlighting.

- **NAD(P)-binding Rossmann-fold domains.** El fold de Rossmann es un motif estructural que se encuentra en proteínas que se unen a nucleótidos, como los cofactores enzimáticos FAD, NAD + y NADP +. Este plegamiento se compone de hojas beta alternadas con segmentos alfa helicoidales donde las cadenas beta están unidas por enlaces de hidrógeno formando una lámina beta extendida y las hélices alfa rodean ambas caras de la lámina para producir un sándwich de tres capas.

En cuanto a las clases de proteínas, la más común en las proteínas moonlighting es la clase "c", que corresponde a las proteínas alfa-beta (α/β), seguidas de la clase "d", las proteínas alfa+beta (α+β) (Figura 13).

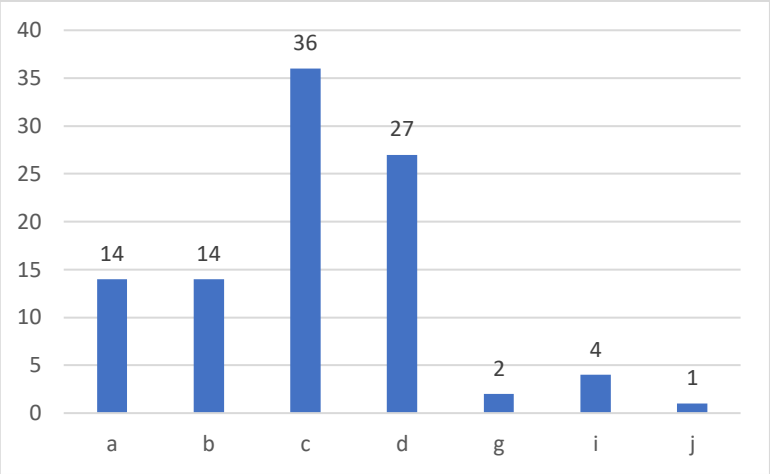


Figura 13: Distribución de las clases de proteínas en las proteínas moonlighting. (a) Toda alfa (b) Toda beta (c) Proteínas alfa-beta (α/β) (d) Proteínas alfa+beta (α+β) (g) Pequeñas proteínas (i) Estructuras proteicas de baja resolución (j) Péptidos.

Sin embargo, estos resultados son diferentes si se habla de proteínas normales (no moonlighting) como se puede ver en la Figura 14.

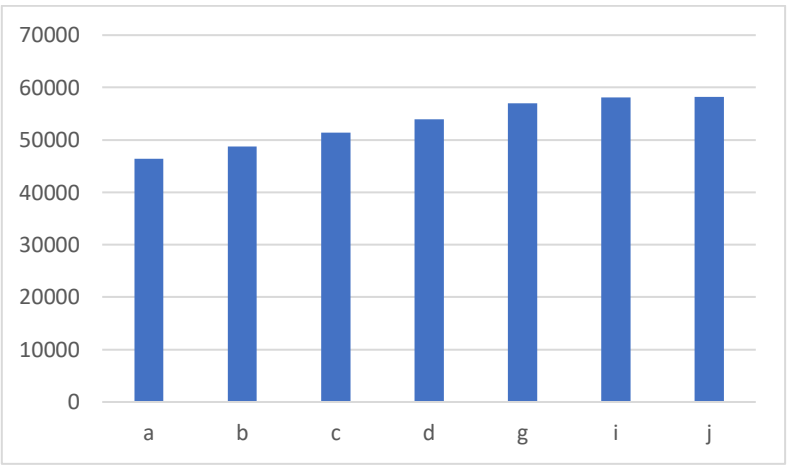


Figura 14: Distribución de las clases de proteínas en el proteoma general (a) Toda alfa (b) Toda beta (c) Proteínas alfa-beta (α/β). (d) Proteínas alfa+beta (α+β). (g) Pequeñas proteínas (i) Estructuras proteicas de baja resolución (j) Péptidos.

Para analizar cualquier preferencia de folds en nuestra base de datos de proteínas moonlighting, todas las proteínas se alinearon con la base de datos astral95, y se hizo un subconjunto de proteínas considerando solo aquellos con menos del 95% de identidad (moon95). Con esta estrategia, quisimos evitar la abundancia de la misma proteína de especies cercanas y evitar proteínas

sobrerrepresentadas debido a la acumulación de la misma proteína con múltiples funciones moonlighting. Para comprobar si la distribución de las frecuencias de fold es similar a la que veríamos si las proteínas moonlighting tuvieran la misma distribución de folds que la observada en la base de datos astral95, la distribución de frecuencias del subconjunto moon95 se comparó con la distribución presente en las proteínas de la base de datos astral95. Esto se hizo usando un G-test calculado a través del paquete estadístico R (www.r-project.org). El p-value proporcionado por R fue menor que 2.2x10-16, que está por debajo del umbral de aceptación de la hipótesis nula. Entonces podríamos concluir que la distribución de frecuencias en las clases estructurales de ambos subgrupos de proteínas es diferente. El listado completo de proteínas con sus clases y folds funcionales se pueden consultar en la Información Suplementaria S2 de este trabajo.

Organismos en MultitaskprotDB-II

Otra estadística interesante de la base de datos de proteínas moonlighting es la abundancia de organismos en ella. Como se puede ver en la Figura 15, el organismo más abundante es *Homo sapiens* con 186 proteínas moonlighting en nuestra base de datos, ese hecho ya se esperaba porque es el organismo más investigado. Los siguientes organismos son *Streptococcus, E. coli* y *Saccharomyces cerevisiae*. Un 25% de las proteínas moonlighting de la base de datos estan involucradas en la virulencia de microorganismos patógenos, como se analizará más adelante en este trabajo.

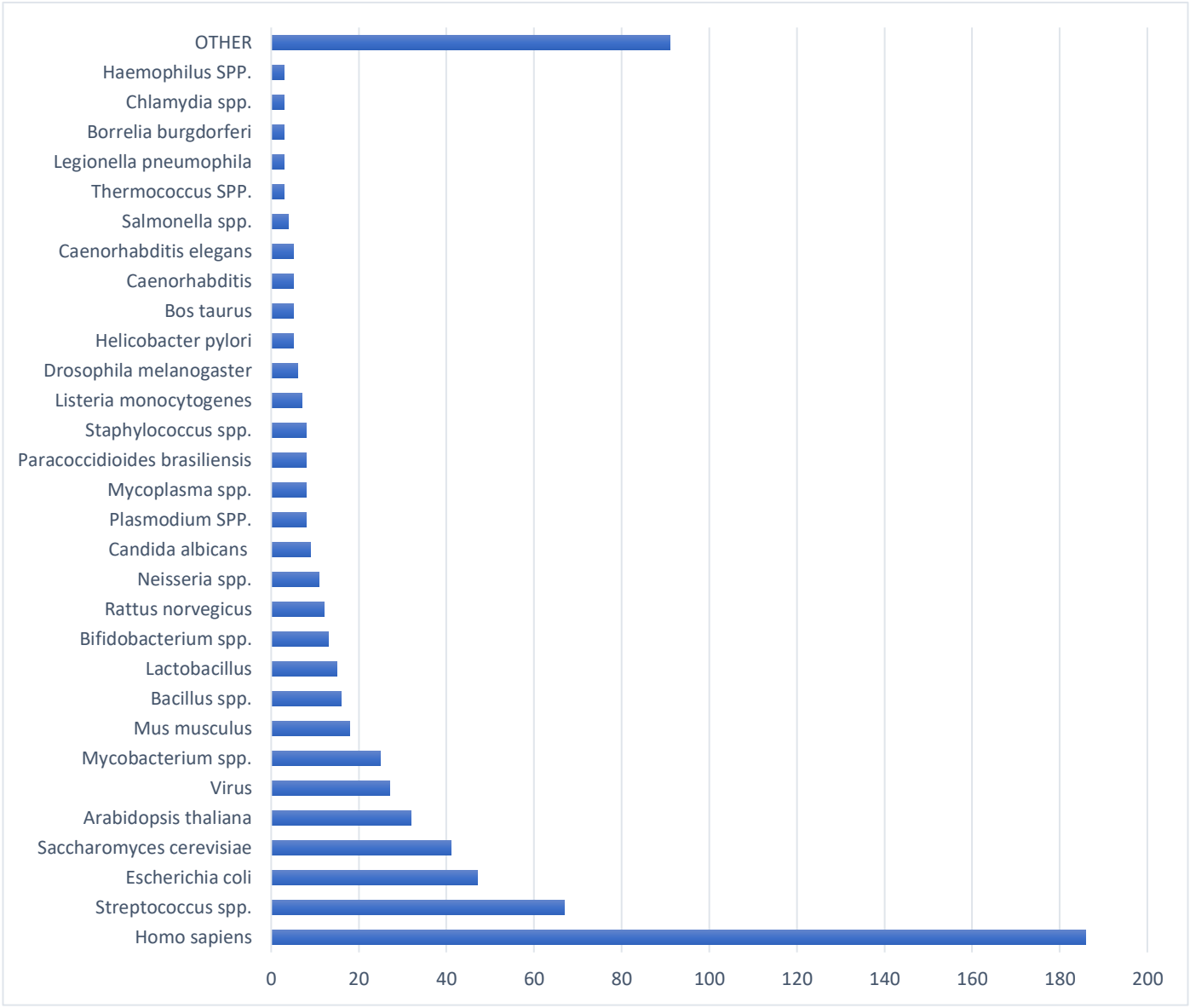


Figura 15: Principales organismos presentes en la base de datos de proteínas moonlighting ordenado por el número de proteínas moonlighting detectadas.

En relación a las funciones de las proteínas moonlighting, hemos analizado y emparejado funciones de todas ellas en nuestra base de datos, resultando que, en organismos Eucariotas, el ser enzima y factor de transcripción es la clase más abundante, seguida por presentar dos actividades enzimáticas diferentes. En el caso de los organismos Procariotas, lo más común es ser una enzima y una proteína de adhesión involucrada en virulencia. Como se muestra en las siguientes Tablas 4 y 5.

Tabla 4: Porcentajes de pares funcionales en las proteínas moonlighting respecto la cantidad total de proteínas en nuestra base de datos

FUNCTION 1	FUNCTION 2	% OF PROTEINS
Enzyme	Transcription factor	22,9%
Enzyme	Adhesion protein	16,0%
Enzyme	Enzyme	14,2%
Enzyme	Structural protein	11,8%
Transcription factor	Transcription factor	9,4%
Chaperone	Cytokine activator	2,8%
Structural protein	Structural protein	2,4%
Enzyme	Enzyme inhibitor	2,1%
Receptor	Receptor	1,7%
Enzyme	Chaperone	1,7%
Enzyme activator	Enzyme inhibitor	1,7%
Structural protein	Adhesion protein	1,7%
Enzyme	Cytokine	1,4%
Enzyme	Apoptosis	1,4%
Enzyme	Signalling	1,4%
Transcription factor	Structural protein	1,0%
Transcription factor	Cytokine	1,0%
Enzyme	Virulence factor (not adhesion)	0,7%
Chaperone	DNA binding	0,7%
Chaperone	Toxin	0,7%
Transcription factor	Adhesion protein	0,7%
Structural protein	DNA binding	0,7%
Receptor	Enzyme	0,7%
Chaperone	Biofilm protein	0,3%
Structural protein	Membrane proteins	0,3%
Enzyme	Halotolerance	0,3%

Tabla 5: Porcentaje de pares funcionales en proteínas moonlighting Procariotas listadas en nuestra base de datos

FUNCTION 1	FUNCTION 2	% OF PROTEINS
Enzyme	Adhesion protein	45,1%
Enzyme	Transcription factor	17,6%
Chaperone	Cytokine activator	7,8%
Enzyme	Enzyme	2,9%
Transcription factor	Transcription factor	6,9%
Enzyme	Structural protein	4,9%
Enzyme	Chaperone	2,9%
Enzyme	Apoptosis	2,0%
Chaperone	Toxin	2,0%
Enzyme	Virulence factor	2,0%
Transcription factor	Structural protein	1,0%
Enzyme	Enzyme inhibitor	1,0%
Receptor	Receptor	1,0%
Enzyme	Cytokine	1,0%
Structural protein	DNA binding protein	1,0%
Chaperone	Biofilm protein	1,0%

Como se puede ver en ambas figuras, la función moonlighting más común en proteínas de organismos Eucariotas y Procariotas son las enzimas. Las principales diferencias entre Eucariotas y Procariotas es que la segunda función en los primeros es de unión a DNA y actuar como factor de transcripción y en los segundos son funciones relacionadas con la virulencia de microorganismos patógenos que son el 45% de todas las proteínas moonlighting de Procariotas. Además, este porcentaje aumenta cada día debido al descubrimiento de factores de virulencia en diferentes organismos. Algunas proteínas actúan como moonlighting en diversas especies. Este hecho será discutido en la Sección IV.E. y esto nos hace pensar que la multifuncionalidad se conserva entre proteínas similares en diferentes especies.

IV.B. PREDICCIÓN E IDENTIFICACIÓN DE PROTEÍNAS MOONLIGHTING Y MAPADO DE DOMINIOS FUNCIONALES

Todos estos resultados provienen del análisis de las proteínas multifuncionales contenidas en la base de datos MultitaskProtDB-II. Como se mencionó anteriormente, las funciones han sido etiquetadas como canónica (la primera históricamente identificada) y moonlighting (o multitasking o multifuncional), las posteriores. La función canónica suele corresponder también con la función más divulgada, pero esto no implica relevancia biológica y simplemente se refiere al orden histórico del descubrimiento de la función biológica. Existen diversas formas de asignar una función a una secuencia de proteína cuya función es desconocida, pero los métodos más utilizados usan la propiedad transitiva. Si una proteína de función desconocida tiene un suficiente grado de similitud con otra proteína con diferente función, entonces se supone que comparten la misma función. Si además tenemos mucha información redundante (es decir, un gran conjunto de secuencias relacionadas), esta redundancia se puede usar para inferir funciones extrayendo patrones o perfiles. En este caso, podemos usar estos patrones (los llamaremos "motifs" y dominios) para inferir la función. El patrón extraído también se puede utilizar para identificar aminoácidos esenciales para la función. Otra forma de identificar aminoácidos importantes para la función de la proteína es mediante el modelado de su estructura tridimensional. Finalmente, a partir del desarrollo de los métodos de interactómica, identificar los partners de interacción compartidos también permite sugerir función (Espadaler et al., 2005, 2008; Gomez et al., 2008). Todas estas estrategias (alineamiento de secuencias, identificación de motivos, partners de interacción, modelado 3D y sus combinaciones) se han utilizado en este trabajo para inferir la función de las proteínas multifuncionales.

IV.A.1. ANÁLISIS MEDIANTE HOMOLOGIA REMOTA

Blast, y especialmente PSI-Blast, puede detectar proteínas multifuncionales, en especial aquellas que provienen de la fusión de dos proteínas o dos dominios diferentes. El algoritmo de homología remota PSI-Blast es especialmente adecuado para identificar proteínas moonlighting porque puede identificar tramos de residuos de aminoácidos de diferentes dominios, aunque no estén aparentemente conservados secuencialmente y que se puedan relacionar con diferentes funciones biológicas (Gomez et al., 2003; Khan et al., 2012). Al igual que en las búsquedas de bases de datos de PPI (ver siguiente apartado), se muestran una gran lista de resultados y el investigador no sabe a priori, cuáles de ellos serán verdaderos positivos, y es el análisis cuidadoso de las diferentes predicciones y de los datos experimentales el que puede sugerir un verdadero positivo. Se utilizó PSI-Blast y ByPass con las proteínas moonlighting listadas en la base de datos MultitaskProtDB-II según los métodos descritos en la Sección III de Métodos para analizar si estos métodos son capaces de detectar las proteínas moonlighting descritas y con qué eficacia lo hacen.

La columna 4 de la Tabla 6 muestra algunos ejemplos de proteínas moonlighting identificadas por homología remota. En la Información Suplementaria S3 de la tesis y de la publicación en Frontiers in Bioengineer. & Biotechnol. se muestran más ejemplos.

Tabla 6: Ejemplos de proteínas moonlighting identificadas combinando homología remota e interactómica

CANONICAL FUNCTION	MOONLIGHTING FUNCTION	PPI partners (only some hits are shown)	PSI-Blast/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

También se han utilizado los programas de búsqueda de motifs/dominios funcionales (Prosite, Pfam, etc) contenidos en el servidor InterPro. La Figura 16 muestra dos predicciones utilizando estos programas. Hay que resaltar que debido al rediseño ("curado") que los creadores de tales programas realizan, con objeto de mejorar la predicción del motif/dominio "canónico", como en la Figura 16(a), se pierden aquellos que sugerirían una segunda función. Por ello,

programas no curados, como PfamB, o no actualizados, como Blocks, suelen ser más útiles para la identificación de proteínas moonlighting como muestra la Figura 16(b). Finalmente, merece ser mencionado que solo el 10% de las proteínas moonlighting de la base de datos son identificadas por PSI-Blast e InterPro al mismo tiempo, lo que se explica por el relativamente escaso número de motifs y dominios descritos (unos 1300).

IV.A.2. BÚSQUEDA EN BASES DE DATOS DE INTERACTÓMICA

El grupo ha propuesto anteriormente que las bases de datos de interacción proteína-proteína (PPI) combinadas con análisis de similitud de secuencia pueden ayudar a predecir la función proteica (Espadaler et al., 2005, 2008) y que las bases de datos PPI también deben contener información sobre las proteínas moonlighting y sugerencias de análisis adicionales para probar sus funciones moonlighting (Gomez et al., 2011; Hernández et al., 2014, 2015). Los partners de interacción de una proteína podrían sugerir la función o funciones de una proteína ("guilty-by-association"), al menos en el nivel del proceso biológico de GO. Consideramos que las bases de datos de interactómica revelan correctamente una segunda función para la proteína moonlighting si la base de datos PPI identifica una Función Molecular o, en algunos casos, un Proceso Biológico según la anotación Gene Ontology, que esté de acuerdo con la función adicional descrita en nuestra base de datos. Luego, para filtrar resultados y mejorar la precisión, es aconsejable realizar un análisis de enriquecimiento de Gene Ontology. Para cada proteína moonlighting incluida la base de datos de interactómica APID, se recopilaron los términos GO de los partners de interacción y se calculó el enriquecimiento del término GO mediante el paquete GOStat R (Beissbarth y Speed, 2004). Esta función calcula p-values hipergeométricos para la sobrerrepresentación de cada término GO en la categoría especificada entre las anotaciones GO para las funciones de interés. Seleccionamos como verdaderos indicadores de función moonlighting estos términos GO con un p-value menor a 0.05; este umbral también nos permite eliminar descripciones inespecíficas de GO (Gomez et al., 2011). La columna 3 de la Tabla 6 muestra algunos ejemplos de identificación de funciones

moonlighting con PPI. Debido a que el número de partners de interacción encontrados en las bases de datos PPI puede ser alto, seleccionar los verdaderos partners no es una tarea fácil si el investigador no tiene pistas adicionales. La lista de resultados debe reducirse adecuadamente al tener en cuenta otras predicciones bioinformáticas como se describe a continuación, o con la ayuda de datos experimentales o clínicos que sugieran correlaciones útiles. En este sentido, hemos encontrado que, al combinar la información de la base de datos PPI y de las búsquedas remotas de homología, la predicción de moonlightings es altamente mejorada. Un problema adicional es que muchas especies no han sido analizadas por interactómica; por lo tanto, la mayoría de las proteínas de MultitaskProtDB-II no tienen partners proteicos en las bases de datos PPI.

En nuestra opinión, el límite principal del nivel de predicción de proteínas moonlighting utitlizando las bases de datos de PPI se debe principalmente a la baja sensibilidad de la interactómica (es decir, muchos falsos negativos) más que a la baja especificidad (es decir, falsos positivos).

IV.A.3. COMBINACIÓN DE ANÁLISIS DE HOMOLOGIA REMOTA CON INTERACTÓMICA

Buscamos proteínas de nuestra base de datos que tengan partners de interacción en la base de datos de PPI, APID (Alonso-López et al., 2016). Como se indicó anteriormente, cada proteína moonlighting puede presentar una gran cantidad de supuestos partners de interacción y también una gran cantidad de presuntos homólogos remotos en el algoritmo PSI-Blast. Hemos inspeccionado manualmente ambos tipos de resultados para verificar la intersección de ambos conjuntos y así limitar la lista de candidatos y mejorar la predicción de las proteínas moonlighting conocidas. Esta inspección manual ha sido necesaria porque hay un problema relacionado con los diferentes descriptores de anotación representados por los dos tipos de resultados. La mayoría de las salidas Blast/PSI-Blast de los alineamientos de secuencia no informan de anotaciones semánticas, mientras que muchas bases de datos PPI usan anotaciones GO. Este hecho complica la automatización en la salida de datos.

Sugerimos tomar como posibles coincidencias positivas a las que describen una función en cualquier posición del resultado PSI-Blast/ByPass que corresponde a un partner de la base de datos de PPI, como se muestra en los ejemplos de la Tabla 6, columnas 3 y 4. Ahora el grupo está diseñando un programa que es capaz de hacer coincidir automáticamente dos o más salidas.

IV.A.4. BÚSQUEDA DE MOTIFS O DOMINIOS ESPECÍFICOS DE FUNCIÓN

La búsqueda de diferentes motifs/dominios vinculados a diferentes funciones en una secuencia de proteína diana utilizando InterPro debería, en principio, ayudar a identificar las proteínas moonlighting. Sin embargo, hay dos problemas principales: (a) el número relativamente bajo de dominios y motifs actualmente conocidos y (b) la versión actual de programas como Prosite, etc., que se han diseñado para una predicción más precisa de los motifs/dominios más comunes, pero no identifica dominios comunes. Esto explicaría el hecho de que el uso de InterPro con las proteínas de nuestra base de datos identifique la función canónica para aproximadamente el 80% de ellas, pero para la función moonlighting en solo el 10% de los casos. Por ejemplo, el clásico ejemplo de proteína moonlighting es la Gliceraldehído-3-fosfato deshidrogenasa/uracil glicosilasa (GAPDH/UDG), la salida PSI-Blast revela ambas funciones con puntuaciones altas, pero InterPro solo identifica un motif para la función canónica (GAPDH) de esta proteína. Sin embargo, ambas funciones están identificadas por Blocks. En el caso de la proteína Arg 2, Blocks identifica las funciones canónicas y moonlighting como las dos puntuaciones más altas de la salida (Figura 16A).

A **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 E-value	Family		Strand	Blocks
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

B **>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 16: Dos ejemplos de las salidas de programas de identificación de motifs/dominios (A) El servidor Blocks identifica ambas funciones de la proteína Arg2 en las posiciones superiores de la salida. (B) El programa ProDom muestra dos dominios relacionados con las funciones de la aconitase, tanto canónica como moonlighting.

