

# Proteínas moonlighting: Relación e implicaciones con el desplazamiento no ortólogo de gen

**Francisco Manuel Ribas García**

Máster de bioinformática y bioestadística  
Área del trabajo final número 2

**Nombre Consultor/a: Dr. Luis Franco Serrano**

**Nombre Profesor/a responsable de la asignatura: Marc Maceira Duch**

8 de junio del 2021



## A) Creative Commons:



Esta obra está sujeta a una licencia de Reconocimiento-NoComercial-SinObraDerivada [3.0 España de Creative Commons](https://creativecommons.org/licenses/by-nc-nd/3.0/es/)

## B) GNU Free Documentation License (GNU FDL)

Copyright © 2021 Francisco Manuel Ribas Garcia.

Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.3 or any later version published by the Free Software Foundation; with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts.

A copy of the license is included in the section entitled "GNU Free Documentation License".

## C) Copyright

© (Francisco Manuel Ribas Garcia)

Reservados todos los derechos. Está prohibido la reproducción total o parcial de esta obra por cualquier medio o procedimiento, comprendidos la impresión, la reprografía, el microfilme, el tratamiento informático o cualquier otro sistema, así como la distribución de ejemplares mediante alquiler y préstamo, sin la autorización escrita del autor o de los límites que autorice la Ley de Propiedad Intelectual.

## FICHA DEL TRABAJO FINAL

<b>Título del trabajo:</b>	<i>Proteínas moonlighting: Relación e implicaciones con el desplazamiento no ortólogo de gen</i>
<b>Nombre del autor:</b>	<i>Francisco Manuel Ribas Garcia</i>
<b>Nombre del consultor/a:</b>	<i>Luis Franco Serrano</i>
<b>Nombre del PRA:</b>	<i>Marc Maceira Duch</i>
<b>Fecha de entrega (mm/aaaa):</b>	06/2021
<b>Titulación:</b>	<i>Máster de bioinformática y bioestadística</i>
<b>Área del Trabajo Final:</b>	Área 2
<b>Idioma del trabajo:</b>	Castellano
<b>Número de créditos:</b>	15
<b>Palabras clave</b>	<i>Árbol filogenético, relación evolutiva y alineamiento múltiple de secuencias</i>
<b>Resumen del Trabajo (máximo 250 palabras): <i>Con la finalidad, contexto de aplicación, metodología, resultados i conclusiones del trabajo.</i></b>	
<p>Por un lado, las proteínas moonlighting presentan la característica de poder realizar diversas funciones bioquímicas con la misma cadena de aminoácidos. Estas funciones no tienen porque estar relacionadas y normalmente una de ellas pertenece al metabolismo primario. Por otro lado, el fenómeno NOGD es el que se produce cuando en una ruta metabólica, desaparece una enzima y su función es realizada por otra proteína totalmente diferente. Por tanto, el objetivo de este estudio ha sido el de intentar conseguir evidencias de una posible relación evolutiva entre las proteínas moonlighting, las cuales están muy poco estudiadas, con el fenómeno de NOGD. Para ello, se han realizado, con herramientas bioinformáticas, árboles filogenéticos de distintas enzimas extraídas de bases de datos de proteínas que encajasen en ambos grupos. En todos los casos se ha encontrado que solamente una proteína NOGD comparte linaje evolutivo con la proteína multifuncional, por tanto, se podría sugerir que el fenómeno NOGD se basa en parte en la multifuncionalidad de las proteínas de diversas formas como la duplicación génica o mediante mutaciones que van produciendo zonas favorables a realizar otra función. Al fin y al cabo, la multifuncionalidad presenta ventajas evolutivas como una mayor versatilidad en el DNA o de protección del individuo de posibles mutaciones que invaliden una enzima. Sea como fuere, el campo de las proteínas multifuncionales se encuentra en un desarrollo que en poco tiempo van a revolucionar la percepción bioquímica actual.</p>	
<b>Abstract (in English, 250 words or less):</b>	
<p>Moonlighting proteins have the characteristic of being able to perform several biochemical functions with the same amino acid chain. These functions do not necessarily have to be related and normally one of them belongs to metabolism. The NOGD phenomenon occurs when an enzyme disappears from a metabolic pathway and its function is carried out by a completely different protein. Therefore, the aim of this study was to try to obtain evidence of a possible relationship and evolution between moonlighting proteins, which are little studied, and the NOGD</p>	