El hecho de que la detección de patrones de una función secundaria por un programa que no se ha actualizado desde 2006 sea mejor que el uso de métodos más modernos y refinados, nos hizo pensar que este fenómeno puede deberse a un problema entre sensibilidad y especificidad. Las herramientas de detección de patrones se han desarrollado tradicionalmente para tener una buena relación entre especificidad y sensibilidad. Cuando se construye un conjunto de datos estándar para entrenar estas aplicaciones, generalmente se supone que todas las proteínas incluidas en la base de datos tienen solo una función. 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 sensibilidad, de modo que las herramientas tienden a detectar un número bajo de funciones secundarias. En este sentido, la tendencia de usar secuencias muy curadas para construir estos patrones podría explicar por qué herramientas no actualizadas,

como Blocks, son más efectivas para detectar funciones secundarias. Si es así, esto indicaría que, para detectar tales funciones secundarias, herramientas como Blast o PSI-Blast pueden ser más apropiadas porque no dependen de la preexistencia de patrones previamente construidos con una semilla limitada. Pero esto también puede deberse a otros factores, como el hecho de que muchas herramientas nuevas, además de un conjunto de proteínas con función conocida, incorporan un conjunto de falsos positivos (las secuencias comparten motifs que no tienen una función asignada). Este conjunto contiene proteínas que contienen el patrón asociado con la función de la proteína, pero en realidad no realizan esta función. Para verificar si algunos de los falsos positivos se descartan erróneamente para las funciones moonlighting secundarias, hemos comparado todas las secuencias de falsos positivos en la base de datos de Prosite con nuestra base de datos de proteínas multifuncionales. Luego, verificamos 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 secundaria de estas proteínas multifuncionales. Este cálculo nos llevó a concluir que, al menos para Prosite, los falsos positivos son verdaderos falsos positivos, porque ninguna de esas funciones coincide con la función secundaria de la proteína.

Pfam es uno de los programas que contiene InterPro y está basado en modelos ocultos de modelos de Markov (HMM). Las familias Pfam (dominios de proteínas agrupados mediante HMM) se basan en alineamientos múltiples de secuencias, agrupados en dominios Pfam. La actividad biológica de estas familias podría describirse como dominios múltiples que cumplen juntos la función principal. Los límites de estos dominios están mejor establecidos que en la familia. Estas características podrían implicar que los dominios Pfam serían una mejor herramienta para identificar las funciones moonlighting. Nuestros resultados muestran que los dominios Pfam son cuatro veces más efectivos en la detección de la función moonlighting que otros métodos, pero la importancia estadística de esta diferencia es baja, el p-value proporcionado por un test $\chi 2$ es 0.02.

Otra consideración importante es que parte de la mejora en la predicción de la función moonlighting por Pfam se debe a la información adicional de la función de dominio dada como documentación complementaria.

Otro punto que hemos explorado es la diferencia entre las bases de datos PfamA y PfamB. PfamA es una base de datos curada manualmente que contiene un conjunto de HMM de más de 14,000 familias. La base de datos PfamB se construye automáticamente con grupos de secuencias producidas por el algoritmo ADDA (Heger et al., 2005), y sus familias generalmente provienen de alineamientos que contienen proteínas con funciones bastante heterogéneas. Esta característica nos animó a probar si PfamA y PfamB son herramientas apropiadas para predecir funciones secundarias. Probamos ambas versiones utilizando el conjunto de proteínas presentes en nuestra base de datos. PfamA predice el 78% de las funciones canónicas, pero solo el 6% de las funciones moonlighting. Con PfamB, encontramos 58 proteínas del conjunto de proteínas de nuestra base de datos que tienen una 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 moonlighting. Sin embargo, este método es difícil de automatizar, ya que el número de anotaciones que se probarán es muy alto, incluso seleccionando los mejores ítems previamente. Hemos realizado una breve lista de anotaciones en cada familia PfamB, priorizando secuencias de cadenas más largas con respecto a las cortas incluidas en las secuencias originales usadas para generar familias PfamB.

También es destacable que PfamA no revela cerca del 80% de las proteínas moonlighting identificadas por PfamB. Obviamente, si tenemos una ligera idea de la función de la proteína, la exploración de la salida de PfamB puede proporcionar sugerencias sobre el proceso de encontrar la función secundaria de nuestra proteína.

IV.A.5. LOCALIZACIÓN DE LAS FUNCIONES CANÓNICAS Y MOONLIGHTING EN LA SECUENCIA/ESTRUCTURA DE LA PROTEÍNA

La predicción de proteínas moonlighting es un trabajo difícil, pero mucho más lo es ir más allá y encontrar un método de predicción que también permita mapar bioinformáticamente las diferentes funciones en la estructura 3D de la proteína, excepto en aquellos casos en que se identifica un motif/dominio específico. Solo

unas pocas proteínas moonlighting tienen sus funciones localizadas, se representan en la siguiente Tabla 7.

Tabla 7: Proteínas moonlighting en las que ambas funciones se han localizado en la estructura de la proteína de un total de 694

PROTEIN NAME	ORGANISM	UNIPROT
DESCRIBED IN THE BIBLIOGRAPHY		
Apartate receptor	*E. coli*	J7R5H8
Peroxiredoxin-6	*Homo sapiens*	P30041
Fatty acid multifunctional protein	*Arabidopsis thaliana*	Q9ZPI5
Aldolase	*Oryctolagus cuniculus*	P00883
DLD (D-Lactate dehydrogenase)	*Homo sapiens*	Q86WU2
Enolase	*S. pneumoniae*	Q97QS2
Citocromo P450	*Streptomyces coelicolor*	Q9K498
S10 ribosomal protein	*E. coli*	P0A7R5
RfaH	*E. coli*	P0AFW0
MAPK1	*Homo sapiens*	P28482
Malate synthase	*Mycobacterium tuberculosis*	P9WK17
BirA	*E. coli*	P06709
GAPDH	*E. coli*	P0A9B2
Phosphoglucose isomerase	*Homo sapiens*	P06744
MAPPED USING PISITE AND SEQUENCE OR 3D ALIGNMENTS		
PPI	*Helicobacter pylori*	D0K1B1
NK tumor recognition factor	*Human*	P30414
Fatty acid multifunctional protein	*Arabidopsis thaliana*	Q9ZPI5
Molybdopterin biosynthesis enzyme	*Arabidopsis thaliana*	Q39054
Hal3 (Halotolerance protein HAL3)	*Saccharomyces cerevisiae*	P36024
Aspartate-kinase–homoserine-dehydrogenase	*Arabidopsis thaliana*	Q9SA18
Histidine biosynthesis bifunctional protein, chloroplastic	*Arabidopsis thaliana*	O82768
Dihydrofolate-reductasethymidylate-synthase	*Arabidopsis thaliana*	Q05762
6-Hydroxymethyl-7,8-dihydropterinpyrophosphokinase–7,8-dihydropteroate-synthase	*Arabidopsis thaliana*	Q1ENB6
Formiminotransferase	*Thermoplasma acidophilum*	Q9HI69

RESULTADOS. PREDICCIÓN.

Existen diferentes algoritmos y programas para la predicción y el modelado de estructuras que pueden dar pistas sobre sitios funcionales en secuencias de proteínas (PiSite, Phyre2, I-Tasser, SiteEngine ...). Hemos aplicado tres de ellos: PiSite (Higurashi et al., 2009), Phyre2 (Kelley y Sternberg et al., 2009) e I-Tasser (Wang et al., 2017) a las proteínas moonlighting de la base de datos MultitaskProtDB-II. La principal limitación de PiSite es que requiere que la proteína de consulta, o un dominio de la misma, tenga una secuencia de aminoácidos significativamente similar a una estructura contenida en la base de datos PDB. Aunque la secuencia de aminoácidos se usa en la mayoría de los métodos de predicción de función, la función de una proteína la realiza realmente la proteína en su conformación nativa, en otras palabras, por su estructura terciaria o incluso cuaternaria. Se sabe que la estructura terciaria de una proteína está más conservada que la secuencia en relación con la función. Para llevar a cabo este análisis, modelamos las proteínas de nuestra base de datos utilizando Phyre2, únicamente se seleccionaron aquellos modelos con un alto grado de confianza y un porcentaje aceptable de identidad de secuencia. Las funciones de las proteínas obtenidas en Phyre2 se compararon con las funciones canónicas y moonlighting de la proteína moonlighting modelada, para comprobar si Phyre2 identificaba ambas funciones, es decir, encontraba proteínas cuyas funciones encajaban con las funciones de la proteína analizada.

Para la mayoría de las proteínas, se identificó la función canónica, en cambio se identificó la moonlighting en solo el 10% de los casos. La tasa de identificación de la función moonlighting no es alta, pero los resultados son similares a los obtenidos por otras técnicas como Pfam. En el caso de PiSite, el programa identificó 266 coincidencias en PDB de un total de 288 proteínas presentes en nuestra primera base de datos MultitaskProtDB y, además de la función canónica, identifica la función moonlighting para 28 proteínas de MultitaskProtDB, por lo tanto, la tasa de detección es similar a la de los otros métodos descritos anteriormente, un 10%. La Figura 17A muestra un ejemplo de una coincidencia para una proteína moonlighting con dos funciones HPPK y DHPS (donde PiSite identifica las funciones tanto canónicas como moonlighting). Además, utilizando la herramienta SwissPDBViewer (Guex y Peitsch et al., 1997), ambas funciones se pueden mapar estructuralmente con un buen RMSD.

PiSite por sí solo no puede identificar tantas proteínas multifuncionales como la combinación de bases de datos PSI-Blast y PPI, pero puede usarse para apoyar supuestos positivos y considerarlos verdaderos positivos después de ejecutar esos programas, y sugiere una ubicación en la estructura tridimensional de la función moonlighting. Además, puede sugerir un origen evolutivo de la doble función como resultado de una fusión gen/dominio.

También en el caso de Phyre2, la tabla de la parte superior de la Figura 17 muestra que el programa se puede usar para modelar e identificar los dominios de la proteína correspondientes a ambas funciones. En esta misma figura, se muestra un ejemplo de utilización de este método de mapado utilizando los programas PiSite y Phyre2 y alineando las proteínas encontradas utilizando el programa SwissPDBViewer con un RMSD de 5,08 y 5,32 Å respectivamente. Este RMSD permite afirmar que las proteínas encontradas se superponen estructuralmente de forma significativa.

Color	Protein name	Uniprot	Functions
Red	6-hydroxymethyl-7,8-dihydropterin pyrophosphokinase/ 7,8-dihydropteroate synthase *(predicted)	Q1ENB6	DHPS / HPPK
Black	Folic acid synthesis protein FOL1 (fragment)	P53848	DHPS / HPPK
Blue	Dihydropteroate synthase	Q81VW8	DHPS
Green	2-amino-4-hydroxy-6-hydroxymethyldihydropteridine pyrophosphokinase	P26281	HPPK

*Predicted 3D structure using PHYRE server. Confidence=100, Identity=35%.

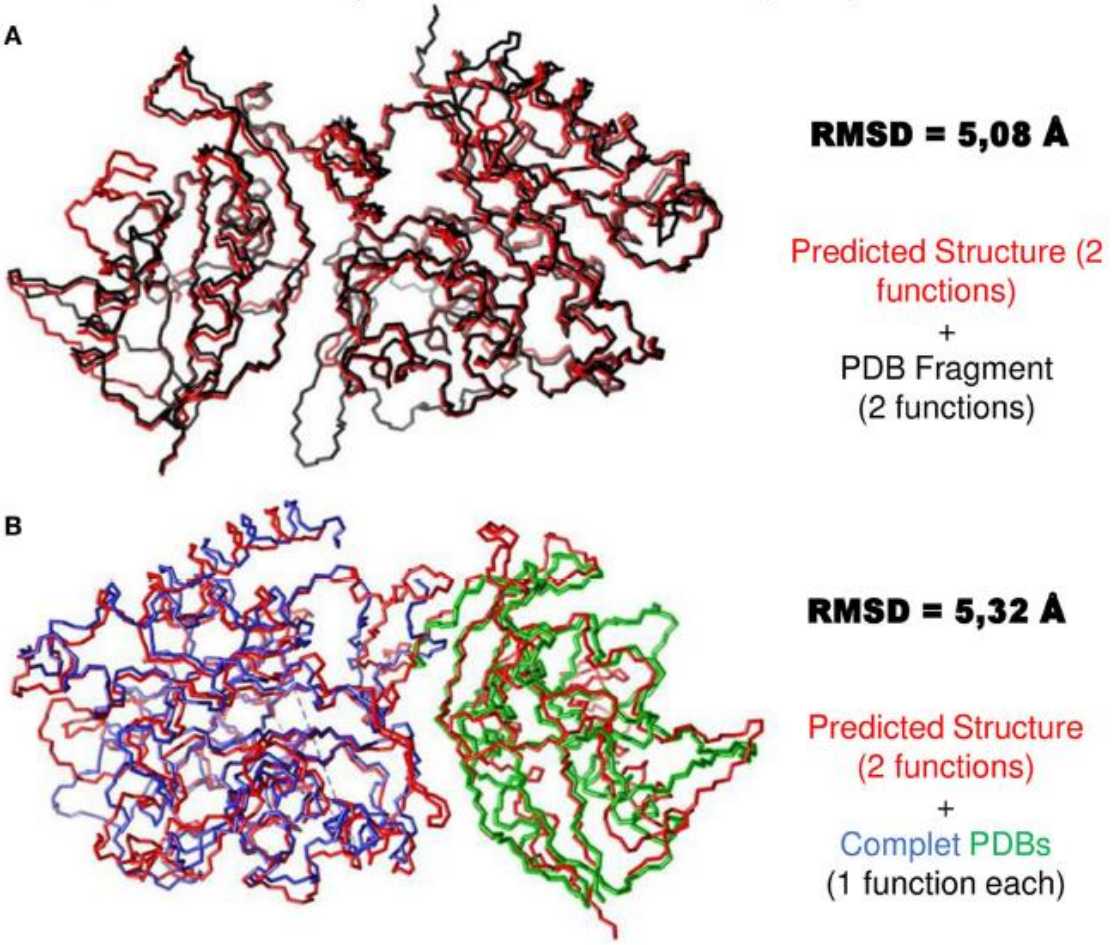


Figura 17: Ejemplo de mapado de las funciones canónica y moonlighting a partir del modelado estructural de la proteína mediante Phyre2. Las proteínas de la figura se utilizaron para hacer una comparación de estructura usando SwissPDBViewer y USCF Chimera. La estructura 3D de la proteína "roja" moonlighting se predijo utilizando Phyre2, mientras que las otras estructuras se encontraron en el PDB. En (A), se corroboró la similitud de secuencia encontrada previamente. En (B), la superposición de la estructura de las tres proteínas alineadas enfatiza la utilidad de estos métodos para mapar las dos funciones de una proteína moonlighting.

Otro ejemplo de una proteína, en este caso identificada por PiSite como una supuesta proteína moonlighting es la Fatty Acid Multifunctional Protein (MFP) que se muestra en la Figura 18. En este caso, MFP tiene dos funciones y PiSite encontró dos proteínas diferentes. Cada una de estas proteínas superpone en una parte diferente de la proteína multifuncional que permite ubicar dónde se encuentran las funciones canónicas y moonlighting.

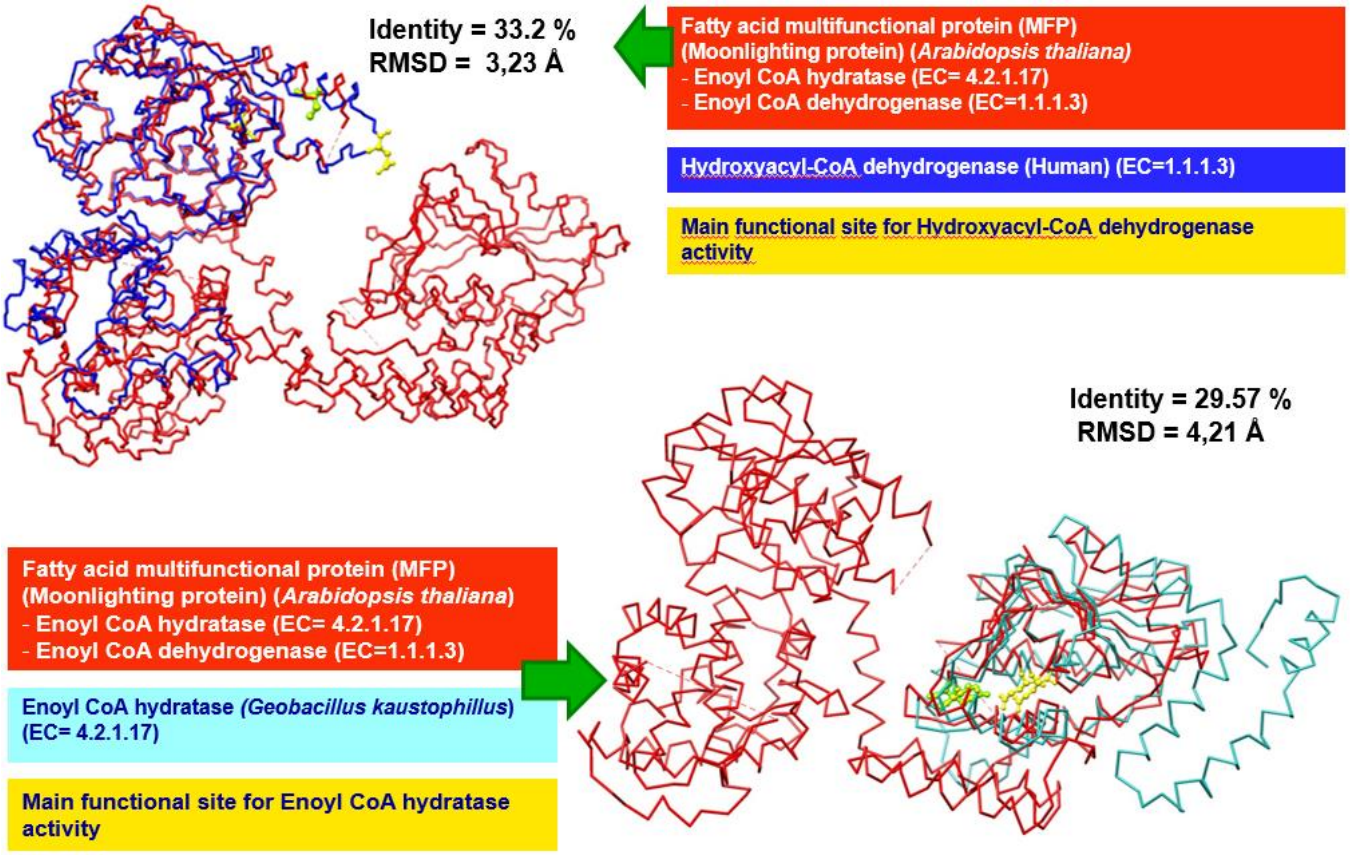


Figura 18: Identificación, modelado y mapado de las dos funciones, canónica y moonlighting, de la proteína MFP. Puede apreciarse que se encontró un nivel significativo de similitud estructural con dos estructuras PDB ya resueltas que representan las áreas de funciones canónicas y moonlighting. Esto permite la superposición estructural y es muy útil mapar ambas funciones. Los residuos activos clave implicados en las funciones canónica y moonlighting se representan en formato de "ball and stick" y de color amarillo. La proteína en rojo es la supuesta proteína moonlighting con dos funciones, MFP. Hidratasa y deshidrogenasa. La proteína en azul tiene solo la función deshidrogenasa y superpone perfectamente en una sola región de la proteína multifuncional y la proteína en azul claro tiene solo la función hidratasa y superpone en otra región de la proteína multifuncional.

Se realizó un análisis de todas las proteínas moonlighting en MultitaskProtDB-II con PiSite. Algunos de los resultados se muestran en la Tabla 8. Por ejemplo, la proteína Leukotriene A4 hydrolase tiene dos funciones diferentes: biosíntesis del

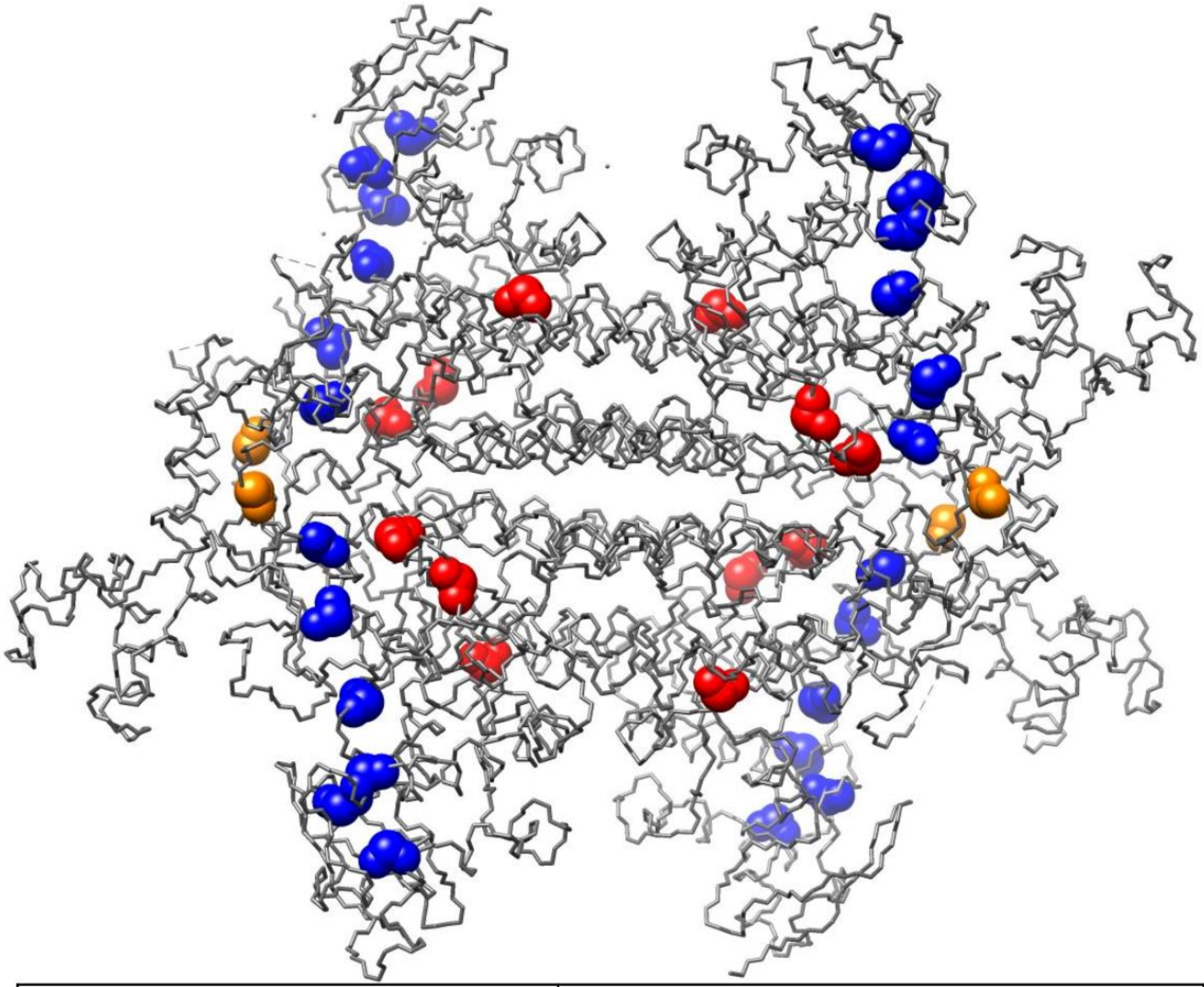
mediador proinflamatorio leukotriene B4 (función canónica) y actúa como aminopeptidasa (moonlighting) y PiSite encontró que la proteína Cold-active aminopeptidase es similar a Leukotriene A4 hidrolasa y su función coincide con la función moonlighting de la proteína moonlighting.