phenomenon. To this end, phylogenetic trees of different enzymes extracted from protein databases that fit into both groups were created using bioinformatics tools. We have found that only one NOGD protein shares evolutionary lineage with the multifunctional protein. Therefore, it could be argued that the multifunctionality of the proteins has arisen from the NOGD phenomenon in different ways, such as gene duplication or through mutations that produce areas that are favourable for another function to be stored. After all, multifunctionality has evolutionary advantages such as greater versatility in DNA or protection of the individual from possible mutations that invalidate an enzyme. Be that as it may, the field of multifunctional proteins is developing in a way that will soon revolutionise the current biochemical perception of protein functionality.



# Índice

<b>LISTA DE FIGURAS .....</b>	<b>1</b>
<b>LISTA DE TABLAS .....</b>	<b>1</b>
<b>RESUMEN .....</b>	<b>1</b>
<b>INTRODUCCIÓN.....</b>	<b>1</b>
CONTEXTO Y JUSTIFICACIÓN DEL TRABAJO .....	1
OBJETIVOS DEL TRABAJO .....	2
ENFOQUE Y MÉTODO SEGUIDO.....	3
PLANIFICACIÓN DEL TRABAJO .....	3
BREVE SUMARIO DE CONTRIBUCIONES Y PRODUCTOS OBTENIDOS .....	5
BREVE DESCRIPCIÓN DE LOS OTROS CAPÍTULOS DE LA MEMORIA .....	5
<i>Estado de arte</i> .....	5
<i>Metodología</i> .....	5
<i>Resultados</i> .....	5
<i>Discusión</i> .....	5
<i>Conclusiones</i> .....	5
<i>Glosario</i> .....	6
<i>Bibliografía</i> .....	6
<b>ESTADO DEL ARTE .....</b>	<b>6</b>
<b>METODOLOGÍA .....</b>	<b>6</b>
BASES DE DATOS EMPLEADAS.....	6
<i>Multitasking / Moonlighting Proteins DataBase</i> .....	6
<i>Lista actualizada de enzimas análogas</i> .....	7
<i>UniProt</i> .....	7
<i>ClustalW y Clustal Omega</i> .....	7
<i>PSI-BLAST (Position-Specific Iterative Basic Local Alignment Search Tool)</i> .....	7
<i>NGPhylogeny</i> .....	8
<i>Interactive Tree Of Life (iTOL)</i> .....	8
ACTUALIZAR EL INVENTARIO DE PROTEÍNAS QUE SON MOONLIGHTING Y NOGD .....	8
ALINEAMIENTO DE SECUENCIAS ENTRE LAS PROTEÍNAS MOONLIGHTING Y NOGD .....	9
MSA Y ANÁLISIS FILOGENÉTICO ENTRE LAS PROTEÍNAS MOONLIGHT Y NOGD .....	9
<b>RESULTADOS.....</b>	<b>10</b>
ACTUALIZAR EL INVENTARIO DE PROTEÍNAS QUE SON MOONLIGHTING Y NOGD .....	10
ALINEAMIENTO MÚLTIPLE DE SECUENCIAS ENTRE LAS PROTEÍNAS MOONLIGHTING Y NOGD .....	11
HOMOLOGÍA REMOTA Y ARBOLES FILOGENÉTICOS.....	12
<i>Oxido nítrico sintasa</i> .....	12
<i>Tirosina fosfatasa</i> .....	13
<i>IMP ciclohidrolasa</i> .....	14
<i>Proteína-glutamina gamma-glutamyltransferasa 2</i> .....	14
<i>Lactato deshidrogenasa</i> .....	15
<i>Glucosa-6-fosfato isomerasas</i> .....	17
<i>Peroxiirredoxina</i> .....	18
<b>DISCUSIÓN.....</b>	<b>19</b>
<b>CONCLUSIONES .....</b>	<b>22</b>
CONCLUSIONES.....	22
LÍNEAS DE FUTURO .....	22
SEGUIMIENTO DE LA PLANIFICACIÓN .....	22
<b>GLOSARIO.....</b>	<b>23</b>
<b>BIBLIOGRAFÍA .....</b>	<b>23</b>





## Lista de figuras

**Figura 1:** Diagrama de Gantt sobre la planificación del trabajo.

**Figura 2:** Esquema sobre la metodología.

**Figura 3:** Árbol filogenético de la enzima oxido nítrico sintasa.

**Figura 4:** Árbol filogenético de la enzima tirosina fosfatasa.

**Figura 5:** Árbol filogenético de la enzima IMP ciclohidrolasa.

**Figura 6:** Árbol filogenético de la enzima proteína-glutamina gamma-glutamilttransferasa 2

**Figura 7:** Árbol filogenético de la enzima lactato deshidrogenasa de upsilon-cristalino en *Ornithorhynchus anatinus*.