Tabla 8: Algunos ejemplos de análisis mediante el programa PiSite en los que se puede ver la que la función de la proteína encontrada por PiSite coincide con la función moonlighting de la proteína analizada

Protein name	Canonical Function	Moonlighting Function	Protein found by PISITE analysis	Function
Fatty acid multifunctional protein (MFP)	Enoyl CoA hydratase	Enoyl CoA dehydrogenase	Enoyl CoA dehydrogenase (Q5KYB2)	Enoyl CoA dehydrogenase
Gephyrin	Microtubule associated protein.	Synthesis of molybdenum cofactor.	Molybdopterin molybdenumtransfe-rase. (P12281)	Synthesis of molybdenum cofactor.
Leukotriene A4 hydrolase	Biosynthesis of the proinflammatory mediator leukotriene B4	Aminopeptidase	Cold-active aminopeptidase (Q7WVY1)	Aminopeptidase
Peptidyl Prolyl cis,trans-Isomerase	Accelerate the folding of proteins	Induces apoptosis. Activates monocyte IL-6 synthesis.	NK-tumor recognition protein (P30414)	Involved in the function of NK cells.
ERK2	Mitogen-activated protein kinase 1	Transcriptional repressor	Cell Division Control protein 2 (Q07785)	Phosphorylates the repetitive C terminus of RNA polymerase II
Hal3 (Halotolerance protein HAL3)	Inhibitor of Protein Phosphatase PPZ1	Coenzyme A biosynthesis.	Phosphopantothenoyl-cysteine decarboxylase (Q96CD2)	Necessary for the biosynthesis of coenzyme A.
Formimidoyl transferase - cyclodeaminase	Formimidoyl transferase activity	Cyclodeaminase activity	Cyclodeaminase (Q9HI69)	Cyclodeaminase activity
Bifunctional aspartokinase homoserine dehydrogenase 1, chloroplastic	Aspartokinase activity	Homoserine dehydrogenase activity	Subtilin transport ATP-binding protein SpaT (P31116)	Homoserine dehydrogenase activity
Bifunctional dihydrofolate reductase, thymidylate synthase 1	Dihydrofolate reductase activity	Thymidylate synthase activity	Thymidylate synthase (P45352)	Thymidylate synthase activity

En la Información Suplementaria S4 de este trabajo se muestran más ejemplos de resultados positivos en los que el programa PiSite identifica, por alineamiento estructural, las funciones moonlighting. Como ya se ha mencionado antes, la principal limitación del método es que numerosas proteínas carecen de estructura 3D en la base de datos PDB. En general, en la bibliografía se describe la identificación de una nueva proteína moonlighting sin vincular (mapar) cada función a una región o dominio específico de la proteína. Incluso una de las principales proteínas moonlighting, la Glyceraldehyde-3-phosphate dehydrogenase, no ha sido mapada funcionalmente (excepto para su función canónica) (Sirover et al., 2014). Por lo tanto, sería muy útil localizar cada función, en la medida de lo posible, en la secuencia/estructura de la proteína. Una aplicación importante de este mapado funcional sería para relacionar cada función con enfermedades humanas. Por ejemplo, en el caso de la Fumarase hidratase.

Esta proteína está asociada con la Fumarase deficiency (FD) y la Hereditary leiomyomatosis and renal cell cancer (HLRCC). La base de datos UniProt informa mutaciones de proteínas involucradas en estas enfermedades. La Figura 19 muestra la estructura 3D del tetrámero, así como las mutaciones relacionadas con FD y HLRCC, que se representan en rojo y azul, respectivamente. En amarillo, se resaltan las mutaciones asociadas con ambas enfermedades. Esta imagen muestra claramente que la función canónica, que está relacionada con FD, se encuentra en el centro del tetrámero, mientras que la función moonlighting, que está relacionada con HLRCC, está en una región de proteína diferente. La estructura y los mutantes de aminoácidos sugieren las bases moleculares de la enfermedad, porque estas mutaciones perturbarían la interacción y la formación correcta del tetrámero, que a su vez puede cambiar indirectamente, en cierto grado, el posicionamiento preciso de los aminoácidos del centro activo, reduciendo su actividad.



Mutations	Associated disease
A308T, F312C, D425V, Q185R, R233H, G282V	Hereditary leiomyomatosis and renal cell cancer
N107T, A117P, H180R	Fumarase deficiency
K230R	Both diseases

Figura 19: Estructura tetramérica de la proteína Fumarate hydratase humana (P07954). Esta proteína tiene dos enfermedades asociadas: deficiencia de fumarasa (FD) y leiomiomatosis hereditaria más cáncer de células renales (HLRCC). Marcados en rojo están las mutaciones relacionadas con FD, y en azul las relacionadas con HLRCC. Las mutaciones asociadas con ambas enfermedades están en amarillo.

Tal y como se ha mencionado en el apartado de Métodos, estos métodos son útiles en aquellos casos en que las funciones adicionales son debidas a la incorporación de otro dominio, es decir, en aquellas proteínas moonlighting en que ambas funciones estan en dominios claramente separados. Solo algunas de las proteínas moonlighting son de este tipo y además hay pocas proteínas cuya estructura tridimensional haya sido descrita y publicada. Por tanto, un 10% de predicción de proteínas moonlighting utilizando métodos estructurales, es un valor relativamente positivo. Como se muestra en los ejemplos anteriores, el mapado de cada función en la estructura 3D de una proteína moonlighting podría ser muy útil para la comprensión de sus funciones normales y patológicas. Además, permite tener una visión más clara de las bases moleculares de las enfermedades y puede ser beneficioso para el diseño de fármacos específicos o incluso para el conocido “Drug Repositioning”.

IV.C. RELACIÓN DE LAS PROTEÍNAS MOONLIGHTING CON ENFERMEDADES HUMANAS Y DIANAS FARMACOLÓGICAS

IV.C.1. PROTEÍNAS MOONLIGHTING Y ENFERMEDADES HUMANAS

Hemos realizado análisis para encontrar cuántas proteínas moonlighting están relacionadas con enfermedades humanas conocidas. Tal y como se ha descrito en el apartado III.F.3, el análisis de la información contenida en UniProt y OMIM en relación a las enfermedades asociadas con las proteínas moonlighting humanas presentes en MultitaskProtDB-II muestra que el 78% de ellas están relacionadas con enfermedades. Estas proteínas, junto con una descripción de sus funciones y las enfermedades en las que están involucradas, se han recopilado en la Información Suplementaria S5 que se adjunta a este trabajo.

El número de proteínas humanas indicadas como entradas revisadas en la base de datos UniProt es de 20.168 en el momento del estudio. Por otra parte, en la base de datos OMIM vimos que solo 3.600 de estas proteínas están relacionadas con enfermedades humanas, lo cual representa un porcentaje del 17.85%. Además, verificamos en OMIM y en la bibliografía si las proteínas de MultitaskProtDB-II estaban relacionadas con enfermedades humanas, y se encontró un número sorprendente que muestra que el 78% de las proteínas analizadas están implicadas en enfermedades humanas. Este porcentaje es mucho más alto que el 17.85% encontrado en las proteínas humanas en general y es claramente significativo como puede verse en la Figura 20. Las probabilidades de que una proteína humana presente en UniProt y una proteína humana de nuestra base de datos MultitaskProtDB-II estuvieran implicadas en una enfermedad OMIM conocida, se calcularon usando la relación ODD. En conjunto, estos resultados sugieren que las proteínas moonlighting son propensas a estar involucradas en enfermedades humanas, según indican los resultados del ODD test realizado: 16.47 (IC 95% 10.95-25.44) en el análisis OMIM siendo altamente significativo (valor exacto de la prueba de Fisher, $p < 2.2e-16$) (Figura 20).

Degree of significance

No. of Moons proteins involved in		111	78,169	
No. of Moons proteins not involved in diseases:		31	21,831	
Total		142		
No. of human proteins both in uniprot and OMIM:		3600	17,8501	
No. of human proteins in Uniprot		16568	82,1499	
Total		20168	Rewied proteins	
Risk estimators				**1,96**
		lower	upper	
Absolute risk in the treatment group	0,78	0,71	0,85	Significant if less than 1
Absolute risk in the control group	0,18	0,17	0,18	Significant if less than 1
Absolute risk reduction (RAR)	-0,60	-0,67	-0,54	Significant if it does not include 1
Relative risk (RR)	4,38	3,99	4,80	Number of times the two groups differ
Relative risk reduction (RRR)	-3,38	-3,80	-2,99	Times Uniprots are more frequent in the OMIM than the Moons
Odds				
Odds in the treatment group	3,58			
Odds in the control group	0,22			
Odds ratio (OR)	16,48	11,05	24,58	Significant if greater than 1
No. required to treat (NNT)	-1,66	-1,87	-1,49	Significant if less than 1

Figura 20: Resultados del análisis estadístico ODD test que muestra el grado de significación, con un intervalo de confianza del 95%. Esto demuestra que las proteínas moonlighting son más propensas a estar involucradas en enfermedades humanas que el grupo general de proteínas.

En la Tabla 9 se pueden ver algunos ejemplos relevantes de proteínas moonlighting que están involucradas en enfermedades humanas. Se identificaron después de cruzar los datos de OMIM y los de HGMD con la base de datos MultitaskProtDB-II. Estos ejemplos se han elegido para mostrar casos en los que el fenotipo puede atribuirse fácilmente a una de las funciones biológicas de la proteína. En negrita, se indican las supuestas funciones involucradas en las enfermedades: (C) para aquellas enfermedades relacionadas con la función canónica y (M) para aquellas relacionadas con la función moonlighting. Hay algunos ejemplos en los que cada función está relacionada con una enfermedad diferente (por ejemplo, Fumarate hydratase). En algunos casos es difícil dilucidar qué función es responsable de la enfermedad, lo que sugiere que ambas funciones pueden contribuir a los

diferentes síntomas (por ejemplo, la proteína Hes1). En la Información Suplementaria S5 podemos encontrar la lista de las 112 proteínas moonlighting que están involucradas en enfermedades humanas. En la Figura 21-A1 podemos ver una representación que muestra los porcentajes de las proteínas moonlightings clasificadas por tipo de enfermedad. Estos resultados podrían compararse con la Figura 21-A2 que muestra los mismos datos, pero relacionados con todo el conjunto de proteínas humanas involucradas en enfermedades según UniProt. Dos trabajos recientes pertenecientes al grupo de Brun predicen que el 3% del interactoma humano corresponde a proteínas moonlighting. También postulan que estas proteínas están significativamente involucradas en más de una enfermedad o comorbilidad (Chapple et al., 2015a, 2015b). Catorce de su conjunto de proteínas moonlighting humanas predichas se pueden encontrar en nuestra lista de proteínas moonlighting implicadas en enfermedades humanas. Debe tenerse en cuenta que nuestra lista solo está compuesta por proteínas moonlighting demostradas experimentalmente, mientras que las proteínas moonlighting de los estudios del grupo de Brun corresponden a las proteínas predichas (Zanzoni et al., 2015). Los resultados anteriores implican que las proteínas moonlighting humanas están significativamente asociadas a las enfermedades humanas en comparación con las no-moonlighting.

Tabla 9: Ejemplos de proteínas moonlighting implicadas en enfermedades humanas

* Columns linked to the corresponding information.
C – Disease related to the canonical function
M – Disease related to the moonlighting function

PROTEIN NAME (*)	CANONICAL FUNCTION	MOONLIGHTING FUNCTION	DISEASE	MOLECULAR PROCESS REFERENCE (*)	DRUG TARGETS (*)
Cyclooxygenase 1	**Prostaglandin G/H synthase**	Heme-dependent peroxidase	**(C) Bleeding disorder type 12**	Brit. J. Haemat. 92: 212-217, 1996.	**YES**
Gephyrin, protein anchor	Microtubule-associated protein	**Synthesis of molybdenum cofactor (MoCo)**	**(M) Molybdenum cofactor deficiency C**	Am J Hum Genet. 2001 Jan;68(1):208-13.	**YES**
Ribosomal S3 protein	Ribosomal protein	**DNA repair**	**(M) Colon adenocarcinomas**	DOI:http://dx.doi.org/10.1016/j.tig.2014.06.003	NO
Succinyl-coA synthetase	Succinyl-CoA synthetase	**Mitochondrial DNA maintenance**	**(M) Mitochondrial DNA depletion syndrome 9 (encephalomyopathic type)**	J. Med. Genet. 47: 670-676, 2010.	**YES**
Fumarate hydratase	**TCA cycle**	**Tumor suppressor**	**(C) Fumarase deficiency** **(M) Leiomyomatosis with Renal cell cancer**	Oncogene. 2015 Mar 19;34(12):1475-86.	NO
ERCC2 - TFIIH	**DNA helicase** **DNA repair damaged by exposure to ultraviolet light**	**It is also a subunit of TFIIH, a basal transcription factor.**	**(C) Xeroderma pigmentosum** **(C) Cerebrooculo facioskeletal syndrome 2**	- Mutat Res. 1992 Mar;273(2):193-202. - Am. J. Hum. Genet. 69: 291-300, 2001	NO
alpha-crystallin A chain	**Lens Crystallin**	**Heat-shock protein**	**(C) Cataract** **(M) Autoimmune diseases** **(C) Uveitis**	Biomed Pap Med Fac Univ Palacky Olomouc Czech Repub. 2005 Dec;149(2):243-9.	**YES**
Hes 1 protein	Transcriptional repressor	**It is able to induce the activation of the NF-kB pathway in T Cell Leukemia**	**(M) Leukemia, myeloid/lymphoid or mixed-lineage**	Cancer Cell. 2004 Sep;6(3):203-8.	NO
PIAS1	Inhibition of activated STAT	**Activation of p53**	**(M) Cancer**	Cold Spring Harb Perspect Biol. 2009 Nov; 1(5): a001883.	NO
Phosphoglucose isomerase	**Glycolysis**	Neuroleukin, differentiation and maturation factor / nerve growth factor / **stimulation of cell migration** / implantation factor / **modulator of tumor progression** and a target for cancer therapy / sperm surface antigen.	**(C) Hemolytic anemia PGI deficiency.** **(C) Angiogenesis in cancer.**	- Harefuah. 1994 Jun 15;126(12):699-702, 764, 763. - Cancer Res 2003;63:242-249.	**YES**

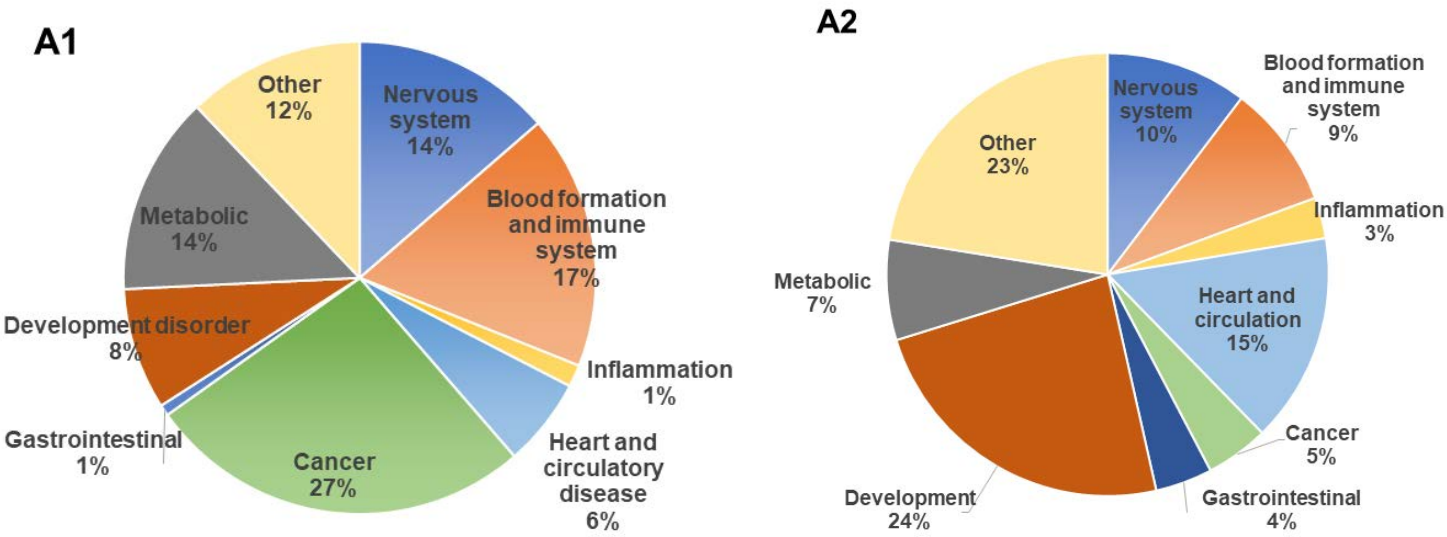


Figura 21: (A1) Distribución de los trastornos asociados con las proteínas moonlighting humanas y sus porcentajes relativos. (A2) Distribución de trastornos asociados a proteínas humanas, en general.

En la Figura 21, se puede observar como algunas clases de enfermedades son más frecuentes en proteínas moonlighting que en el proteoma humano en general. Por ejemplo, el 5% de las proteínas humanas involucradas en enfermedades, estan relacionadas con cáncer, mientras que en el caso de las proteínas moonlighting es de un 27%. Este hecho puede estar relacionado con que un gran número de proteínas moonlighting tienen funciones relacionadas con la regulación genética, especialmente como factores de transcripción. Algo similar pasa con las enfermedades de la sangre y el sistema inmunitario, un gran número de proteínas moonlighting tienen como función la regulación del sistema inmunitario y la angiogénesis. Por otro lado, en sentido contrario, muy pocas proteínas moonlighting estan implicadas en enfermedades del desarrollo embrionario. Es posible que, debido a que las proteínas moonlighting son mayoritariamente proteínas del metabolismo primario y otros procesos clave para la supervivencia, una alteración de estas proteínas pueda significar la no supervivencia del feto en desarrollo.

IV.C.2. PREDICCIÓN DE PROTEÍNAS CANDIDATAS A SER MOONLIGHTING A PARTIR DE LA INFORMACIÓN CONTENIDA EN LA BASE DE DATOS OMIM

En general, las proteínas moonlighting se descubren experimentalmente por serendipia. Por lo tanto, en la medida de lo posible, sería muy interesante identificarlas bioinformáticamente. Los equipos de Brun (Chapple et al., 2015a, Zanzoni et al., 2015), Kihara (Khan et al., 2012; 2014a, 2014b) y nosotros (Gómez et al., 2003; 2011; Hernández et al., 2014; 2015) hemos propuesto varios métodos de predecir bioinformáticamente las proteínas multitarea, como ya se ha descrito en el apartado IV.B. Una pregunta interesante que surge es si las bases de datos de enfermedades genéticas humanas, como OMIM, podrían ser un recurso útil para encontrar algunas proteínas moonlighting. Se ha iniciado un análisis manual de las bases de datos de enfermedades humanas para identificar posibles proteínas moonlighting y, además, para tratar de sugerir las bases moleculares de la enfermedad asociada. La Tabla 10 muestra algunos ejemplos de predicción de las posibles proteínas moonlighting presentes en la base de datos OMIM. Como ya se ha descrito en el Apartado IV.B., la información presente en las bases de datos de interacción proteína-proteína (PPI) y el uso de la herramienta de comparación de estructuras PiSite pueden ayudar a explicar la relación previamente desconocida entre la función canónica/moonlighting y la enfermedad asociada. Por ejemplo, en el caso de la anemia de Fanconi group J (Q9BX63), la búsqueda en la base de datos APID muestra que esta proteína interactúa con proteínas de reparación de DNA, pero también con la proteína relacionada con cáncer de mama BRC1. Otro ejemplo es la Calcium-independent phospholipase A2 (O60733), en la que la función canónica de la proteína está relacionada con el metabolismo de los ácidos grasos, pero parece estar también implicada en las enfermedades neurodegenerativas del cerebro. Además, un análisis en APID muestra una relación entre esta proteína y BAG, proteína implicada en la apoptosis y la supervivencia celular. Por otra parte, utilizando el servidor PiSite, encontramos que su estructura es similar a la proteína de apoptosis Caspasa-2. En la Tabla 10 se muestran más ejemplos sobre cómo las proteínas moonlighting pueden predecirse utilizando bases de datos de enfermedades, junto con análisis estructurales e interactómicos.

Tabla 10: Ejemplos de predicción de proteínas moonlighting humanas a partir de la información contenida en la base de datos OMIM y sus enfermedades genéticas asociadas

Protein & UniProt Descriptors (*)	Canonical (C) & predicted moonlighting functions (M)	Associated diseases & reference (*)	Interactomics partners (*)	PISITE models (*)
3-hydroxyacyl-CoA dehydrogenase type-2 Q99714 HCD2_HUMAN	C) mitochondrial ribonuclease P M) beta-oxidation of Fatty acids	1) 2-methyl-3-hydroxybutyryl-CoA dehydrogenase deficiency (MHBD deficiency) 2) Mental retardation X-linked, syndromes	1)Amyloid beta A4 protein 2) Mitochondrial ribonuclease P protein 1 3) Sulfatase-modifying factor 1 4) Mitochondrial ribonuclease P protein 3	1) 3-hydroxyacyl-CoA dehydrogenase type-2 PDBID: 1so8 CHAIN:A 2) 3-alpha-(or 20-beta)-hydroxysteroid dehydrogenase PDBID: 1nfq CHAIN:C
Pyruvate kinase PKLR P30613 KPYR_HUMAN	C) Glycolysis M) May participate in red cell survival.	1) Pyruvate kinase hyperactivity (PKHYP) 2) Pyruvate kinase deficiency of red cells (PKRD)	1) Myocilin 2) Kinesin-like protein KIF23 3) Rho guanine nucleotide exchange factor 7 4) Rho guanine nucleotide exchange factor 6 5) Paxillin 6) Serine/threonine-protein kinase PAK 1	*No relevant matches found.*
Fanconi anemia group J protein Q9BX63 FANCJ_HUMAN	C) DNA-dependent ATPase and 5' to 3' DNA helicase. M) Involved in the repair of DNA double-strand breaks	1) Breast cancer (BC) 2) Fanconi anemia complementation group J (FANCJ)	1) Breast cancer type 1 susceptibility protein 2) DNA mismatch repair protein Mlh1 3) Mismatch repair endonuclease PMS2 4) POZ-, AT hook-, and zinc finger-containing protein 1	*No relevant matches found.*
85/88 kDa calcium-independent phospholipase A2 O60733 PLPL9_HUMAN	C) Catalyzes the release of Fatty acids from phospholipids. M) May participate in apoptosis.	1) Neurodegeneration with brain iron accumulation 2B (NBIA2B) 2) Neurodegeneration with brain iron accumulation 2A (NBIA2A) 3) Parkinson disease 14 (PARK14)	1) BAG family molecular chaperone regulator 3	1) CASPASE-2 PDBID: 2p2c CHAIN:U
Alpha-aminoadipic semialdehyde synthase, mitocondrial Q9UDR5 AASS_HUMAN	C) Lysine-ketoglutarate reductase. M) Saccharopine dehydrogenase	1) Hyperlysinemia, 1 (HYPLYS1) 2) 2,4-dienoyl-CoA reductase deficiency (DECRD)	1) mRNA-decapping enzyme 1A 2) Peptidyl-tRNA hydrolase ICT1, mitochondrial 3) Myc proto-oncogene protein 4) Telomeric repeat-binding factor 2	1) SACCHAROPINE DEHYDROGENASE PDBID: 2axq CHAIN:A 2) SACCHAROPINE REDUCTASE PDBID: 1e5q CHAIN:H

Una pregunta intrigante es si los mutantes de los partners de interacción de una proteína moonlighting relacionada con una enfermedad también pueden contribuir a la misma patología, o otra muy similar. Esta idea refuerza mucho la participación de estos mutantes en la enfermedad. Con respecto a este tema, hay algunos ejemplos en la Tabla 10, como la 3-hidroxiacil-CoA deshidrogenasa. Esta proteína está involucrada en un trastorno de retraso mental y su partner de interacción, la proteína Amyloid beta A4 mutante, está involucrada en dos enfermedades relacionadas con el cerebro como el Alzheimer y la angiopatía cerebral. Además, como se menciona en el párrafo anterior, la interacción de la Fosfolipasa A2 con BRC1 es un buen ejemplo de dos mutantes asociados que

causan la misma enfermedad que la de la proteína moonlighting predicha. Al contrario, se pueden encontrar proteínas partners de interacción implicadas en diferentes enfermedades, lo que sugiere que la proteína moonlighting predicha puede participar en otra enfermedad aún no identificada. Un ejemplo de este caso es la Alpha-aminoadipic semialdehyde synthase (Q9UDR5), que presenta dos partners de interacción relacionadas con cáncer: el proto-oncogén Myc y el Factor 2 telomeric repeat-binding factor 2.

En resumen, se puede decir que la predicción de las proteínas moonlighting utilizando bases de datos OMIM y HGMD puede mejorar nuestro conocimiento a nivel molecular de la base clínica de una serie de enfermedades. Otras aplicaciones de todos estos estudios podrían ayudar a revisar y reinterpretar algunos fenotipos de enfermedades humanas, generando nuevas estrategias terapéuticas. Además, podrían explicarse algunos efectos secundarios de los fármacos (Butler et al., 2008).

IV.C.3. UN NÚMERO SIGNIFICATIVO DE PROTEÍNAS MOONLIGHTING SON DIANAS FARMACOLÓGICAS

En la clínica humana actual se requiere identificar la base molecular de una enfermedad y diseñar la terapia adecuada para ella. En la mayoría de los casos, el proceso terapéutico requiere el uso de medicamentos como tratamiento principal o complementario. Hemos investigado hasta qué punto las proteínas moonlighting humanas son dianas de fármacos actuales, que según Drews, representan “a small and biased set of the potential universe of the druggable genome” (Drews et al., 2000). Hemos determinado que el 48% de las proteínas moonlighting humanas de nuestra base de datos son dianas farmacológicas para fármacos actuales, mientras que solo el 9,8% de las proteínas humanas presentes en UniProt están descritas como dianas farmacológicas. Estos cálculos se realizaron como se explica en la Sección III.G., pero aquí tomamos en consideración las 1.969 proteínas humanas que son dianas farmacológicas presentes en las bases de datos TTD y DrugBank (Quin et al., 2014; Wishart et al., 2018). Este resultado es de nuevo claramente significativo como se puede ver en la Figura 22. Además, las probabilidades de que una proteína humana

presente en UniProt y una proteína moonlighting humana de MultitaskProtDB-II sean dianas farmacológicas enumeradas en las bases de datos TTD y DrugBank, se calcularon utilizando la relación ODD, observando una relación aumentada de la proporción de proteínas "druggables" 8.49 (IC 95% 5.99-12.00) en el subconjunto de moonlightings, siendo esta diferencia altamente significativa (Figura 22).

Degree of significance

No. of Druggable Human Proteins in MoonDB:	68	47,887324
No. of not Druggable Human Proteins in MoonDB:	74	52,112676
Total	142	
No. of Druggable Human Proteins in Uniprot:	1969	9,7629909
No. of not Druggable Human Proteins in Uniprot:	18199	90,237009
Total	20168	Rewied proteins

Risk estimators				**1,96**
		lower	upper	
Absolute risk in the treatment gro	0,48	0,40	0,56	Significant if less than 1
Absolute risk in the control group	0,10	0,09	0,10	Significant if less than 1
Absolute risk reduction (RAR)	-0,38	-0,46	-0,30	Significant if it does not include 1
Relative risk (RR)	4,90	4,11	5,85	Number of times the two groups differ
Relative risk reduction (RRR)	-3,90	-4,85	-3,11	Times Uniprots are more frequent in drugbank than the Moons

Odds		**Confidence interval**		
Odds in the treatment group	0,92			
Odds in the control group	0,11			
Odds ratio (OR)	8,49	6,09	11,84	Significant if greater than 1
No. required to treat (NNT)	-2,62	-3,34	-2,16	Significant if less than 1

Figura 22: Resultados del análisis estadístico ODD test que muestra el grado de significación con un intervalo de confianza del 95% y que demuestra que las proteínas moonlighting son más propensas a ser dianas para fármacos actuales que el grupo general de proteínas.

Los resultados descritos en este y en los dos apartados anteriores, IV.C.1 y IV.C.2, resaltan el interés de las proteínas moonlighting para obtener información de las bases moleculares de las enfermedades genéticas y para el diseño racional de fármacos con la identificación precisa de la diana farmacológica.

En la última columna de la Tabla 9, se pueden ver algunos ejemplos de los fármacos actuales relacionados con la proteína moonlighting que tienen por diana. En la Información Suplementaria S6, se enumera el conjunto completo de

68 proteínas moonlighting identificadas actualmente como dianas farmacológicas, con las correspondientes referencias a las enfermedades humanas y bases de datos de fármacos. La Figura 23-B1 muestra una representación con los porcentajes de las proteínas moonlighting que son dianas de fármacos clasificadas por clases funcionales. Estos resultados pueden compararse con la Figura 23-B2 que muestra los mismos datos, pero relacionados con el conjunto completo de proteínas humanas que son dianas de fármacos actuales, según la base de datos DrugBank (Wishart et al., 2008). Teniendo en cuenta que muchas enfermedades implican a las proteínas moonlighting, debería pasar lo mismo con los fármacos. Aun así, esto no es tan simple como parece. "Druggable" no significa ser una diana farmacológica porque las dianas de los medicamentos actuales son tan solo aquellas para las cuales hay un fármaco probado con éxito (Santos et al., 2017). Además, es bien sabido que existen muchas dianas no druggables. Se debe considerar que Drews (Drews et al., 2000) estimó que la cantidad de posibles dianas farmacológicas oscilaba entre 5,000 y 10,000 (Drews et al., 2000).

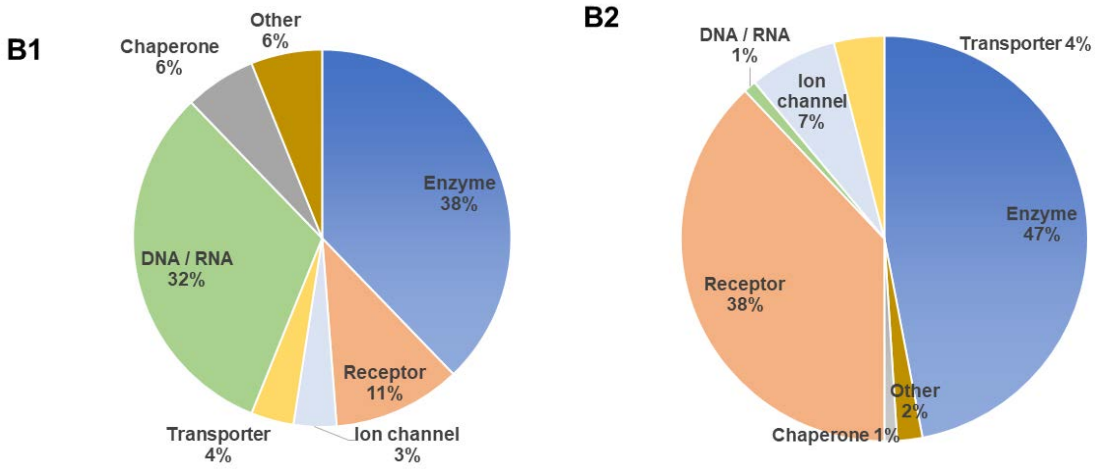


Figura 23: (B1) Clasificación funcional de dianas farmacológicas que son proteínas moonlighting y sus porcentajes relativos. (B2) Clasificación funcional de dianas farmacológicas de proteínas en general y sus porcentajes relativos.

Este 48% de las proteínas moonlighting humanas que son dianas farmacológicas, junto con el hecho de que el 78% de los genes OMIM corresponden a proteínas moonlighting, respalda la opinión compartida por muchos autores de que la multifuncionalidad de las proteínas no es un fenómeno raro y, por lo tanto, muchas proteínas humanas serían multitasking. En un párrafo anterior, se ha sugerido la inspección de las bases de datos de genes de enfermedades con el fin de descubrir nuevas proteínas moonlighting, y también sugerimos la búsqueda de proteínas moonlighting para la selección de dianas de fármacos.

IV.D. RELACIÓN DE LAS PROTEÍNAS MOONLIGHTING CON LA INFECCIÓN Y VIRULENCIA DE MICROORGANISMOS PATÓGENOS

IV.D.1. PROTEÍNAS MOONLIGHTING IMPLICADAS EN VIRULENCIA

Una revisión de nuestra base de datos (MultitaskProtDB-II) muestra que el 25% (232 proteínas) de toda la base de datos son proteínas de microorganismos en las que la segunda función se relaciona con la patogénesis y la virulencia del mismo. Este número aumenta cada día. En la siguiente Tabla 11 pueden verse las proteínas más comunes implicadas en la patogénesis. La lista completa de 232 proteínas implicadas en virulencia está disponible en la Información Suplementaria S7 de este trabajo.

Tabla 11: Proteínas de microorganismos implicadas en patogénesis y el número de especies que las utilizan

Protein name	Number of entries
Enolase	42
Glyceraldehyde-3-phosphate dehydrogenase	27
60 kDa chaperonin	20
Chaperone protein DnaK	13
Aldolase	8
Elongation factor Tu	8
Phosphoglycerate kinase	7
Glutamine Synthetase	5
Phosphoglycerate mutase	5
Triosephosphate isomerase	5
Diacylglycerol acyltransferase	4
Peptidyl Prolyl cis,trans-Isomerase	4
Phosphoglycerate mutase	3
Phosphoglucose isomerase	2

Como puede verse en la Tabla 12, la proteína que se ha relacionado más veces con la virulencia de organismos patógenos es la Enolasa. Como se discutió en la Introducción, todas estas proteínas son del metabolismo primario, altamente conservadas y ancestrales y normalmente se encuentran en el núcleo o el citoplasma de los microorganismos. Todavía se desconoce cómo se pueden secretar estas proteínas a la membrana, teniendo en cuenta que no tienen los aminoácidos apropiados que constituyen la señal de exportación. Se ha sugerido que serían exportados por el sistema independiente de péptido señal SecA2 (Feltcher et al., 2013). Según una revisión reciente sobre los mecanismos para la exportación de proteínas bacterianas, las bacterias Gram negativas usarían el sistema T5SS (Costa et al., 2015). En la superficie bacteriana, su acción de virulencia es principalmente mediante interacción con Plasminógeno (PLG) o con componentes de matriz extracelular (ECM) o también contribuyendo a la evasión del sistema inmune del huésped (inmunidad mediada por complemento, etc.) (Greiciely et al., 2017).

En nuestra base de datos actualizada, las Enolasas de 48 especies diferentes (Tabla 12) se unen al Plasminógeno u otras proteínas del huésped para facilitar la adhesión al huésped y la invasión tisular, el mecanismo general de virulencia de estas proteínas es unir proteínas diana del huésped como Plasminógeno, Laminina, Fibronectina o Mucina. Entre ellos, el más común es el Plasminógeno. Sin embargo, se ha demostrado que la Enolasa y otras proteínas de organismos no patógenos pueden unirse al Plasminógeno y a otras proteínas del huésped, incluso más eficazmente que las patógenas. Este mecanismo puede ser utilizado por la microbiota normal para competir con el organismo patógeno por los mismos tejidos, protegiendo al huésped de la infección (Castaldo et al., 2009). En la Sección IV.D.3. analizamos los motifs responsables de esa unión y encontramos que los organismos patógenos y no patógenos los tienen.

Tabla 12: Enolasas de microorganismos implicadas en la adhesión del microorganismo al huésped con sus funciones moonlighting explicadas

MOONLIGHTING FUNCTION	ORGANISM
It may be involved in a host of other biological functions.	*Plasmodium falciparum*
Plasminogen binding/plasminogen receptor, heat shock protein, cytoskeletal/chromatin structure binding. Binding to C4b-binding proteins. Inhibits C3 convertase formation and depletes complement.	*Streptococcus pneumoniae*
Plasminogen binding	*Aeromonas hydrophila*
Plasminogen binding	*Bacillus anthracis*
Plasminogen binding	*Bifidobacterium animalis subsp. lactis*
Binding to plasminogen.	*Borrelia burgdorferi*
Inhibitor of Neisseria binding. Plasminogen and Laminin binding.	*Lactobacillus johnsonii*
Fibronectin binding. Binding to Plasminogen. Binding to intestinal epithelial cells	*Lactobacillus plantarum*
Plasminogen binding	*Mycoplasma fermentans*
Plasminogen binding	*Neisseria meningitidis*
A role in virulence: it could be involved in larval degradation, maybe through Plasminogen system	*Paenibacillus larvae subsp. larvae*
Laminin binding protein. Binding to Plasmin(ogen).	*Staphylococcus aureus*
Binding to salivary mucin MUC7	*Streptococcus gordonii*
Fibronectin binding protein/adhesin	*Streptococcus suis*
Plasminogen binding	*Trichomonas vaginalis*
EgEno1 expression patterns in hydatid cyst components showed that these proteins are exposed in the host–parasite interface raising evidence of their involvement in molecular signaling pathways important for parasite survival and development.	*Echinococcus granulosus*
Molecular chaperone	*Saccharomyces cerevisiae*
It has neurotrophic and neuroprotective effects on rather a broad spectrum of neurons in the central nervous system (promotes cell survival)	*Homo sapiens*
Plasminogen binding. Binding to Albumin.	*Streptococcus pyogenes*
Fibronectin binding	*Paracoccidioides brasiliensis*
Plasminogen binding	*Candida albicans*
Plasminogen and Laminin binding	*Lactobacillus crispatus*
Plasminogen binding	*Onchocerca volvulus*
Plasminogen binding	*Oral streptococci*
Plasminogen binding. Binding to salivary mucin.	*Streptococcus mutans*
Plasminogen binding	*Leishmania mexicana*
Tau lens crystallin.	*Petromyzon marinus*
Adhesion to host.	*Lactobacillus gasseri*
Required for vacuole homotypic membrane fusion protein	*Saccharomyces cerevisiae*
Binding to Plasmin(ogen)	*Bifidobacterium bifidum*
Binding to Plasmin(ogen)	*Bifidobacterium breve*
Binding to Plasmin(ogen)	*Bifidobacterium longum*
Binding to Plasmin(ogen)	*Mycoplasma pneumoniae*
Binding to Plasmin(ogen)	*Streptococcus anginosus*
Binding to Plasmin(ogen)	*Listeria monocytogenes*
Binding to Plasmin(ogen)	*Streptococcus anginosus*
Binding to Plasmin(ogen)	*Streptococcus oralis*
Binding to intestinal epithelial cells	*Streptococcus suis*

A peptide that binds to the surface of the mosquito midgut epithelium and inhibits the invasion of *Plasmodium berghei.*	*Plasmodium berghei*
Secreted Enolase has been suggested to suppress the immune response in insects	*Steinernema glaseri*
Associated with epithelial cell binding	*Enterococcus faecalis*
Specifically binds to a TTTTCT DNA motif present in the cyst matrix antigen 1 (TgMAG1) gene promoter. Provides a potential lead in the design of drugs against *Toxoplasma* brain cysts.	*Toxoplasma gondii*
Binding to Plasminogen.	*Taenia pisiformis*
Adhesion to host. It is an Effective Protective Antigen in Mice.	*Streptococcus iniae*
Involved in the encystation mechanism	*Entamoeba invadens*
Plasminogen binding.	*Streptococcus dysgalactiae subsp equisimilis*
Is an immunosuppressive protein	*Streptococcus sobrinus*
Binds to Plasminogen.	*Mycobacterium tuberculosis*

Otro hecho importante sobre la unión al Plasminógeno es que los microorganismos patógenos no solo se unen al Plasminógeno, sino que activan su paso a Plasmina, causando una lisis del tejido que puede facilitar la invasión de los tejidos del huésped. Todavía se desconoce si esta es la diferencia entre un microorganismo patógeno y uno no patógeno. Todos estos hechos plantean tres preguntas principales: (a) ¿Por qué la función canónica de estas proteínas de virulencia proviene principalmente de funciones biológicas ancestrales clave (especialmente del metabolismo primario)? (b) ¿Por qué las mismas proteínas son compartidas por muchas especies de patógenos?, y (c) ¿Cómo ese conjunto de proteínas secuencial y estructuralmente diferentes interaccionan con un pequeño grupo de dianas de los componentes de la matriz extracelular?

IV.D.2. ¿POR QUÉ NUMEROSOS FACTORES DE VIRULENCIA DE LOS MICROORGANISMOS PATÓGENOS SON PROTEÍNAS MOONLIGHTING?

En un trabajo anterior, nuestro grupo (Amela et al., 2017) sugirió que, dado que una de las tareas más importantes del sistema inmune es la diferenciación entre antígenos propios y extraños, este sistema descartaría la producción de anticuerpos protectores contra las proteínas del microorganismo patógeno que compartan epítopos con proteínas del huésped (mimetismo epitópico). Esto podría ser causa de enfermedades autoinmunes (Benoist et al., 2001). Significa que, aunque muchas proteínas patógenas pueden ser antigénicas, solo unas pocas de ellas generarían una respuesta inmune protectora.

Como puede verse en la Tabla 11, la mayoría de las proteínas moonlighting pertenecen al metabolismo primario (glucólisis, ciclo de Krebs ...) estando sus secuencias aminoacídicas altamente conservadas a lo largo de la evolución (Henderson et al., 2011; Amblee et al., 2015). Por lo tanto, probablemente compartan epítopos.

Como puede verse en la Figura 24, muchos epítopos humanos y de la proteína homóloga del patógeno se superponen entre si y con los segmentos de secuencia de aminoácidos altamente conservados presentes en todas las Enolasas. De acuerdo con nuestra hipótesis, el sistema inmune humano no generaría anticuerpos protectores contra las Enolasas del patógeno. Como hemos descrito anteriormente, hay ejemplos en los que incluso un número menor de epítopos compartidos puede ser responsable de una respuesta autoinmune (Amela et al., 2007).

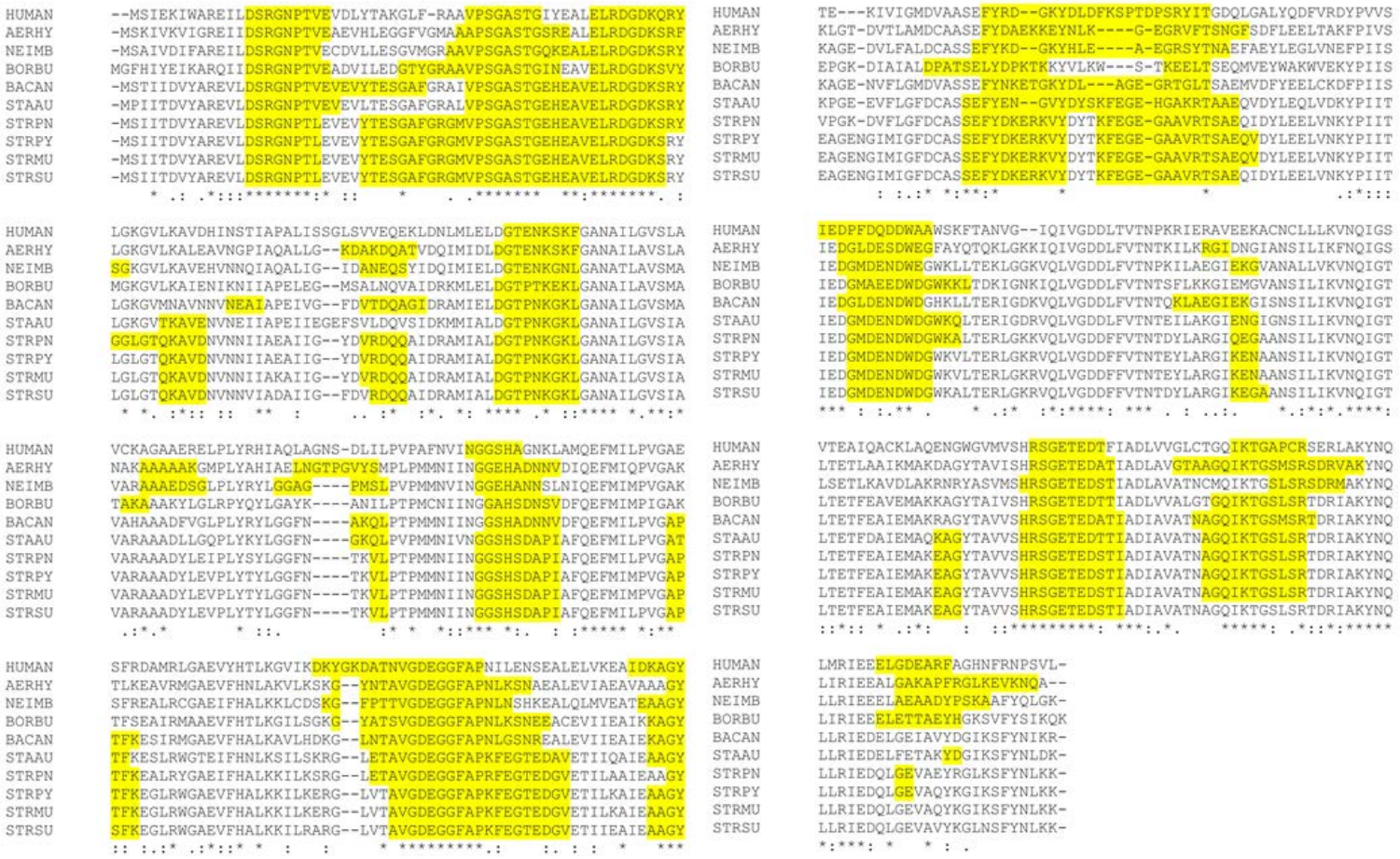


Figure 24: Alineamiento múltiple con el programa Clustal Omega de Enolasas humanas y de diversos patógenos. En amarillo se muestran los epítopos predichos para las diferentes Enolasas usando el servidor de BepiPred. Los asteriscos representan aminoácidos completamente conservados. La mayoría de los epítopos predichos humanos coinciden con los aminoácidos altamente conservados. Los microorganismos son: AERHY = *Aeromonas hydrophila*; NEIMB = *Neisseria meningitidis*; BORBU = *Borrelia burgdorferi*; BACAN = *Bacillus anthracis*; STAAU = *Staphylococcus aureus*; STRPN = *Streptococcus pneumoniae*; STRPY = *Streptococcus pyogenes*; STRMU = *Streptococcus mutans*; STRSU = *Streptococcus suis*.

En la Figura 25 y 26 se muestran otros ejemplos de enzimas moonlighting del patógeno (GAPDH, PDK, PMD ...) alineadas con la proteína ortóloga humana. Como puede verse, también coinciden con los segmentos de la secuencia de aminoácidos y los epítopos predichos se superponen. En la Información Suplementaria S8 se muestran más ejemplos.

De hecho, se ha descrito que la Enolasa estreptocócica reacciona de forma cruzada con la Enolasa humana y puede estar involucrada en afecciones autoinmunes y complicaciones posteriores a la infección (Cole et al., 2005; Fontan et al., 2000).