**Figura 8:** Árbol filogenético de la enzima lactato deshidrogenasa de upsilon-cristalino en *Anas platyrhynchos*

**Figura 9:** Árbol filogenético de la enzima lactato deshidrogenasa cadena A

**Figura 10:** Árbol filogenético de la enzima glucosa-6-fosfato isomerasa en *Mus musculus*.

**Figura 11:** Árbol filogenético de la enzima glucosa-6-fosfato isomerasa en *Mustela putorius furo*.

**Figura 12:** Árbol filogenético de la enzima peroxirredoxina.

**Figura 13:** Hipotesis sobre la resolución del conflicto de las proteínas moonlighting.

## Lista de tablas

**Tabla 1:** Planificación detallada con nombres, fechas de inicio y de fin de las diferentes tareas e hitos del trabajo.

**Tabla 2:** Selección de proteínas para realizar homología remota.

## Resumen

### Introducción

#### Contexto y justificación del Trabajo

Históricamente se ha podido secuenciar genomas con herramientas cada vez mejores, donde gracias a los análisis incluso se ha llegado a estimar que las funciones de entre el 40-60% de las proteínas se podría llegar a predecir solamente basándose en la gran afinidad con secuencias similares que se tienen de otras proteínas. Pero uno de los diversos problemas existentes con la predicción de función de estas proteínas es el desplazamiento no ortólogo de gen (NOGD, Non-Orthologous gene Displacement).<sup>1</sup> El NOGD es el fenómeno en la que se produce una alteración en el organismo donde la función de una proteína en un lugar específico es reemplazada por otra con un origen evolutivo totalmente distinto y que en bacterias se realiza mediante la transferencia horizontal de genes (HGT, horizontal gene transfer).<sup>2,3</sup> Es aquí cuando encontramos a las proteínas moonlight o multifuncionales que son un subconjunto de proteínas, las cuales, presentan más de una función bioquímicamente relevantes y en donde estas funciones no están relacionadas entre sí. Lo curioso de la multifuncionalidad que presentan las proteínas viene debido a que la cadena peptídica es para ambas funciones la misma, sin que haya splicing alternativo, fusión con otras proteínas o cortes en la cadena para crear otra proteína nueva que derive de la anterior.<sup>4,5</sup> Estas proteínas moonlighting se encuentran en toda la filogenia desde bacterias hasta los mamíferos o plantas e incluso virus, en donde la combinación de funciones puede ser cualquiera pero la más habitual es la existente entre las funciones enzimáticas del metabolismo como canónica y funciones extracelulares como secundarias, como por ejemplo, la primera en ser identificada en estreptococos patógenos la gliceraldehído-3-fosfato deshidrogenasa (GAPDH, Glyceraldehyde 3-phosphate dehydrogenase), que al estudiarse en mamíferos se ha observado como función primaria en la glucólisis pero que la secundaria es receptor de transferrina para la absorción de hierro.<sup>4,6</sup>

El descubrimiento de las proteínas multifuncionales fue debido a la casualidad de la convergencia de dos investigaciones. Se cree que la existencia de estas proteínas se debe a patrones de expresión inusuales que hacen que la proteína sustituida en el NOGD no pueda realizar la función como es debido en el tejido o durante ciertos estados del desarrollo. Es por ello por lo que el mecanismo, se considera que es una forma inteligente de generar complejidad en los organismos sin que tener que expandir el genoma. Es por tanto que el estudio de las moonlighting es un campo prometedor en el ámbito de la farmacología y que en un futuro pueda llegar a usar este fenómeno en tratamientos, donde por ejemplo queramos inhibir o sobreexpresar una sustancia para que tenga un efecto terapéutico en el organismo.<sup>7</sup>