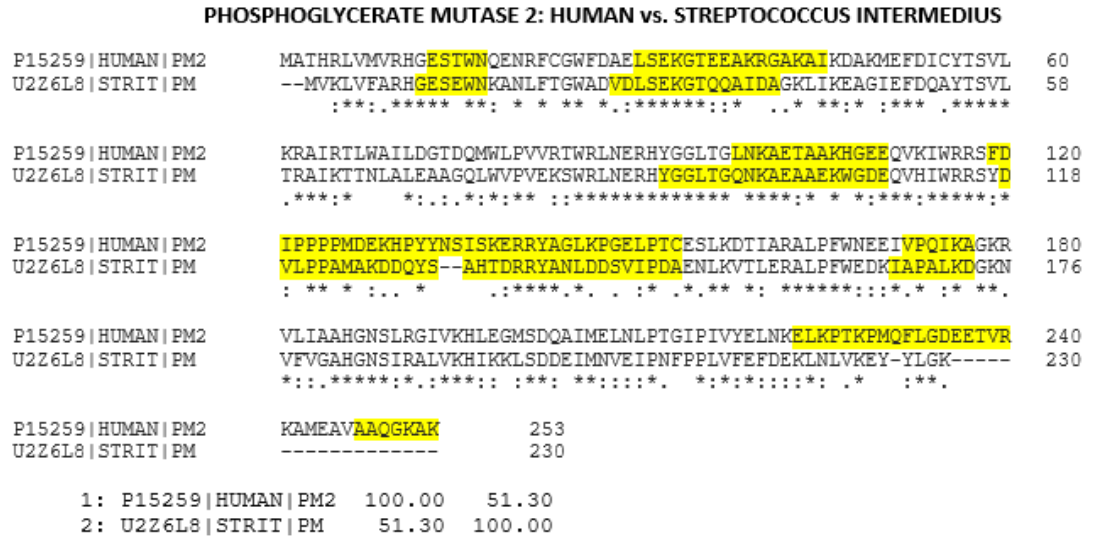


Figura 25: Alineamiento BLASTP entre la proteína moonlighting Phosphoglycerate mutase 2 de *Streptococcus intermedius* implicada en la patogénesis y su homóloga en humanos.

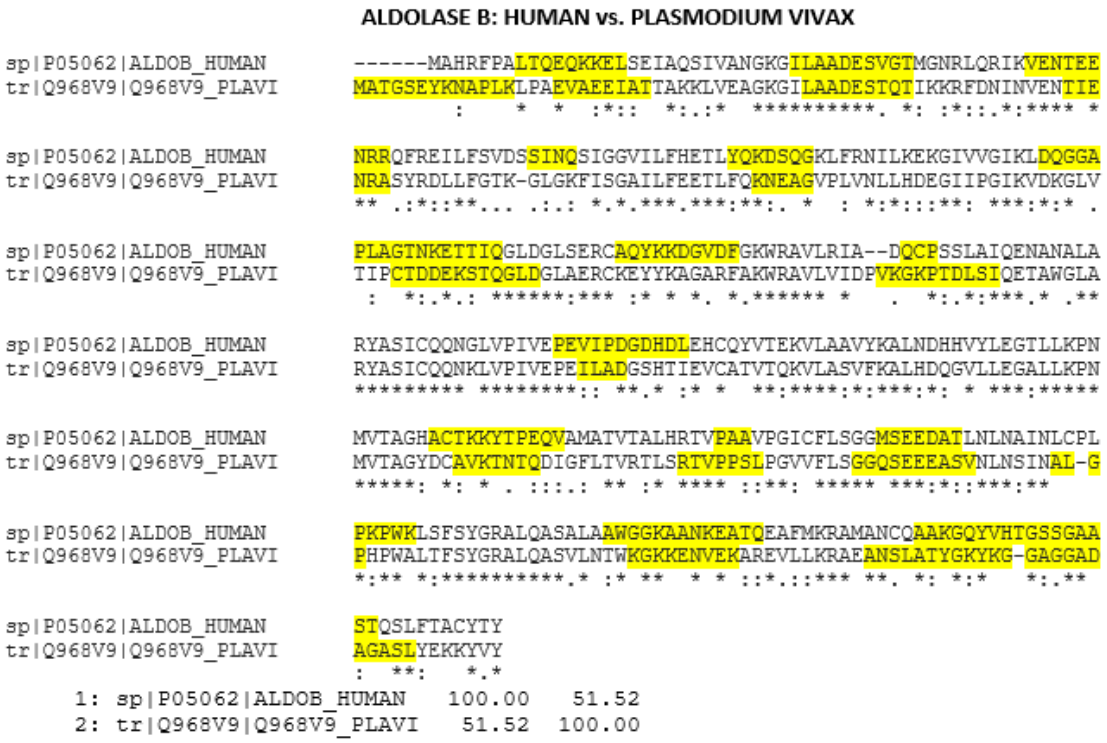


Figura 26: Alineamiento BLASTP entre la proteína moonlighting Aldolase B de *Plasmodium falciparum* implicada en la patogénesis y su homóloga en humanos.

Obviamente, los resultados anteriores tienen una importancia particular para el diseño y desarrollo de vacunas recombinantes por subunidades, para las que es crítica la identificación y selección de las dianas proteicas del patógeno adecuadas antes de diseñar la vacuna. En este sentido, una inspección exhaustiva de la base de datos Violinet (He et al., 2014) muestra que, en el mercado no existe ninguna proteína moonlighting elegida como vacuna. Encontrar vacunas en el mercado que generen actividad protectora, sería la mejor prueba de que un antígeno es eficaz. Pocas (algunas chaperonas como GroEL y Hsp70) son vacunas en algún estado de investigación según Violinet, pero ninguna está ya aprobada y comercializada y, por lo tanto, probadamente eficaz. Además, en todos los casos que implican proteínas moonlighting, los ensayos se han realizado en ratones, cobayas o conejos mostrando un nivel de protección bastante bajo (≤ 20%). En algún caso los autores simplemente indican la presencia de "algún grado" de respuesta inmune. Según Violinet, las proteínas recombinantes que han llegado al mercado (o están a punto de hacerlo, como una vacuna contra el Ébola) son:

Tabla 13: Vacunas de proteínas recombinantes actuales, enlazadas con la correspondiente referencia en la base de datos Violinet

PATHOGEN	HOST	RECOMBINANT PROTEINS
Neisseria meningitidis	Human	NHBA, NadA, FHbp
Borrelia burgdorferi	Human	OspA
Bordetella pertussis	Human	fhaB
Human papilloma virus	Human	L1
Hepatitis B virus	Human	Capsid protein
Zaire ebola virus	Human	Vp35
Pig circovirus	Pig	Capsid protein
Pasteurella multocida serotype D	Pig	Dermonecrotoxin

Alineamientos por pares de estas proteínas de los patógenos frente al proteoma humano, o de los correspondientes mamíferos, utilizando BLASTP en el servidor del NCBI, y utilizando los parámetros por defecto, indican "*No significant similarity found*".

Estos resultados concuerdan con nuestra hipótesis de que el huésped, para evitar una respuesta autoinmune, evita la producción de anticuerpos protectores contra proteínas patógenas con las que comparte epítopos. Por lo tanto, la evolución del patógeno seleccionaría positivamente aquellas proteínas de virulencia cuya secuencia de aminoácidos se conserva en cierto grado. Nuestra hipótesis explicaría la falta de vacunas de subunidades exitosas presentes en el mercado basadas en proteínas moonlighting. Por otro lado, debido al grado de secuencia conservada y epítopos compartidos, debería existir inmunidad protectora entre cepas, e incluso entre especies, utilizando proteínas moonlighting como vacunas de subunidad, lo que no es el caso. Por todas estas razones, una estrategia basada en el diseño de una vacuna que use una proteína moonlighting como el antígeno principal podría no tener éxito.

IV.D.3. IDENTIFICACIÓN DE MOTIFS COMUNES EN PROTEÍNAS MOONLIGHTING DE VIRULENCIA

Como se describió anteriormente, varias proteínas secuencial y estructuralmente diferentes interaccionan con un pequeño grupo de dianas, principalmente componentes de la matriz extracelular. Si no comparten la estructura 3D, ¿Comparten motivos o dominios de aminoácidos? Por ejemplo, muchas proteínas pueden unirse al Plasminógeno, pero si tienen un motif común para hacerlo, aún se desconoce. En el caso de la Enolasa se han descrito tres pequeños motifs (Kornblatt et al., 2011). Estos han sido relacionados con la actividad de unión del Plasminógeno. Diseñamos un programa para encontrar en qué proteínas y organismos se pueden encontrar estos motifs. Un resultado sorprendente es que encontramos que estos motifs se pueden encontrar en organismos patógenos y no patógenos, incluso en aquellos en los que no se ha descrito la función de unión al Plasminógeno. Entre todos los motifs de aminoácidos descritos en la bibliografía (Itzek et al., 2010) los más importantes son: KK (en el N-terminal de la proteína), KxxK y FYDKERKVY. Usando el programa mencionado anteriormente, encontramos que los motifs KK y KxxK se encuentran aleatoriamente en una gran lista de microorganismos, patógenos y no patógenos, lo que no es sorprendente ya que la Lys es un aminoácido muy abundante y la longitud del motif es muy corta. Además, el motif FYDKERKVY

solo se encuentra en el género *Streptococcus*. Por lo tanto, debe ser otra característica de la proteína la que va a determinar qué organismos van a ser patógenos y van a activar el Plasminógeno. Por otra parte, una búsqueda de estos motifs en otras proteínas relacionadas con virulencia (GAPDH, PGI, etc) no encuentra resultados, aparte de la presencia de Lys.

Una búsqueda mediante InterPro de las proteínas moonlighting relacionadas con virulencia, no identifica motifs/dominios significativos. Como InterPro identifica motifs suficientemente largos, hemos procedido a rastrear esas secuencias con el programa MinimotifMiner (Mi et al., 2012) que busca motifs cortos, generalmente ligados a sitios de interacción, de modificación post-traduccional, etc. Usando el programa MinimotifMiner, buscamos pequeños motifs que se encuentran en la Enolasa de *Streptococcus pneumoniae* y no en la de *Lactobacillus plantarum*, un microorganismo no patógeno cercano genéticamente. Utilizando un proceso de selección explicado en la Sección III.H. obtuvimos 47 motifs (enumerados en la Información Suplementaria S9). A continuación, buscamos estos motifs en diferentes microorganismos patógenos y encontramos que dos de ellos están presentes mayoritariamente en Enolasas de patógenos (Tabla 14).

Tabla 14: Motifs que se encuentran presentes principalmente en la Enolasa de microorganismos patógenos

Motif	Function	Observations
YTAV	Binds SH2 domain of Protein Tyrosine Phosphatase Non-Receptor Type 11.	Most of the organism that share this motif are pathogenic including microorganism phylogenetically far like: *Escherichia, Bacillus, Burkholderia, Enterococcus, Salmonella, Staphylococcus, Streptococcus, Treponema, Vibrio, Clostridium, Listeria, Lactobacillus, Pseudomonas, Streptococcus, Yersinia.*
Y[LIV]E[LIV] Y[VIL][ED][PIV] Y[VIL][ED][VIL]	This consensus motif binds the SH2 domain of PLCgamma;	Most of the organism that share this motif are pathogenic. *Streptococcus, Staphylococcus, clostridium, campylobacter*.

Como se muestra en la Tabla 14, estos dos motifs también son importantes debido a la función que suelen desempeñar. El primero, YTAV, se une a la Protein Tyrosine Phosphatase Non-Receptor Type 11 (PTPN11). PTN11 participa en procesos biológicos como la coagulación sanguínea, la regulación de la respuesta inmune y la señalización. El segundo motif mostrado en la Tabla 14 se une a la proteína PLCgamma, que se ha demostrado que es un sustrato principal para la Heparin-binding growth factor 1 (acidic fibroblast growth factor). Ambas proteínas están estrechamente relacionadas con el Plasminógeno, funciones en la sangre y la regulación inmune. Los microorganismos en los que sus Enolasas presentan estos motifs se encuentran listados en la Información Suplementaria S10. Estos motifs están representados en la estructura tridimensional de la Enolasa en la Figura 27.

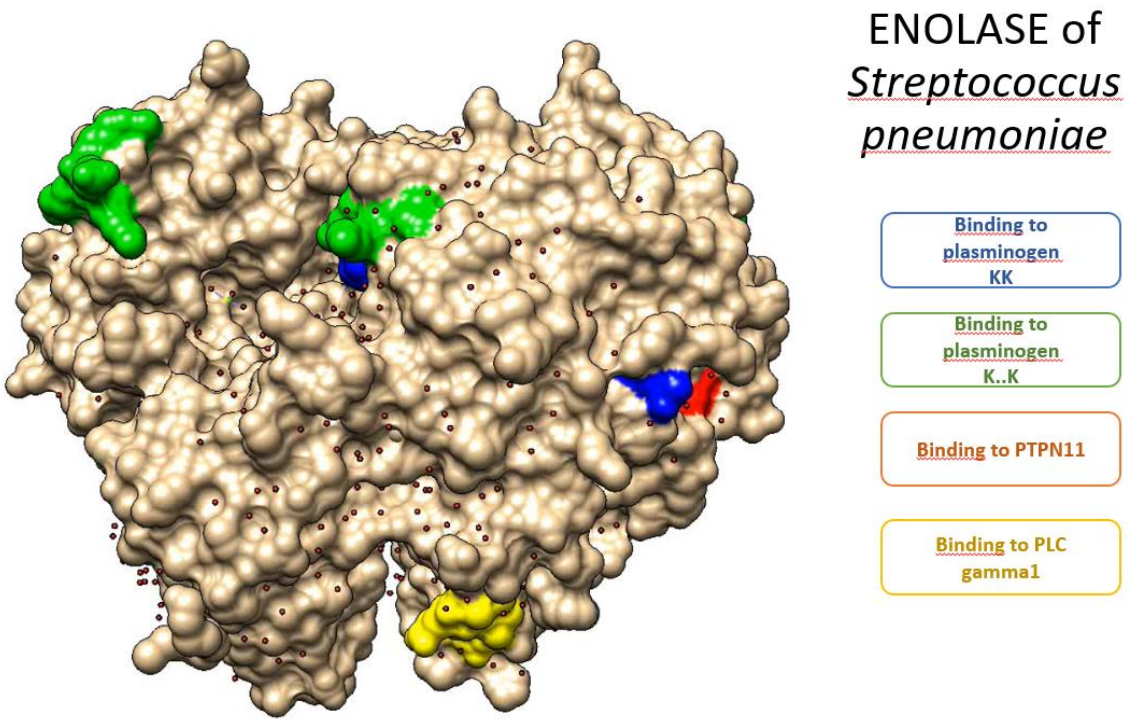


Figura 27: Estructura 3D de la Enolasa con algunos motifs marcados en diferentes colores en la superficie. Los motifs marcados en azul y verde son descritos en la bibliografía por participar en la unión al Plasminógeno. Los resaltados en rojo y amarillo son aquellos descubiertos usando el análisis de minimotifs descrito anteriormente.

Además, analizamos las interacciones proteína-proteína de ambas proteínas usando la base de datos BioGrid y obtuvimos que ambas proteínas interaccionan con una gran lista de proteínas involucradas en la coagulación sanguínea, regulación de la respuesta inmune, reacción Plasminógeno-Plasmina, etc (Figura 28).

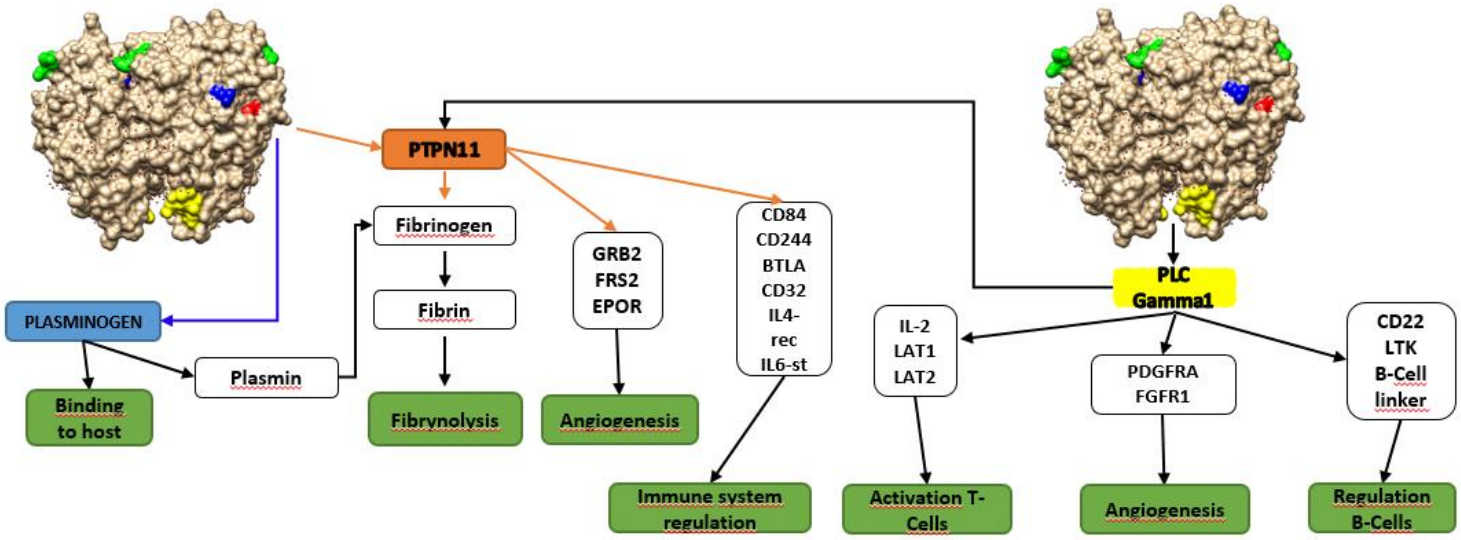


Figura 28: Partners de interacción proteína-proteína y funciones relacionadas del motif YTAV (marcado en rojo en la estructura de la Enolasa) e Y[LIV]E[LIV] (marcado en amarillo en la estructura de la Enolasa).

PTPN11, interacciona con proteínas implicadas en la fibrinólisis, la angiogénesis, la regulación del sistema inmune y la interacción con el Plasminógeno. Por otro lado, PLC Gamma1 interacciona con proteínas involucradas en la regulación del sistema inmune, la activación de las células T, la angiogénesis y la regulación de las células B. Este motif, presentado principalmente en la Enolasa de microorganismos patógenos, permite no solo unirse al Plasminógeno sinó también activarlo, además de regular la angiogénesis y el sistema inmunitario, facilitando así la infección. Ambos motifs están localizados en la superficie de la Enolasa, lo que les permite ser funcionales, como se puede ver en la Figura 27 y 28.

Únicamente el hecho de analizar la secuencia de las proteínas moonlighting de virulencia puede no ser suficiente para identificar las regiones implicadas en la interacción con PLG y otras proteínas diana del huésped. En todo caso, cabe mencionar que ambas regiones de unión al PLG de la Enolasa y de la M protein de *S. pyogenes* comparten la conformación en forma de una hélice alfa (Figura 29).

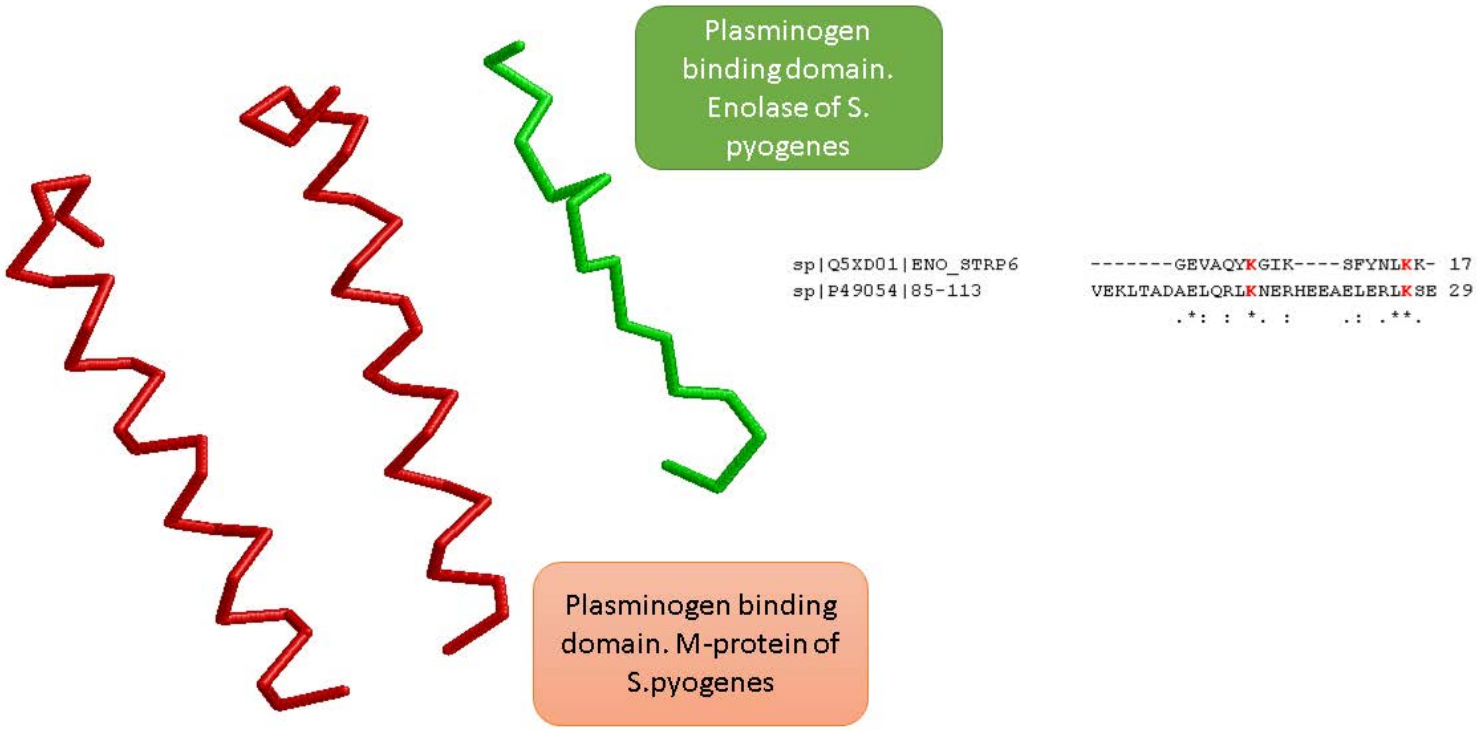


Figura 29: Las regiones de unión a Plasminógeno de la Enolasa de S. pyogenes (verde) y la M-protein (roja) muestran una estructura tridimensional conservada. A la derecha, el alineamiento secuencial entre ambas proteínas no muestra conservación, solo en dos Lys, descritas con frecuencia en la bibliografía como implicadas en la unión del Plasminógeno.

Esto refuerza la idea de que únicamente con un análisis de las secuencias no se puede tener una idea clara de si las funciones de varias proteínas están o no conservadas, ya que como se puede observar en la Figura 29, dos regiones que conservan función y estructura tridimensional no conservan la secuencia.

IV.E. PROTEINAS MOONLIGHTING Y EVOLUCIÓN

En este apartado IV.E. mostramos diversas aproximaciones que hemos utilizado para responder algunas de las preguntas que plantea la existencia de proteínas moonlighting, como por ejemplo: si abundan, si se conserva la multifuncionalidad, etc.

IV.E.1. ¿ES LA MULTIFUNCIONALIDAD DE LAS PROTEÍNAS MUCHO MÁS GENERAL DE LO QUE SE CRÍA HASTA AHORA?

Esta es una pregunta que se hacen la mayoría de los investigadores en el campo. Muchos de ellos, incluidos nosotros, pensamos que las proteínas moonlighting conocidas actualmente son solo "la punta del iceberg" y que un gran número de proteínas están realizando más de una función molecular. A partir de las siguientes aproximaciones sugerimos que las proteínas multifuncionales abundarían. Esta pregunta adquiere más relevancia a partir de los recientes resultados de proteómica en que el número de proteínas del proteoma humano, ratón, etc, es mucho menor de lo que se esperaba a partir de los resultados de transcriptómica/splicing (Tress et al., 2017). En la Discusión General se ampliará este aspecto.