Gracias a técnicas como la cristalografía por rayos X se ha podido estudiar como las proteínas presentan esta multifuncionalidad. Y se ha podido observar que los sitios funcionales en esta proteína pueden en unos casos estar localizados en distintos lugares de la superficie de la proteína, pudiendo así realizar las 2 funciones simultáneamente dependiendo así de la localización celular donde se realice; mientras que para otras proteínas los lugares donde se realizan las funciones se solapan y es ahí donde la proteína cambia su conformación para dar lugar al lugar funcional secundario; y finalmente hay proteínas que realizan una función como monómero, pero al ser incorporadas en un complejo multiproteico realizan una segunda función.<sup>8</sup>

Desde aquella primera proteína multifuncional descubierta se han llegado a realizar muchos avances. El grupo del Dr. Franco ha creado una base de datos de proteínas multifuncionales que hasta la fecha presenta 694 proteínas. De esta base de datos se ha llegado a estimar que el 78% de las proteínas moonlight humanas están relacionadas con alguna enfermedad, donde el 48% de estas proteínas multifuncionales humanas son dianas de fármacos actuales. Del mismo modo, se ha llegado a concluir que el 25% de la base de datos corresponde a proteínas que están relacionadas con mecanismos de virulencia de microorganismos patógenos.<sup>9</sup>

Es, por tanto, que el estudio de la multifuncionalidad de las proteínas junto con el NOGD es tan importante, ya que desgraciadamente actualmente hay muy poca investigación e información sobre el tema y por tanto quedan demasiados asuntos por dilucidar como puede ser el origen evolutivo o los mecanismos por los cuales las proteínas han conseguido estas funciones secundarias o porque una y no otra proteína es la que sustituye la función. Cuanta más información se llegue a obtener sobre la multifuncionalidad, más ayudará a poder identificar estos sucesos en más proteínas, llegar a introducir estas variables en bases de datos y poder por ejemplo diseñar fármacos con mejores dianas terapéuticas y que así sean necesarias dosis mucho más bajas de estos.

### Objetivos del Trabajo

El objetivo principal es el de profundizar en la relación del moonlighting con el fenómeno NOGD debido a que actualmente es un campo científico del cual no hay mucha información.

Los objetivos específicos son:

- Actualizar y ampliar la base de datos de proteínas moonlighting y NOGD del grupo de Biología Molecular del Instituto de Biotecnología y Biomedicina.
- Comparar las bases de datos de proteínas moonlighting y de proteínas NOGD para obtener proteínas coincidentes en ambas bases de datos.
- Observar el posible impacto que tiene o puede llegar a tener tanto en las aplicaciones genómicas, bioingeniería o el desarrollo de fármacos para tratar enfermedades.

- Realizar un análisis evolutivo de las proteínas moonlighting que participan en fenómenos NOGD para determinar su origen en casos concretos.

### Enfoque y método seguido

A partir de dos bases de datos de proteínas diferentes que son las de proteínas moonlighting y las de NOGD, las proteínas que se encontraban en ambas bases de datos se les realiza un estudio de homología para poder llegar a determinar su origen y si presentan relaciones filogenéticas entre ambas. Luego si de una proteína NOGD que presenta diversas estructuras, existe una de ellas que coincide con la estructura de la proteína moonlighting, se demuestra de esta forma que hay diversas líneas evolutivas y que la proteína multifuncional habría surgido de un mecanismo NOGD, porque hay que recordar que esta función es la secundaria y no la canónica.

El primer paso implica una actualización de un inventario de proteínas que son tanto NOGD como moonlighting, para realizarlo se comparan tanto la base de datos de ambos tipos de proteínas y se buscan las que estén en ambas listas.

Posteriormente, se escogen unas cuantas proteínas que se pueda pensar que tienen una relación y se les realizan análisis mediante herramientas de alineamiento de secuencias, para observar de que variante evolutiva pueden proceder y realizar con esta información arboles filogenéticos.

Finalmente, cuando se tengan los arboles filogenéticos hechos, analizar si las estructuras de las proteínas proceden de distintas líneas evolutivas y si alguna de estas procede de una que anteriormente no realizaba esta función para suplir a otra y además mantiene la función original, pasa a ser una proteína moonlighting.

### Planificación del Trabajo

Las primeras semanas del trabajo se han dedicado a la búsqueda de información para la definición del trabajo, la búsqueda bibliográfica y la planificación de las tareas a lo largo de las semanas.