IV.E.1.a. APROXIMACIÓN BASADA EN LA EXISTENCIA DE MULTIPLICIDAD DE DESCRIPTORES GO PARA *MOLECULAR FUNCTION* UTILIZANDO LA INFORMACIÓN CONTENIDA EN LA BASE DE DATOS UNIPROT

Como se ha descrito en la Sección de Métodos, la base de datos GO utiliza tres descriptores para clasificar las proteínas según su función: (a) *Biological process (b) Molecular function* y (c) *Cellular component*. Una proteína que desempeña más de una función molecular se puede considerar como moonlighting. Sin embargo, esto no es directamente aplicable a las funciones biológicas, porque una proteína realizando la misma función molecular puede estar implicada en diferentes procesos biológicos. Por ejemplo, una quinasa que realiza una única función molecular (es decir, la transferencia de grupos fosfato) puede participar en diferentes rutas dependiendo de la proteína que esté fosforilando. Este caso

no se consideraría moonlighting, ya que únicamente está desarrollando una función molecular que interviene en distintos procesos.

A partir de esta premisa, se ha realizado un análisis de data mining de la base de datos UniProt buscando el número de proteínas que tienen identificadores GO con diferentes funciones moleculares, comenzando con GO asociadas a las funciones principales en nuestra base de datos de proteínas moonlighting (Ver Tablas 4 y 5). Como se puede ver en las siguientes tablas 15-17, los resultados muestran que habría un gran número de posibles proteínas multifuncionales, a las que los investigadores han otorgado descriptor GO multifuncional, pero que en la bibliografía no se han incorporado los términos *moonlighting*, *multitasking*, etc, términos que utilizamos para capturar proteínas moonlighting a partir de la bibliografia para ser añadidas a la base de datos MultitaskProtDB-II.

Tabla 15: Número de proteínas que tienen dos funciones diferentes (de acuerdo con los códigos GO anotados en la base de datos UniProt)

GO Function 1	GO Function 2	Number of proteins
isomerase activity [0016853]	DNA binding [0003677]	166.833
DNA binding transcription factor activity [0003700]	catalytic activity [3824]	128.026
catalytic activity [0003824]	protein folding [0006457]	108.037
isomerase activity [0016853]	oxidoreductase activity [0016491]	79.566
protein folding [0006457]	isomerase activity [0016853]	78.219
chaperone binding [0051087]	cytokine activity [0005125]	56.072
ligase activity [0016874]	DNA binding [0003677]	33.878
oxidoreductase activity [0016491]	DNA binding [0003677]	20.690
catalytic activity [0003824]	pathogenesis [0009405]	11.510
isomerase activity [0016853]	ligase activity [0016874]	2.992
protein folding [0006457]	DNA binding [0003677]	2.843
protein folding [0006457]	oxidoreductase activity [0016491]	1.149

Como se muestra en la Tabla 15, buscando en UniProt proteínas con actividad catalítica (enzimas) y actividad de factor de transcripción, que es el par funcional más común en las proteínas moonlighting, obtuvimos 128.026 resultados. Y un resultado similar se encuentra cuando buscamos proteínas con actividad catalítica e involucradas en el plegamiento de proteínas (chaperonas), en este caso se encuentran 108.037 proteínas.

Continuando con esta idea, buscamos proteínas localizadas en diferentes compartimentos celulares, dado que en muchas ocasiones las dos funciones, canónica y moonlighting, tienen lugar en dos compartimientos diferentes. Los resultados se muestran en la Tabla 16. Destacan las 98.596 proteínas que están tanto en el núcleo como en el citoplasma, o las 83.423 proteínas localizadas tanto en la membrana celular como en el citoplasma. Consideramos que, si una proteína está localizada en diferentes compartimentos celulares, es porque puede estar realizando una función molecular diferente.

Tabla 16: Número de proteínas presentes en diferentes compartimentos celulares (de acuerdo con los códigos GO anotados en la base de datos UniProt)

GO Compartment 1	GO Compartment 2	Number of proteins
membrane [0016020]	cytoplasm [0005737]	83.423
cytoplasm [0005737]	nucleus [0005634]	98.596
membrane [0016020]	nucleus [0005634]	16.694
cytoskeleton [0005856]	nucleus [0005634]	2.353
outer membrane [0019867]	cytoplasm [5737]	362

Tabla 17: Número de proteínas que están en compartimentos celulares no asociados con su función canónica descrita (de acuerdo con los códigos GO anotados en la base de datos UniProt)

GO Canonical Function	GO Compartment	Number of proteins
DNA binding transcription factor activity [0003700]	cytoplasm [5737]	137.615
protein folding [0006457]	membrane [0016020]	12.009
catalytic activity [0003824]	cytoskeleton [0005856]	7.531
isomerase activity [0016853]	membrane [0016020]	6.837
isomerase activity [0016853]	nucleus [5634]	3.510
DNA binding transcription factor activity [0003700]	membrane [0016020]	2.399

Además, combinamos ambos análisis buscando proteínas que están localizadas en un compartimento celular, en el que no están realizando su función canónica. Por ejemplo, encontramos que 137.615 proteínas involucradas en la transcripción del ADN, que ocurre en el núcleo, también se localizan en el citoplasma. Además, encontramos que 7.531 proteínas con actividad catalítica (enzimas) se encuentran en la membrana, que también es una localización inesperada para una enzima.

En nuestra base de datos actual, hemos descrito 694 proteínas moonlighting, en las cuales la bibliografía describía específicamente ambas funciones moleculares. Pero teniendo en cuenta los números anteriores, esperamos que haya un gran número de proteínas moonlighting aún por descubrir. Estas cifras sugieren que el moonlighting no es un fenómeno raro, sinó un mecanismo común, ampliamente utilizado en diferentes especies, para expandir sus funciones disponibles, sin expandir el número de genes y proteínas.

Se debe tener en cuenta que no todas las proteínas tienen su función descrita en la base de datos de UniProt o GO y que estos números pueden ser incluso mayores, ya que cada día se descubren nuevas funciones. Para estimar el número de proteínas que tienen todas sus funciones descritas en UniProt, analizamos manualmente todas las proteínas en nuestra base de datos actualizada. Los resultados se muestran en la Tabla 18 y la Figura 30.

Tabla 18: Porcentaje de funciones detectadas utilizando la descripción de UniProt, descriptores de GO y combinando ambos parámetros

	UniProt Description		GO Descriptors		UniProt & GO	
% OF DETECTION	**Canonical**	**Moonlighting**	**Canonical**	**Moonlighting**	**Canonical**	**Moonlighting**
GENERAL	80,5	46,9	95,3	54,9	95,3	57,4
ONLY REVIEWED	93,2	61	97,1	69,8	100	72,7

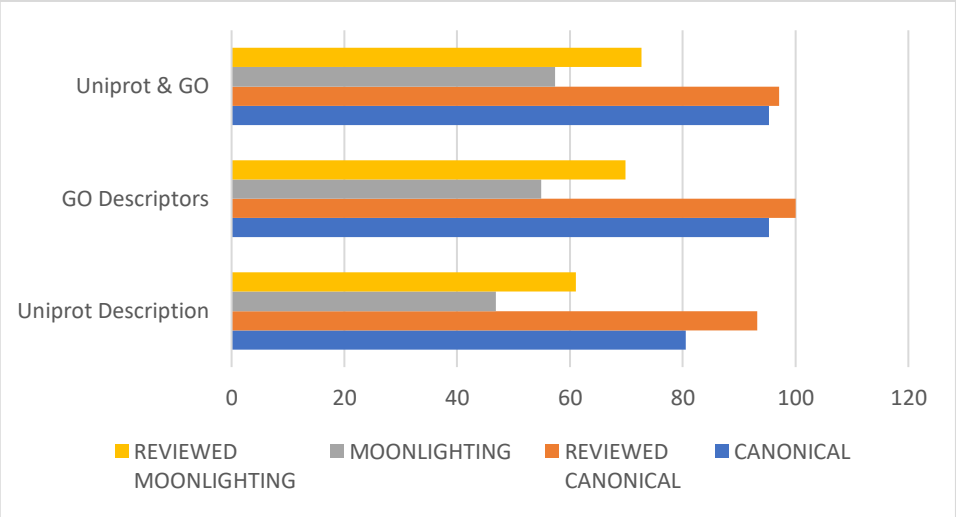


Figura 30: Porcentaje de detección de la función tanto canónica como moonlighting usando UniProt, GO y ambos.

En este análisis, separamos las entradas generales en UniProt y las revisadas. Las entradas revisadas en la base de datos de UniProt son aquellas que han pasado un proceso de revisión por el personal de UniProt, y generalmente contienen más información, y más fiable, sobre las proteínas, por eso el porcentaje de detección es más alto en todos los métodos comparados.

Como puede verse, el uso de GO es un mejor método para detectar el fenómeno moonlighting y las funciones canónicas comparado con la mera descripción de texto en UniProt. UniProt detecta la función moonlighting en el 47% de las proteínas de nuestra base de datos, mientras que GO permite identificar el 55% de las funciones moonlighting. Combinando ambos métodos, este porcentaje aumenta hasta un 57%. Podemos decir que casi la mitad de las proteínas moonlighting de nuestra base de datos no se describen adecuadamente ni en UniProt ni en la base de datos de GO. Este hecho resalta la importancia de tener una base de datos de proteínas moonlighting como la que hemos diseñado. Además, sugiere que las proteínas moonlighting esperadas según el análisis con los códigos GO en las Tablas 14, 15 y 16, podrían ser incluso mayores.

IV.E.1.b. APROXIMACIÓN BASADA EN LAS PALABRAS UTILIZADAS EN LOS DESCRIPTORES QUE UTILIZA LA BASE DE DATOS UNIPROT

En bioinformática, la búsqueda de patrones en las bases de datos existentes es una buena herramienta para el datamining. Al leer la descripción funcional en forma de texto en UniProt, tratamos de identificar una palabra común en las descripciones de las proteínas moonlighting. Las más comunes son las palabras: "also" y "may". Normalmente se usa para decir que estas proteínas tienen una función, pero "also have another function" o "may act performing another function". Analizamos el número de proteínas moonlighting en nuestra base de datos que tienen estas palabras en su descripción de UniProt y encontramos que un 9% tiene la palabra "also" y un 10% tiene la palabra "may". Yendo más allá en esta idea, buscamos en UniProt todas las entradas de proteínas que tienen estas palabras en su descripción. Encontramos que 778.445 proteínas tienen la palabra "also" y 1.176.554 tienen la palabra "may" en la descripción de la función. Obviamente, es un método inexacto que no permite predecir las proteínas moonlighting, pero nos permite imaginar lo grande que es el número de proteínas moonlighting que aún esperan ser reveladas. A pesar de esto, con todos estos números, podemos esperar que se descubra un gran número de proteínas moonlighting en un futuro cercano.

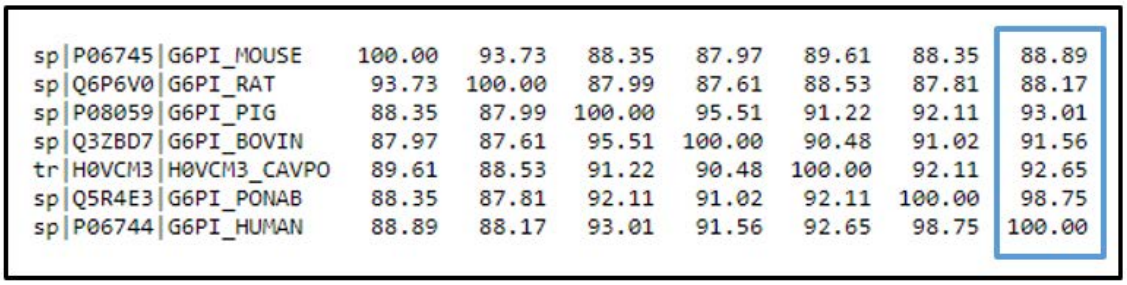
Sin embargo, un problema importante en la ciencia y específicamente en bioinformática es que no tenemos métodos de anotación estandarizados y cada grupo o consorcio está diseñando bases de datos con sus propias reglas. Esto dificulta el uso de métodos automatizados de análisis de datos y es uno de los grandes desafíos del Big Data.

IV.E.2. ¿ESTÁN CONSERVADAS LAS FUNCIONES DE LAS PROTEÍNAS MOONLIGHTING ENTRE ESPECIES?

Esta también es una pregunta ampliamente discutida en el campo de las proteínas moonlighting. En general se considera que no se puede decir que una proteína es moonlighting si esta nueva función no se ha demostrado experimentalmente.

IV.E.2.a. APROXIMACIÓN BASADA EN ALINEAMIENTOS MÚLTIPLES DE SECUENCIAS DE DIFERENTES ORGANISMOS

Si el porcentaje de identidad de las secuencias de proteínas moonlighting en especies cercanas es muy alto, esto sugiere que pueden conservar su multifunción. Por ejemplo, en la Figura 31 podemos observar la matriz de porcentajes de identidades de la proteína Glucose 6 isomerase (G6PI), descrita como proteína moonlighting humana, con las homólogas de organismos cercanos como ratón, rata, etc. A pesar de eso, no podemos decir que G6PI es moonlighting en las otras especies, porque las funciones no han sido descritas experimentalmente.



sp|P06745|G6PI_MOUSE	100.00	93.73	88.35	87.97	89.61	88.35	88.89
sp|Q6P6V0|G6PI_RAT	93.73	100.00	87.99	87.61	88.53	87.81	88.17
sp|P08059|G6PI_PIG	88.35	87.99	100.00	95.51	91.22	92.11	93.01
sp|Q3ZBD7|G6PI_BOVIN	87.97	87.61	95.51	100.00	90.48	91.02	91.56
tr|H0VCM3|H0VCM3_CAVPO	89.61	88.53	91.22	90.48	100.00	92.11	92.65
sp|Q5R4E3|G6PI_PONAB	88.35	87.81	92.11	91.02	92.11	100.00	98.75
sp|P06744|G6PI_HUMAN	88.89	88.17	93.01	91.56	92.65	98.75	100.00

Figura 31: Matriz de identidad de un alineamiento múltiple de secuencias de Glucose 6 isomerase. La proteína humana, además de su función en la glucólisis, actúa como neuroleuquina, factor de motilidad autocrina, en diferenciación y cómo factor de crecimiento nervioso, de estimulación de la migración celular e implantación (en el hurón). También es un importante modulador de la progresión tumoral y una diana para el tratamiento del cáncer, además de mediar en la aglutinación espermática.

Este es un ejemplo de lo que sucede con una gran cantidad de proteínas moonlighting, sus secuencias están altamente conservadas entre especies genéticamente cercanas. Pero, como se describe en la Introducción para la chaperona GroEL, solo un cambio en unos pocos aminoácidos es suficiente para

obtener una nueva función, por lo que es difícil decir que las funciones moonlighting se conservan, incluso en el caso de altas tasas de conservación de secuencias.

Una aproximación para resolver este problema consistiría en comparar secuencias de las regiones específicas involucradas en las diferentes funciones. Sin embargo, los aminoácidos clave para las funciones moonlighting se desconocen (o al menos no se describen) en la mayoría de casos, especialmente para aquellas funciones que no son enzimas (tal y como se ha descrito en el Apartado IV.A.5. y la Tabla 7). Incluso contactando directamente con los investigadores que descubrieron las proteínas moonlighting de nuestra base de datos, en la mayoría de casos no han determinado en qué región de la proteína se localiza la función moonlighting. A pesar de esto, algunas proteínas moonlighting sí que tienen mapadas y descritas ambas funciones en la bibliografía y fueron usadas para estudiar la conservación de las regiones específicas para las funciones moonlighting.

Por ejemplo, la Malate synthase es una enzima que actúa como adhesina y factor de virulencia en *Mycobacterium tuberculosis* y se ha identificado la región responsable de esta función moonlighting. La Figura 32 muestra esta región comparada en algunos microorganismos y se puede observar que esta región está perfectamente conservada en *Mycobacterium bovis* y *Mycobacterium leprae*.

```
sp|P37330|MASZ_ECOLI      FKAASDLIFLGVKQPNGYTEPLLHAWRLREKESH----------------
sp|P42450|MASZ_CORGL      FQAAKDLIFEGTKSPSGYTEPILHARRREFKAKN----------------
sp|Q9I636|MASZ_PSEAE      FQAALELVVEGTRQPNGYTEPVLHRRRREFKAKNGL--------------
sp|P9WK17|MASZ_MYCTU      FLAAQELILSGAQQPNGYTEPILHRRRREFKARAAEKPAPSDRAGDDAAR
sp|P0A5J5|MASZ_MYCBO      FLAAQELILSGAQQPNGYTEPILHRRRREFKARAAEKPAPSDRAGDDAAR
sp|B8ZSN3|MASZ_MYCLB      FLAAQELILSGAQQPNGYTEPILHRRRREFKAQNR----------------
                          * ** :*:. *.:.*.*****:**  * . *
```

Figura 32: Alineamiento de secuencias de la región específica para la función de adhesión al huésped descrita en *Mycobacterium tuberculosis*. Se puede observar como esta región está conservada en microorganismos cercanos.

Otro ejemplo es la Mitogen-activated protein kinase 1, que se describe como proteína moonlighting en humanos y que regula la señalización del interferón gamma además de su propia función de unión a DNA y regulación de la transcripción.

```
sp|O94737|MAPK1_PNECA      CGLKYIHSANVLHRDLKPSNLLINADCKLKICDFGLSRGISVNQGQGTEYMTEYVTTRWY
sp|P26696|MK01_XENLA       RGLKYIHSANVLHRDLKPSNLLLNTTCDLKICDFGLARVADPDH-DHTGFLTEYVATRWY
sp|P63086|MK01_RAT         RGLKYIHSANVLHRDLKPSNLLLNTTCDLKICDFGLARVADPDH-DHTGFLTEYVATRWY
sp|P63085|MK01_MOUSE       RGLKYIHSANVLHRDLKPSNLLLNTTCDLKICDFGLARVADPDH-DHTGFLTEYVATRWY
sp|P28482|MK01_HUMAN       RGLKYIHSANVLHRDLKPSNLLLNTTCDLKICDFGLARVADPDH-DHTGFLTEYVATRWY
sp|P46196|MK01_BOVIN       RGLKYIHSANVLHRDLKPSNLLLNTTCDLKICDFGLARVADPDH-DHTGFLTEYVATRWY
sp|Q84UI5|MPK1_ORYSJ       RGLKYIHSANVLHRDLKPSNLLLNANCDLKICDFGLARTTS-----ETDFMTEYVVTRWY
sp|Q39021|MPK1_ARATH       RGLKYIHSANILHRDLKPGNLLVNANCDLKICDFGLARASNT----KGQFMTEYVVTRWY
                            *********:*******.***:*: *.********:*  .        ::****.****

sp|O94737|MAPK1_PNECA      KIKSASAQSYIRSLPTLPKMPYSKIFPYANPDALDLLNCLLTFDPYDRISCEEALEHPYL
sp|P26696|MK01_XENLA       CIINLKARNYLLSLPHKNKVPWNRLFPNADPKALDLLDKMLTFNPHKRIEVEAALAHPYL
sp|P63086|MK01_RAT         CIINLKARNYLLSLPHKNKVPWNRLFPNADSKALDLLDKMLTFNPHKRIEVEQALAHPYL
sp|P63085|MK01_MOUSE       CIINLKARNYLLSLPHKNKVPWNRLFPNADSKALDLLDKMLTFNPHKRIEVEQALAHPYL
sp|P28482|MK01_HUMAN       CIINLKARNYLLSLPHKNKVPWNRLFPNADSKALDLLDKMLTFNPHKRIEVEQALAHPYL
sp|P46196|MK01_BOVIN       CIINLKARNYLLSLPHKNKVPWNRLFPNADSKALDLLDKMLTFNPHKRIEVEQALAHPYL
sp|Q84UI5|MPK1_ORYSJ       FVNE-NARRYIRQLPRHARQSFPEKFPHVHPLAIDLVEKMLTFDPRQRITVEGALAHPYL
sp|Q39021|MPK1_ARATH       FIDNPKAKRYIRSLPYSPGMSLSRLYPGAHVLAIDLLQKMLVFDPSKRISVSEALQHPYM
                            : . .*: *: .**        . :* ..  *:**:: :*.*:* .**  . ** ***:
```

Figura 33: Alineamiento múltiple de secuencias (parcial) de la proteína MAPK1 de diferentes especies. Marcados en azul están los aminoácidos importantes para la función canónica (represor de la transcripción) y en amarillo la región involucrada en la función moonlighting (para regular la señalización del interferón gamma).

Como puede verse en la Figura 33, todos los organismos comparados tienen los aminoácidos clave para la función canónica conservada. Los aminoácidos relacionados con la función moonlighting también se conservan en organismos genéticamente cercanos como: *Xenopus laevis, Rat, Mouse* y *Bovin*, pero no se conservan en organismos como *Arabidopsis thaliana* o *Pneumocystis carinii*. Esto se debe a que la función moonlighting (regulación de la señalización del interferón gamma) no se puede realizar en microorganismos o vegetales, pero sí en animales. Que la conservación de la región involucrada en la función moonlighting aparezca solo en los organismos en los que se puede realizar, sugiere que la función moonlighting podría estar conservada en organismos genéticamente cercanos.

Otro ejemplo similar es la proteína Methylthioribose-1-phosphate isomerase (Figura 34), que además de su función canónica está involucrada en la invasión de tejidos en melanoma. Los aminoácidos clave para la función canónica son cys168, asp248, y para la invasión en melanomas lo son ser283, arg109. Los aminoácidos correspondientes a la función canónica se conservan en todos los organismos analizados, pero los relacionados con el cáncer solo se conservan entre humanos y ratas, que son los organismos en los que existe el melanoma. No tiene sentido que un microorganismo o un vegetal tenga esa función relacionada con el cáncer. Pero de nuevo, los aminoácidos clave para la función moonlighting se conservan entre especies genéticamente cercanas, especialmente en especies en las que se puede realizar la función moonlighting. Este hecho refuerza de nuevo la idea de que el fenómeno moonlighting puede conservarse en especies cercanas y de que hay un gran número de proteínas moonlighting aún por descubrir.

Figura 34: Alineamiento múltiple de secuencias de la proteína Methylthioribose-1-phosphate isomerase, donde los aminoácidos clave para la función moonlighting están marcados en verde y para la función canónica en amarillo.

Además, en algunos casos la función moonlighting está más conservada entre diferentes especies que la función canónica. Un ejemplo de esto es la proteína ribosomal S10 de *E. coli,* donde estan conservados los motifs relacionados con las dos funciones e incluso está más conservado el correspondiente a la función moonlighting (Figura 6 de la Introducción).

IV.E.2.b. APROXIMACIÓN BASADA EN LA ORTOLOGIA DE INTERACTOMAS DE DIFERENTES ORGANISMOS

Otra estrategia para comprobar que existe conservación de la multifuncionalidad es la interactómica comparada. Si existe conservación del interactoma, 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). Por ejemplo, si una determinada proteína moonlighting tiene dos funciones A y B, probablemente en su interactoma encontraremos proteínas involucradas en la función A y otras en la función B. Si analizando el interactoma de la misma proteína, pero en otro organismo, encontraremos partners de interacción relacionados con la función A y otros con la B, podríamos pensar que la proteína en este organismo también fuera moonlighting y realizara las dos funciones. En la Tabla 19, se muestran algunos ejemplos del análisis realizado partiendo de las proteínas moonlighting de nuestra base de datos MultitaskProtDB-II. Con fondo azul se muestran las proteínas moonlighting en MultitaskProtDB-II y debajo, sin color de fondo, las proteínas equivalentes en otras especies. Por ejemplo, la Aconitasa cuya función canónica es el paso de citrato a isocitrato, en humano se la ha identificado como proteína de unión a mRNA, si analizamos su interactoma, en efecto encontramos proteínas como la Nuclease-sensitive element-binding protein 1, cuya función es la unión a RNA. Analizando la Aconitasa de otras especies en las que no se ha descrito específicamente la función de unión a mRNA, también encontramos que tienen partners de interacción cuyas funciones implican la unión a mRNA. Es el caso de la Aconitate hydratase de *E. coli, Saccharomyces cerevisiae* y *Rattus norvegicus*. Este hecho sugiere, que las aconitasas de esas especies conservan la función de unión a RNA.

Otro ejemplo que se puede observar en la Tabla 19, es el del Cytochrome c. Esta proteína en humano, además de su función canónica está involucrada en el control de la apoptosis. Analizando sus partners de interacción encontramos proteínas relacionadas con la apoptosis como el TNF receptor-associated factor 6. Siguiendo el mismo proceso que con la proteína anterior, al analizar la misma proteína en otras especies, encontramos que en *Mus musculus* y *Drosophila melanogaster* estas proteínas también presentan partners de interacción

relacionados con el control de la apoptosis. Sugiriendo, de nuevo, que la función moonlighting está conservada en especies cercanas.

Otro caso representativo es la proteína ATF2, cuya función moonlighting en humano es la respuesta al daño en DNA. Tanto la proteína descrita como moonlighting en humano, como las equivalentes en diferentes especies (*Mus musculus* y *Saccharomyces cerevisiae* presentan partners de interacción relacionados con la respuesta celular al daño en el DNA.

En la Tabla 19, se muestran 5 ejemplos en los que la ortología en el interactoma, sugiere que la función moonlighting está conservada en organismos cercanos filogenéticamente. En la última columna de la tabla se puede observar el porcentaje de conservación de las secuencias de las proteínas analizadas respecto a la descrita como moonlighting. Se puede observar como no es necesario un alto porcentaje de identidad para que se conserven las funciones moonlighting de las proteínas. Por ejemplo, la proteína ATF2 cuyos partners de interacción sugieren que en Saccharomyces se conserva la función de reparación de DNA, únicamente tiene un 20,4% de identidad con la secuencia de la misma proteína en humano.

Un total de 61 proteínas moonlighting han mostrado ortología de interactoma con otras proteínas equivalentes en otras especies. (Ver Información Suplementaria S11).

Tabla 19. Ortologia de interactomas en diferentes especies

Protein name	Uniprot	Moonlighting function	Organism	Interactomic proteins	Identity
Aconitase	P21399	mRNA binding	Human	**Nuclease-sensitive element-binding protein 1.** Binds to RNA.	
Aconitate hydratase A	P25516		*E. coli*	**RNA-binding protein YhbY.** RNA binding function. **Carbon storage regulator.** Binds to RNA.	52%
Aconitate hydratase mitochondrial	P19414		*Saccharomyces cerevisiae*	**Importin subunit beta-2.** RNA binding protein.	30,5%
Aconitate hydratase mitochondrial	Q9ER34		*Rattus norvegicus*	**Pre-mRNA-splicing factor ATP-dependent RNA helicase PRP43.** Pre-mRNA processing factor.	28,5%
ATF2 protein	P15336	DNA damage response	Human	**Spliceosome RNA helicase DDX39B.** Negative regulation of DNA damage checkpoint	
ATF2 protein	P16951		*Mus musculus*	**Mitogen-activated protein kinase 1.** Cellular response to DNA damage stimulus. **Histone acetyltransferase p300.** Damaged DNA binding. **Brca1E3.** DNA repair by facilitating cellular responses to DNA damage.	99,2%
ATF2 protein	P53296		*Saccharomyces cerevisiae*	**DNA repair protein RAD33.** DNA repair.	20,4%
Cytochrome c	P99999	Controlling apoptosis	*Homo sapiens*	**TNF receptor-associated factor 6.** Negative regulation of apoptosis.	
Cytochrome c1	Q9D0M3		*Mus musculus*	**Alpha-synuclein.** Reduces neuronal responsiveness to various apoptotic stimuli	16,5%
Cytochrome c1	P04657		*Drosophila melanogaster*	**Serine protease HTRA2.** Promotes or induces cell death.	66%
Presenilin	A9SVI7	Cytoskeletal function	*Physcomitrella patens*	*No interactions in APID*	
Presenilin-1	P49768		*Homo sapiens*	**Arc.** Regulation of cytoskeleton.	23,5%
Presenilin-1	P49769		*Mus musculus*	**Arc.** Regulation of cytoskeleton.	23,5%
Thioredoxin	P0AA25	Omega DNA Polymerase subunit	*E. coli*	**DNA-directed DNA polymerase.**	
Thioredoxin	P10599		*Homo sapiens*	**Genome polyprotein**. RNA polymerase.	26,9%
Thioredoxin-1	P47938		*Drosophila melanogaster*	**DNA-binding protein Ewg.** DNA binding function.	24,7%
Thioredoxin-2	Q9V429		*Drosophila melanogaster*	**Proliferating cell nuclear antigen.** Auxiliary protein of DNA polymerase delta	32,4%
Thioredoxin	P22217		*Saccharomyces cerevisiae*	**DNA-directed RNA polymerase III.** RNA polymerase component.	40,2%

IV.E.3. PROTEÍNAS MOONLIGHTING Y DESPLAZAMIENTO DE GEN NO ORTÓLOGO

El Non-Orthologous Gene Displacement, como se describe en la Introducción, o NOGD, describe una forma variante de un sistema o vía en la que un componente esperado se reemplaza por un equivalente funcional que difiere en su origen evolutivo. Estos huecos funcionales se encuentran bioinformáticamente al secuenciar el genoma completo de un organismo. Por ejemplo, en una vía metabólica encontramos que falta el gen para una de las enzimas clave, en cambio sabemos que esa vía metabólica está funcionando en ese organismo. Así pues, se supone que la función de ese gen la está realizando otra proteína, que no es la esperada. Este hecho plantea una relación clara con la moonlighticidad. Estas enzimas esperadas en una ruta pueden ser reemplazadas por otra proteína, que está haciendo otra función, pero tiene una estructura similar a la proteína que va a sustituir. En estos casos estaríamos hablando claramente de una proteína moonlighting.

Galperin y Koonin crearon una base de datos de proteínas NOGD (Galperin et al., 1998). Esta base de datos enumera las proteínas que difieren en sus estructuras, pero realizan las mismas funciones, por lo que se espera que sean ejemplos de NOGD con un origen evolutivo distinto. Hemos comparado las proteínas de ambas bases de datos, NOGD y MultitaskProtDB-II, y obtuvimos 20 proteínas que son a la vez ejemplos de moonlighting y NOGD, como se muestra en la Tabla 20.

Tabla 20: Ejemplos de coincidencia entre proteínas moonlighting y NOGD

Protein. Name	Organisms in which is moonlighting	NOGD described organisms
Peptidyl Prolyl cis,trans-Isomerase (PPI)	Helicobacter pilory	E. coli
Phosphofructokinase	Pichia pastoris, Bacillus subtilis	E. coli
DNA primase	Pyrococcus abyssi	Ecoli, Sulso, Human
cPrxI (Peroxiredoxin TSA1)	Saccharomyces cerevisiae, Candida albicans	Aerpe, Myctu, Ecoli
Phosphoglucose isomerase	**Human, G. stearothermophilus, Xanthomonas oryzae Echinococcus multilocularis, Lactobacillus crispatus**	**Homo sapiens, Theli**
Peroxiredoxin-6	Homo sapiens	Aerpe, Myctu, Ecoli
Fumarate hydratase	Homo sapiens	E. coli
DHPR (peptide C) dihydropteridine reductase	**Homo sapiens**	**Homo Sapiens, E. coli**
Peptidyl-prolyl cis-trans isomerase A	Homo sapiens	Ecoli
Upsilon-crystallin	Ornithorhynchus anatinus	Human, Raleh
Dihydrofolate-reductasethymidylate-synthase (DHFR–TS)	Arabidopsis thaliana	E.coli, Bacillus subtilis
Glutamate dehydrogenase	**Bacillus subtilis**	**Bacsu, Achkl**
Chorismate mutase	Xanthomonas oryzae	Entag, Bacsu
Alcohol acetaldehyde dehydrogenase	Listeria monocytogenes	Psepu, Ecoli
cAMP phosphodiesterase	Mycobacterium tuberculosis	Human, Yeast
Superoxide dismutase	Mycobacterium avium, Mycobacterium tuberculosis	Ecoli, Strso
LytC (Lysozyme)	Streptococcus pneumoniae	Cloab, BPP1, Baciu
Endothelial Nitric Oxide Synthase	Homo sapiens	Rat, Arath
Developmental protein eyes absent	Drosophila melanogaster	Human, Strco, Bacsu, Mouse
Dihydrofolate-reductasethymidylate-synthase (DHFR–TS)	Arabidopsis thaliana	Bacsu, Ecoli

Por ejemplo, en el caso de la DHPR Dihydropteridine reductase, que es moonlighting en humanos, se encuentran dos formas diferentes de proteínas, aquellas similares a la estructura de *Escherichia coli* y aquellas similares a la proteína de *Homo sapiens*. A pesar de que ambas proteínas están realizando la misma función molecular, tienen una estructura totalmente diferente, como se puede ver en el siguiente alineamiento (Figura 35).

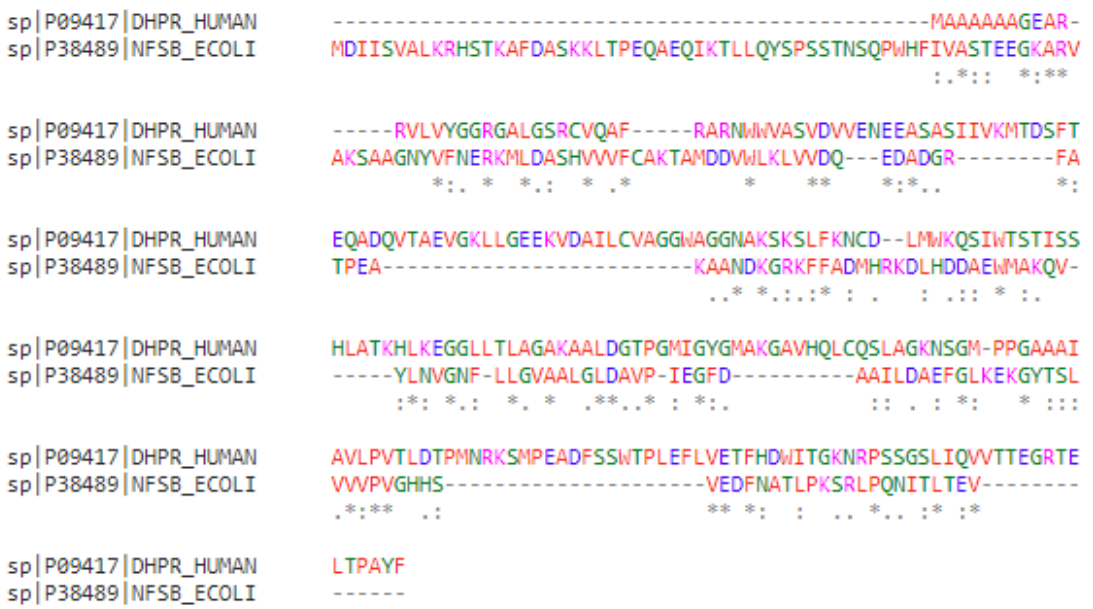


Figura 35: Alineamiento entre las secuencias de las proteínas DHPR de humano y *E. coli*, que determina solo un 24% de identidad entre ambas secuencias de proteínas.

A pesar de eso, la DHPR está perfectamente conservada en especies cercanas a la forma humana, como puede verse en la Figura 36, lo que sugiere que en este caso provienen evolutivamente de la misma proteína ancestral, pero difiere evolutivamente de las proteínas similares a las de *E. coli*. En todo caso, podemos considerar que las proteínas multifuncionales son solo aquellas similares a la forma en la que se ha probado la moonlighticidad.

Figura 36: Alineamiento entre las secuencias de la proteína DHPR de humano y las de organismos genéticamente cercanos. Se puede apreciar una alta homología de secuencia, cercana al 100%.

Los ejemplos anteriores se obtuvieron buscando qué proteínas estaban a la vez en nuestra base de datos de proteínas moonlighting y en la de NOGD. El nombre de una proteína habitualmente viene dado por su función canónica, que fue la primera descrita, pero se quiso buscar si también habría proteínas cuya función moonlighting coincide con el nombre de alguna proteína considerada NOGD. Los resultados encontrados se encuentran en la Tabla 21. Por ejemplo, la Cyclooxygenase-1 es además una Heme-dependent peroxidase y estas son proteínas NOGD.

Tabla 21: Coincidencias entre las funciones moonlighting y las proteínas identificadas como NOGD

Prot. Name	Moonlighting function	Non orthologous displacement protein	Moonlighting in	NOGD in
Upsilon-crystallin	Lactate dehydrogenase A	D-lactate dehydrogenase	Ornithorhynchus anatinus	E. Coli
Cyclooxygenase-1	**Heme-dependent peroxidase**	**Peroxidase Heme-dependent EC: 1.11.1.7**	**Homo sapiens**	**Homo sapiens**
Sarcosine oxidase	NADH dehydrogenase; L-proline dehydrogenase	NADH dehydrogenase EC: 1.6.99.3	Thermococcus kodakarensis	E. Coli, Homo sapiens
Ure2 (Transcriptional regulator URE2)	Glutathione peroxidase activity	Glutathione transferase EC: 2.5.1.18	Saccharomyces cerevisiae	E. Coli, Homo sapiens, Serma
Peroxiredoxin-6	**Phospholipase aiPLA2**	**Phospholipase A2 EC: 3.1.1.4**	**Saccharomyces cerevisiae, Human**	**Human, E. coli**
Trex (Alpha-1,4-Transferase)	Alpha-1,6-Glucosidase	Alpha-glucosidase	Sulfolobus solfataricus	Canal, Thema, Strmu

Nos proponemos realizar posteriormente un análisis evolutivo siguiendo esta relación entre las proteínas moonlighting y el NOGD, lamentablemente el número de NOGD, es limitado.

V. DISCUSIÓN GENERAL

Parece lógico pensar que una proteína pueda presentar más de una función ya que la vida es económica e intenta reutilizar lo que ya ha sido previamente ensayado por la evolución y funciona correctamente. El diseño de una nueva función proteica no es fácil y una buena estrategia es el reclutamiento de alguna estructura preexistente, cuyo plegamiento, estructura y estabilidad ha sido previamente ensayada por la naturaleza. Crear una nueva proteína desde cero es más complejo que crear nuevos genes que no tienen tales restricciones de plegamiento y estabilidad. Los genes, el ADN, tienen una química simple y reiterativa y, por lo que respecta al conocimiento actual sobre ellos, la función principal es lineal y, en realidad, los efectos de sus modificaciones se confrontan no a nivel del propio ácido nucleico sinó a nivel de la proteína en él codificada. Por otra parte, existen mecanismos de cortar y pegar ácidos nucleicos desde los organismos procariotas, aunque el splicing es, por ahora, el más sofisticado de los mismos.

Las proteínas carecen de mecanismos similares. Las proteínas presentan un proceso complejo para su biosíntesis y plegado, de hecho, un tercio de los polipéptidos abortan durante el proceso de la síntesis proteica. También hacen frente a un ambiente celular muy concurrido, han de situarse en uno o más compartimentos concretos, han de disponer de modificaciones post-traduccionales altamente específicas, y deben establecer interacciones precisas con los partners adecuados antes de poder realizar una función biológica concreta. Probablemente, un punto esencial para adquirir una nueva función sea adaptarse a un nuevo sitio de interacción. Por lo tanto, incorporar una segunda función requiere superar y ajustar muchas restricciones para adaptarse a las nuevas propiedades biológicas. Finalmente, si una proteína es multifuncional existirá el denominado *conflicto adaptativo*, por el que incorporar cambios en una secuencia de proteína para mejorar una función, puede causar un deterioro de la otra función. Por todo ello una pregunta obvia es ¿Cuál es la ventaja de que una proteína tenga más de una función? Aparte de lo comentado de reutilizar secuencias y estructuras ya ensayadas por la naturaleza se requiere un menor número de genes para codificar las funciones requeridas por un organismo. Esto

es especialmente útil en células mínimas, como son los micoplasmas. Incluso en organismos eucariotas, aunque aparentemente “sobra” DNA, cada vez hay más evidencias que el DNA no codificante de proteínas tiene importantes funciones reguladoras. La multifuncionalidad de las proteínas aumenta la cantidad de funciones sin aumentar el número de genes, tener un número de componentes o una complejidad excesiva. En todo caso, aumentar el número de funciones aumentando el número de genes conllevaría un incremento de la carga mutacional.

En un tiempo en que el mecanismo de splicing podría dar lugar a genes que codifican diversas proteínas, el que una proteína pueda tener más de una función no da lugar a una gran sorpresa. Pero ya a partir de los resultados del Primer Borrador y el Atlas del Proteoma Humano, no parecía haber tantas proteínas como se pensaba (primero se estimó al menos un millón) (Kim et al., 2014; Wilhelm et al., 2014; Uhlen et al., 2015). Pero a partir de los recientes análisis proteómicos a gran escala, a pesar de que a nivel de RNA se detecten transcritos alternativos para el 72% de los genes humanos, no se encuentran a nivel de proteína las correspondientes isoformas. De hecho, en la mayoría de casos se encuentra tan sólo una isoforma proteíca preferente. Únicamente se encuentran alternativas proteícas para 246 genes (Tress et al., 2017). Esto implica que el mecanismo de splicing puede no ser la única clave para explicar la complejidad del proteoma humano (y mamíferos en los que se ha experimentado, como el ratón). Esto da un valor adicional a la multifuncionalidad de las proteínas y sugiere que debe ser un fenómeno más general de lo que se pensaba hasta ahora y puede contribuir a explicar la complejidad de los organismos vivos.

Todo esto hace necesario optimizar las proteínas para lograr el máximo número de funciones biológicas. El concepto de función biológica de una proteína no es algo simple y se suele utilizar una definición instrumental, operativa. En general se considera una única función bioquímica (la *Molecular Function* del servidor GO) pero en realidad –y especialmente tras descubrir la existencia de las proteínas moonlighting– corresponde a un criterio jerárquico y es además dependiente del contexto en que está, cosa que no se suele tener en cuenta. El

éxito de las técnicas de DNA recombinante ha inducido a la idea de que la función biológica de un gen es trasladable a otro entorno, y esto puede funcionar en la clonación y expresión de proteínas en forma heteróloga, por ejemplo, una proteína humana en *E. coli* (si no es tóxica y si se consigue expresar). Pero una evidencia de que el contexto es clave, sucede en el caso de genes/proteínas altamente conservadas que en diferentes organismos tienen funciones diferentes o algo diferentes pero relacionables. Por ejemplo, el gen que determina la dirección de la gravedad en plantas, en humano va ligado a sordera (en el oído hay el sentido del equilibrio), y el gen que codifica un sensor de la luz en plantas, en humano está involucrado en el ritmo circadiano (McGary et al., 2010).

En el presente trabajo hemos actualizado la que ha sido la primera y más amplia base de datos de proteínas multifuncionales existente, MultitaskProtDB (Hernández et al., 2014) que hemos denominado MultitaskProtDB-II (Franco-Serrano et al., 2018). Como ya se ha descrito a lo largo del presente trabajo, esta base de datos es de gran ayuda para identificar las características de las proteínas multifuncionales y profundizar en su función biológica, su evolución y su relación con los mecanismos moleculares de las enfermedades humanas, las dianas de fármacos y los factores de virulencia de microorganismos patógenos –y por ello el diseño de vacunas recombinantes por subunidades– (Franco-Serrano et al., 2018). Representa una fuente de información experimentalmente contrastada y a partir de la cual se pueden establecer hipótesis así como sugerir nuevas aproximaciones. En el trabajo se han presentado algunas de estas ideas y en el futuro estamos diseñando nuevas, especialmente desde el punto de vista evolutivo y de aparición de las funciones moonlighting a lo largo de la escala filogenética, del fenómeno del non-orthologous-gene-displacement, etc.

Otro objetivo del grupo es la predicción bioinformática de las proteínas moonlighting y de sus características como el interés evolutivo y clínico. La predicción y determinación experimental de la función de una proteína es una tarea difícil. Tan solo vale la pena recordar que, a tres proteínas con una gran importancia clínica, investigadas a lo largo de muchos años en numerosos

grupos académicos e industriales, como son las proteínas relacionadas con el Alzheimer, el Prion y la Corea de Huntington, aún a estas alturas se desconoce en detalle su función biológica. Y todavía es más difícil identificar las funciones 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 investigadores inmersos en tareas como la anotación funcional de nuevos genomas, la interpretación de experimentos de noqueo de genes en casos en que la eliminación de un gen produce resultados fenotípicos inesperados, como hemos descrito. También es clave para la identificación de dianas terapéuticas, diseño de fármacos y vacunas, estudio efectos secundarios de los fármacos, etc.

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 utilizando tan sólo este algoritmo es difícil para el investigador identificar las mejores dianas candidatas de entre la larga lista de dianas que contiene 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. Por ello, la combinación de diferentes algoritmos y bases de datos bioinformáticas 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 homología remota (PSI-Blast o HMMER) con los resultados existentes en las bases de datos de interactómica (PPIs). Las bases de datos de motifs/dominios funcionales (InterPro rastrea la mayoría), especialmente la no curadas (Blocks, PFamB), son también de ayuda en algunos casos (Gómez et al, 2003 y 2011; Hernández et al., 2014b, 2015). Hemos determinado (Tabla 6) que esta combinación 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). La sensibilidad aumenta utilizando únicamente o PSI-Blast o PPIs, pero en estos casos disminuye también la especificidad, dando lugar a más falsos positivos. También hay que tener en cuenta que en nuestros

análisis hemos utilizado una base de datos de proteínas donde previamente se ha demostrado que las proteínas son moonlighting. Desarrollar esta tarea para proteínas desconocidas es más difícil y además en último término requiere determinación experimental. En todo caso, PSI-Blast identifica mejor las proteínas moonlighting que son multidominio y en las que cada dominio presenta una función independiente en vez de aquellas en que las dos funciones se solapan. La Figura 16 que pertenece a 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. De todas formas, 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. Y en todo caso, la mayoría de los métodos utilizados no son fáciles de implementar como métodos automáticos, ya que requieren la interpretación de un lenguaje muy ambiguo y lleno de sinónimos. Hay que tener en cuenta que, con el fin de encontrar la función correcta, suele ser necesario leer la documentación complementaria asociada y no sólo la del descriptor principal de la función asociada con el código GO.

Otro enfoque adicional para predecir una proteína como verdadera moonlighting es la alineación de la secuencia y estructura de la proteína 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 los programas PiSite y Phyre2, siendo el primero el que nos ha dado lugar a los mejores resultados, y como ya se ha descrito en el trabajo, presenta una interesante aplicación para comprender los mecanismos moleculares de algunas enfermedades humanas. Como en el caso del algoritmo PSI-Blast, PiSite 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 se solapan. Finalmente, a partir de este trabajo podemos sugerir otro criterio adicional para la predicción de la multifuncionalidad y es el de que si una proteína es moonlighting sus ortólogas con buen nivel de similitud secuencial probablemente también lo serán.

DISCUSIÓN GENERAL

Como se ha descrito en el Apartado IV.C., algunas bases de datos contienen información sobre proteínas moonlighting que ha pasado desapercibida. Una de ellas es la base de datos de enfermedades genéticas OMIM. Otro ejemplo son los descriptores funcionales GO de la base de datos UniProt. Probablemente otras bases de datos, además de éstas y PubMed, contienen información útil que el datamining puede revelar. Pero como ya se ha comentado, la demostración final para determinar que una proteína es moonlighting debe ser experimental. 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.

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 moonlighting. A algunas de estas preguntas hemos intentado dar respuesta, proponer sugerencias o incluso, plantear especulaciones, durante la tesis. Por ejemplo: ¿Qué ventaja evolutiva representa el que una proteína tenga más de una función? A veces ambas funciones son indispensables, y estan en la misma proteína en vez de utilizar la duplicación de un gen y evolución independiente del parálogo, lo cual evitaria el denominado “conflicto adaptativo”. O: ¿Cuál es el mecanismo que conduce a la aparición de la segunda función? ¿Hay conservación filogenética de la multifuncionalidad? ¿Es la multifuncionalidad un fenómeno general o solo aparece en casos excepcionales? ¿Qué relación hay entre las proteínas moonlighting, las enfermedades humanas y las dianas terapéuticas de fármacos? ¿Por qué los microorganismos patógenos utilizan tanto las proteínas moonlighting como factores de virulencia? Aunque en el presente trabajo no se puede dar una respuesta completa a todas estas preguntas sí merece la pena dedicarles algunos comentarios. Vamos a ampliar lo descrito en la Introducción y Resultados y, especialmente, 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 comentado al inicio de esta discusión que existe una doble respuesta: (a) que con menos genes el organismo pueda 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. Esto lleva a plantear otra pregunta: ¿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 con las proteínas ortólogas 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 por lo que no interfieren con la función canónica. De hecho, los resultados de la ingeniería de proteínas demuestran que la mayor parte de las mutaciones creadas 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 (método de correlación mutacional), como hemos descrito previamente en algunos trabajos (Hernandez et al., 2014 y 2015). En un cierto número de casos ni siquiera se requiere cambio alguno, bastando una nueva localización celular, por ejemplo, en el caso de las numerosas proteínas relacionadas con la virulencia de los microorganismos patógenos. 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. Esto tiene una importante repercusión para la Biología de Sistemas.

Respecto a cómo aparece la segunda función, esto aún no se ha descrito, pero ya se ha mencionado anteriormente el caso de la chaperona GroEL en la que mutando 4 aminoácidos se da lugar a una segunda, una función toxina de insectos. También tendrían lugar casos de Non Orthologous Gene Displacement o de reclutamiento enzimático. Como se ha descrito en el Apartado IV.E., se trata de un mecanismo que permite explicar un cierto número de casos, pero la limitación tanto en las bases de datos de proteínas moonlighting como en las de NOGDGs que no son suficientemente amplias, dificulta encontrar suficientes solapamientos atribuibles al proceso NOGD. Este es un modelo y mecanismo de evolución que se proseguirá estudiando después del presente trabajo, tratando de encontrar nuevas aproximaciones computacionales a esta hipótesis. En un cierto número de casos, la adquisición de la segunda función tiene lugar a través 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 funciones en la secuencia/estructura de las proteínas de nuestra 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. Hemos descrito otra aproximación basada en mapar las funciones mediante programas de superposición de estructuras 3D de proteínas mediante PiSite. Por otra parte, del análisis de nuestra base de datos se ha constatado que los tipos y frecuencias de folds que presentan las proteínas multifuncionales son estándar. Una conclusión evidente es que existen diversos mecanismos por los cuales se consigue la segunda, o más, funciones. Pero no es un tema sencillo pues no podemos trazar fácilmente procesos evolutivos, ni incluso con aproximaciones bioinformáticas y experimentales combinadas. Este es otro tema en el que se proseguirá investigando tras la finalización del presente trabajo.

Respecto a la conservación filogenética de la multifuncionalidad, no se conoce apenas nada dado que habría que realizar experimentación en cada organismo en que se sospeche. Tras los resultados que hemos descrito en el Apartado IV.E., podemos sugerir que el fenómeno de la multifuncionalidad es más común de lo que por el momento conocemos. A partir de tres aproximaciones descritas en los Apartados IV.E., (a) alineamiento múltiple de secuencias de organismos

cercanos, (b) ortología de interactoma y (c) descriptores GO de las proteínas depositadas en la base de datos UniProt, todo indicaría que hay muchísimas proteínas moonlighting aunque no haya experimentación directa acerca de ellas. Dadas las ventajas que hemos descrito es lógico esperar que la vida utilice esta posibilidad, facilitada incluso por la compartimentación celular. En muchos casos bastaría con favorecer una nueva interacción (y no se requieren muchos aminoácidos) para adoptar una nueva función. La inmensa cantidad de áreas cóncavas y convexas y la gran variedad de aminoácidos expuestos en ellas lo facilitarían. Cabe resaltar los casos de las enzimas de la glicólisis o del ciclo de Krebs relacionadas con virulencia en los que se encuentran numerosos casos conocidos que presentan ambas funciones en diferentes microorganismos patógenos. Por ejemplo, la Enolasa es una proteína moonlighting en 42 especies de microorganismos y GADPH en 27 especies (Tabla 11 e Información Suplementaria S7). Asimismo, en la Figura 6 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. Además, 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”. Como se ha descrito en el presente trabajo hemos abordado el problema a partir de la información existente en la base de datos UniProt. Especificamente, a partir de la existencia de duplicidad de los identificadores GO para *Molecular Function* en una muestra de pares de funciones (por ejemplo para *enzyme and transcription factor*. Hemos encontrado miles de ejemplos en los que los investigadores han descrito posibles funciones moonlighting sin utilizar este término, o equivalentes como multitasking, y por ello han escapado a la búsqueda bibliográfica en PubMed, principal aproximación para crear nuestra base de datos. Los resultados de ortología de interactoma también apuntarían sobre la hipótesis de que las proteínas multifuncionales serían abundantes. Pero

en todo caso la 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 4 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. Por ejemplo, en el caso de microorganismos, en los mecanismos de formación de biofilms, en la supervivencia a condiciones anormales del medio, en la infección, etc, que no se identifican en los medios de cultivo "normales". O sea, en muchos casos las funciones moonlighting corresponden a funciones de difícil identificación funcional. En este sentido cabe mencionar que los microorganismos modelo de célula/genoma mínimo, los micoplasmas, presentan, dependiendo de la especie, entre un 25-42% de sus genes huérfanos y sin función conocida.

Un interesante resultado del presente trabajo es la relación entre proteínas moonlighting, enfermedades humanas, dianas de fármacos y virulencia de microorganismos patógenos. Al cotejar las bases de datos de enfermedades genéticas (OMIM, HMGD) con las proteínas moonlighting humanas de nuestra base de datos, vemos que un 78% de las mismas están involucradas en enfermedades. Asimismo, un 48% de las proteínas moonlighting humanas son dianas de fármacos existentes, de acuerdo con las bases de datos de fármacos TTD y DrugBank. Respecto al primer resultado se puede considerar como lógico ya que si una proteína tiene más de una función tendrá más probabilidad de estar relacionada con más de una enfermedad. Pero lo segundo es más difícil de explicar dado que las proteínas que son dianas de fármacos, *druggable proteome*, son un subconjunto del proteoma que presenta un cierto número de restricciones como el que sea una proteína accesible, capaz de una interacción

eficiente y no tóxica con el fármaco. Aunque parte de la respuesta va ligada a lo anterior: Si hay, significativamente, muchas proteínas moonlighting humanas relacionadas con enfermedades, también deberá haber más fármacos específicos para ellas. Pero aun así no queda claro que se justifique el 48% de dianas farmacéuticas. En todo caso, el que un número tan importante de fármacos existentes interaccionen con proteínas diana moonlighting implica que el fenómeno de la multifuncionalidad es más común de lo que se creía. En conjunto, estos datos sugieren otra vez que las proteínas moonlighting no son una excepción, sinó que deben de ser muy abundantes. Finalmente, constatar que en el caso de las enfermedades hay predominancia de las que están relacionadas con la función canónica que con la moonlighting, si bien hay muchas que presentan dos enfermedades, una con cada función, canónica y moonlighting.

Otro resultado interesante es el que un 25% de todas las proteínas de la base de datos son factores de virulencia de microorganismos patógenos. En realidad, este hecho requiere explicar varios aspectos: (a) ¿Por qué los microorganismos patógenos utilizan tanto las proteínas moonlighting como factores de virulencia? (b) ¿Cómo son secretadas siendo proteínas que carecen de péptido señal u otros motifs para anclarse o atravesar las membranas? (c) ¿Cómo es que una misma especie utiliza más de una proteína moonlighting como factor de virulencia? Y a la vez, ¿Cómo es que diferentes especies utilizan las mismas proteínas moonlighting como factor de virulencia? (d) ¿Cómo proteínas secuencial y estructuralmente diferentes interaccionan con las mismas dianas del huésped (PLG, ECM…)? En el Apartado IV.D. hemos sugerido la respuesta al punto (a), que explica que al tratarse de enzimas y proteínas del metabolismo ancestral están muy conservadas entre organismos Procariotas y Eucariotas y, de acuerdo con un trabajo publicado anteriormente por nuestro grupo (Amela et al., 2007), esto inhibiría la respuesta inmune protectiva por parte del huésped. La pregunta (b) no requiere respuesta muy elaborada pues se conocen otros mecanismos de secreción proteíca no estándar. Las preguntas (c) y (d) son de difícil respuesta. Como ya se ha descrito en el apartado IV.D. de Resultados, hemos estado buscando motifs y minimotifs, dominios, etc, compartidos entre las diferentes proteínas moonlighting conocidas. Tan sólo hay descrito un motif triple en las

Enolasas, pero no lo comparten las otras enzimas. Nuestra hipótesis es que el tapizado de la membrana externa del microorganismo lo camufla por la razón descrita en (a) y, a la vez, inhibe la respuesta inmune protectiva del huésped, vía IL10 u otras citoquinas e interleucinas. Es una hipótesis que trataremos de estudiar en una etapa posterior. En todo caso, se trata de un resultado con importantes consecuencias en el campo de la vacunología inversa y en el diseño de vacunas recombinantes por subunidades: No es aconsejable utilizar como proteínas vacunales aquellos factores de virulencia que sean proteínas moonlighting o que tengan un cierto nivel de homología de secuencia con alguna proteína del huésped.

CONCLUSIONES

1.- Se ha actualizado la base de datos MultitaskProtDB de proteínas moonlighting o multifuncionales, creada por nuestro grupo en 2014. En la nueva versión de la base de datos, denominada MultitaskProtDB-II y publicada en NAR Database Issue, se ha más que duplicado el número de entradas e incluido información importante sobre la relación de cada proteína con enfermedades humanas y dianas de fármacos. Además, la base de datos permite conexiones a muchas otras bases de datos como UniProt, GO o PDB. Por su relación con otros resultados y conclusiones, es de resaltar que el análisis de las características de las proteínas de la base de datos indica que el par de clases funcionales (función canónica/función moonlighting) más numeroso es el de enzima/factor de transcripción seguido por el de enzima/factor de adhesión.

Franco-Serrano L., *Hernández S., Calvo A., Severi MA., Ferragut G., et al., (**2018**) MultitaskProtDB-II: an update of a database of multitasking/moonlighting proteins. Nucleic Acids Res. 46: D645-D648*

2.- Desde el punto de vista de la predicción bioinformática de la función de las proteínas y como continuación de los resultados previos del grupo, se ha corroborado que la mejor estrategia para el propósito es combinar el análisis de homología remota con los resultados existentes en bases de datos de interactómica. Tanto en el caso de las bases de datos de interactómica como en las de dominios funcionales, predicen mejor aquellas que son poco curadas como por ejemplo Blocks, PFamB, APID o Biogrid. Esto es así dado que en las bases de datos más curadas se han eliminado aquellas dianas que, desde los actuales conocimientos de la Bioquímica, se consideran falsos positivos. Dentro de las bases de datos de identificación de “motifs” ha resultado también útil MinimotifMiner que, además de secuencias relacionadas con función, identifica aquellas relacionadas con interacción, con modificación post-traduccional, etc.

*Hernández S., Calvo A., Ferragut G., **Franco L**., Hermoso A. et al. (**2014**) Can bioinformatics help in the identification of moonlighting proteins? Biochem. Soc. Trans. 42:1692-7*

CONCLUSIONES

***Franco L**., Hernández S., Calvo A., Ferragut G., Amela I., Cedano J. (**2016**) Moonlighting proteins: a bioinformatics analysis of their biochemical characteristics. New Biotechnology. 33: 432-433*

3.- También desde el punto de vista de la predicción bioinformática de la función moonlighting y utilizando la estructura tridimensional de la proteína, se ha propuesto una estrategia basada en el uso de los programas de modelado PiSite y Phyre2 que nos permite mapar las posibles funciones de la proteína moonlighting sobre su secuencia y estructura. Esto tiene especial interés para el análisis de la relación entre proteínas multifuncionales, enfermedades humanas y dianas de fármacos.

*Hernández S., **Franco L**., Calvo A., Ferragut G., Hermoso A. et al. (**2015**) Bioinformatics and Moonlighting Proteins. Frontiers Bioeng. Biotechnol. 3:90*

4.- El 78% de las proteínas moonlighting humanas presentes en la base de datos MultitaskProtDB-II están relacionadas con enfermedades, de acuerdo con la información existente en las bases de datos OMIM y HGMD. En algunos casos se ha descrito tan sólo una enfermedad, pero en un cierto número de ellas hay descrita una enfermedad diferente para cada función moonlighting. En algunos casos de los que presentan enfermedades relacionadas con funciones diferentes, el programa PiSite identifica las bases moleculares de ambas condiciones patológicas.

5.- El 48% de las proteínas moonlighting humanas presentes en la base de datos MultitaskProtDB-II son dianas de fármacos existentes en el mercado. Esta cifra adquiere mucha importancia dado que los fármacos actuales no están dirigidos contra un conjunto representativo del proteoma relacionado con enfermedades sinó desviado hacia unas dianas denominadas “druggables” (ya sea por ser dianas más accesibles o con menos efectos tóxicos al interferir con la función biológica).

Las conclusiones 4 y 5 incrementan la importancia del estudio de las proteínas moonlighting y su relación con las patologías humanas.

***Franco-Serrano L**., Huerta M., Hernández S., Cedano J., Pérez-Pons J. et al. Multifunctional Proteins: Involvement in Human Diseases and Targets of Current Drugs. Protein J. DOI: 10.1007/s10930-018-9790-x*

6.- Un 25% de las proteínas de la base de datos MultitaskProtDB-II corresponden a factores de virulencia de microorganismos patógenos. Se trata de proteínas, generalmente enzimas del metabolismo primario, que son secretadas y facilitan la colonización del huésped tras interaccionar con proteínas del mismo, tales como el Plasminógeno. En el presente trabajo se ha desarrollado la hipótesis de que al tratarse de proteínas secuencial y estructuralmente muy conservadas desde Procariotas a mamíferos y que por ello comparten epítopos, el sistema inmune del huésped evita desencadenar una respuesta inmune que podría dar lugar a enfermedades autoinmunes. A partir de un análisis mediante el programa MinimotifMiner, hemos encontrado algunos "motifs" relacionados con el sistema inmune que nos han permitido proponer una explicación de cómo tendría lugar el fenómeno de inhibición de la respuesta inmune del huésped.

***Franco-Serrano L**., Cedano J., Perez-Pons JA., Mozo-Villarias A., Piñol J. et al. (**2018**) A Hypothesis Explaining Why So many Pathogen Virulence Proteins are Moonlighting Proteins. Pathog Dis. DOI: 10.1093/femspd/fty046.*

7.- La existencia de la multifuncionalidad de las proteínas es un fenómeno más común de lo que se cree. Para estudiar este fenómeno se ha utilizado la información existente en las bases de datos UniProt y GO. Muchos investigadores, al introducir la(s) función(es) de las proteínas que analizan de acuerdo con los tres identificadores GO, han incorporado multifuncionalidades (*molecular function*) y localizaciones celulares (*cellular component*) múltiples, pero no han usado términos como multifuncional o moonlighting en sus publicaciones. A partir de la amplísima información sobre proteínas de la base de datos UniProt y de los identificadores GO, se ha buscado cuantas proteínas presentan múltiples descriptores que las hacen candidatas a ser multifuncionales. Esta cifra es elevadísima, para algunos pares de funciones como, por ejemplo, isomerase y DNA binding es de 166.833 proteínas. Esto nos hace pensar que hay un gran número de proteínas moonlighting aún por descubrir.

Work in progress.

8.- Sugerimos que la multifuncionalidad de las proteínas se conserva filogenéticamente. Aunque la respuesta a esta pregunta requiere corroboración

experimental, hemos desarrollado algunas estrategias bioinformáticas que corroborarían esta conservación evolutiva. Aparte de que la elevada conservación de secuencia entre proteínas moonlighting y sus ortólogas en otras especies lo sugiere se han desarrollado otras aproximaciones que también lo sugieren, como por ejemplo una basada en la ortología del interactoma. Aunque existe la limitación de que todavía no existen interactomas detallados para la mayoría de las especies, si encontramos que en aquellos que presentan abundantes ejemplos las proteínas ortólogas de diferentes especies comparten partners en su interactoma, esto apunta que la multifuncionalidad se conserva evolutivamente. Otra aproximación se basa en lo descrito en la anterior conclusión 7, donde decíamos que si proteínas ortólogas comparten identificadores GO múltiples seguramente también compartirian la multifuncionalidad, cosa que hemos visto que ocurre.

Work in progress.

9.- A pesar de los pocos ejemplos existentes, hemos encontrado que hay una relación evolutiva entre la multifuncionalidad y el fenómeno NOGD (*Non Orthologous Gene Displacement*).

10.- Como conclusión general se puede decir que el fenómeno de la multifuncionalidad de las proteínas presenta importantes retos evolutivos de relación estructura/función, de aplicación al análisis de las bases moleculares de las enfermedades humanas y al diseño de fármacos y vacunas. Todo ello las hace un fascinante objetivo de investigación.

REFERENCIAS

- Andreeva A., Howorth D., Chothia C., Kulesha E., Murzin A. (2014) SCOP2 prototype: a new approach to protein structure mining. Nucl. Acid Res. 42: D310-D314

- Alonso-López D, Gutiérrez M.A., Lopes K.P., Prieto C., Santamaria R., De Las Rivas J. (2016) APID interactomes: providing proteome-based interactomes with controlled quality for multiple species and derived networks. Nucleic Acids Research. 44: W529-535

- 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 of an enzyme. HSFP J. 1: 67-78

- 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

- H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne. (2000) The Protein Data Bank. Nucleic Acids Research. 28: 235-242

- 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

- Castaldo C., Vastano V., Siciliano RA., Candela M., Vici M., Muscariello L., Marasco R., Sacco M. (2009) Surface displaced alfa-Enolase of *Lactobacillus plantarum* is a fibronectin binding protein. Microbial Cell Factories. 8:14

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

- 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

- Chen C., Zabad S., Liu H., Wang W., Jeffery C. (2018) MoonProt 2.0: an expansion and update of the moonlighting proteins database. Nucleic Acid Res. 46: D640-D644.

- Cole J.N., Ramirez R.D., Currie B.J., Cordwell S.J., Djordjevic S.P., Walker M.J. (2005) Surface analyses and immune reactivities of major cell wall-associated proteins of Group A Streptococcus. Infection and Immunity 73: 3137–3146

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

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

- Espadaler J., Aragües R., Eswar N., Martí-Renom M., Querol E., Avilés F. X., Sali A., Oliva B. (2005) Detecting remotely related proteins by their Interactions and sequence similarity**.** Proc. Natl. Acad. Sci. USA. 102: 7151-7156

- Espadaler J., Eswar N., Querol E., Aviles F.X., Sali A., Martı-Renom M., Oliva B. (2008) Prediction of enzyme function by combining sequence similarity and protein interactions. BMC Bioinformatics. 9:249

- Feltcher M.E., Sullivan J.T., Braunstein M. (2010) Protein export systems of *Mycobacterium tuberculosis*: novel targets for drug development? 5: 1581-1597

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

- Fontan P.A., Pancholi V., Nociari M.M., Fischetti V.A. (2000) Antibodies to streptococcal surface Enolase react with human alpha-Enolase: implications in poststreptococcal sequelae. Journal of Infectious Diseases. 182: 1712-1721

- Franco-Serrano L., Hernández S., Calvo A., Severi M.A., Ferragut G., Pérez-Pons J.A., Piñol J., Pich O., Mozo-Villarias A., Amela I., Querol E., Cedano J. (2018) MultitaskProtDB-II: an update of a database of multitasking/moonlighting proteins. Nucleic Acids Research. 46: D1 D645-D648

- Franco-Serrano L., Cedano J., Perez-Pons J., Mozo Villarias A., Piñol J., Amela I., Querol E. (2018b) A hypothesis explaining why so many pathogen virulence proteins are moonlighting proteins. Patho. Dis. DOI: 10.1093/femspd/fty046

- Fulde M., Steinert M., Bergmann S. (2013) Interaction of streptococcal plasminogen binding proteins with the host fibrinolytic system. Frontiers Cell Infect. Microbiol. 3: 85. DOI: 10.3389/fcimb.2013.00085

- 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

- Guex N. and Peitsch MC. (1997) SWISS MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling. Electrophoresis. 18: 2714-2723

- 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

- He Y., Racz R., Sayers S., Lin Y., Todd T., Hur J., Li X., Patel M., Zhao B., Chung M., Ostrow J., Sylora A., Dungarani P., Ulysse G., Kochhar K., Vidri B., Strait K., Jourdian G.W., Xiang Z. (2014) Updates on the web-based VIOLIN vaccine database and analysis system. Nucleic Acids Research. 42: D1124-D1132

- 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. (2013a) Bacterial Moonlighting Proteins and Bacterial Virulence. 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. 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. (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

- Itzek A., Guillen C., Fulde M., Friedrichs C., Rodloff A. (2010) Contribution of Plasminogen Activation towards the Pathogenic Potential of Oral Streptococci. Plos One. 5: e13826

- 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. 73226-234

- 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. Moonlighting cell stress proteins in microbial infections. 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 G.E. (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: 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

- Kornblatt M., Kornblatt J., Hancock a. (2011) The Interaction of Canine Plasminogen with *Streptococcus pyogenes* Enolase: They Bind to One Another but What Is the Nature of the Structures Involved? Plos One. 6: e28481

- 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 R.F. (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 C.J., 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

- McGary K.L., Park T.J., Wood J.O., Cha H.J., Wallingford J.B., Marcotte E.M. (2010) Systematic discovery of nonobvious human disease models through orthologous phenotypes. Proc Natl Acad Sci USA. 107: 6544-9

- 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

- Robert D. Finn, et al. (2017) InterPro in 2017 - beyond protein family and domain annotations. Nucleic Acids Research. DOI: 10.1093/nar/gkw1107

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

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

- 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 J.A.., McCabe E.R., Liao J.C., Dipple K.M. (2005) Single-gene disorders: what role could moonlighting enzymes play? A. J. Human Genet. 76: 911-924

- The Gene Ontology Consortium (2017) Expansion of the Gene Ontology knowledgebase and resources. Nucleic Acids Res. 4;45: D331-D338

- The UniProt Consortium. (2017). UniProt: the universal protein knowledgebase. Nucleic Acids Res., 45: D158-D169

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

- Tress M.L., Abascal F., Valencia A. (2017) Alternative splicing may not be the key to proteome complexity. Trends Biochem. Sci. 42: 98-110

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

- Tsai C.J., Ma B., Sham Y.Y., 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 B.F. (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: 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., Feunang Y.D., Guo A.C., Lo E.J., Marcu A., Grant J.R., Sajed T., Johnson D., Li C., Sayeeda Z., Assempour N., Iynkkaran I.,4, Liu Y., Maciejewski A., Gale N., Wilson A., Chin L., Cummings R., Le D., Pon A., Knox C., Wilson M. (2018) DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 46: D1074-D1082

- 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. 4: 11 44