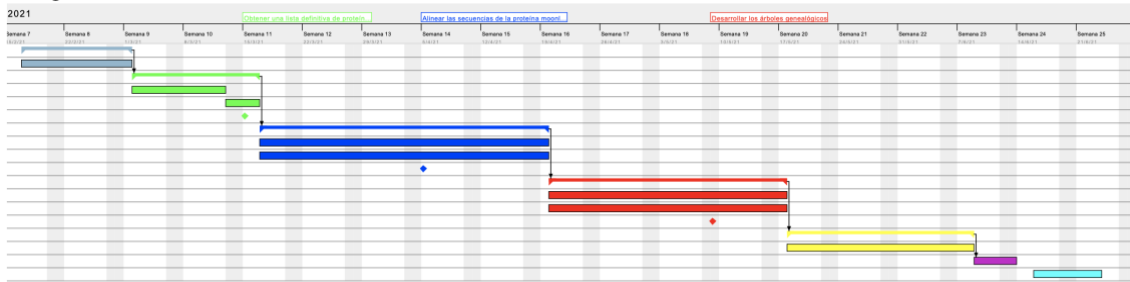
Durante las siguientes semanas, las cuales involucran las PECs 1, 2 y 3 se han ido realizando las diversas tareas de bioinformática, repartidas por los objetivos específicos del trabajo los cuales son:

- Actualizar la base de datos de las proteínas.
- Realizar los análisis de secuencias necesarios.
- Realizar los análisis filogenéticos y análisis de resultados.

Durante los tiempos que integran entre la PEC 1 y la PEC 4 se ha ido desarrollando la memoria del trabajo.

**Tabla 1:** Encontramos la planificación del trabajo durante periodos determinados con las diversas tareas e hitos que se han de ir realizando en cada PEC. Los diferentes hitos que se desarrollan en este proyecto son: “Obtener una lista definitiva de proteínas que son moonlight y NOGD”, “Alinear las secuencias de las proteínas moonlighting frente a sus homólogas” y “Desarrollar los árboles genealógicos”.

<b>Nombre</b>	<b>Fecha de inicio</b>	<b>Fecha de fin</b>
PEC 0 - Definición de los objetivos del trabajo	17/2/21	1/3/21
Busqueda de bibliografía y objetivos	17/2/21	1/3/21
PEC 1 - Plan de trabajo	2/3/21	16/3/21
Iniciar la memoria	2/3/21	12/3/21
Actualizar el listado de proteínas moonlight y NOGD	13/3/21	16/3/21
Obtener una lista definitiva de proteínas que son moonlight y NOGD	15/3/21	15/3/21
PEC 2 - Desarrollo del trabajo - Fase 1	17/3/21	19/4/21
Análisis de proteínas	17/3/21	19/4/21
Continuar redactando la memoria	17/3/21	19/4/21
Alinear las secuencias de la proteína moonlighting frente a sus homólogas NOGD	5/4/21	5/4/21
PEC 3- Desarrollo del trabajo - Fase 2	20/4/21	17/5/21
Analizar los resultados	20/4/21	17/5/21
Continuar redactando la memoria	20/4/21	17/5/21
Desarrollar los árboles genealógicos	9/5/21	9/5/21
PEC 4 - Cierre de la memoria	18/5/21	8/6/21
Acabar y perfeccionar la memoria	18/5/21	8/6/21
PEC 5a - Elaboración de la presentación	9/6/21	13/6/21
PEC 5b -Defensa pública	16/6/21	23/6/21



**Figura 1:** Encontramos representado un diagrama de Gantt, en el que cada línea de color fina indica el título de las diversas PECs ordenadas en orden cronológico. Dentro de cada PEC encontramos las diferentes tareas e hitos (que aparecen tanto como un rombo como encima del diagrama el nombre que tienen) que hemos descrito anteriormente en la **tabla 1**.

### Breve resumen de contribuciones y productos obtenidos

De este trabajo de final de máster se obtienen en primer lugar y cronológicamente, un listado de proteínas que son moonlighting y NOGD al mismo tiempo. En segundo lugar, un árbol filogenético gracias a los alineamientos de secuencias, con los que descubrir la evolución de las proteínas.

### Breve descripción de los otros capítulos de la memoria

#### Estado de arte

El trabajo se va realizando dentro de los tiempos establecidos y que aparecen en el cronograma anteriormente detallado tanto en la tabla 1 como en la figura 1.

#### Metodología

Se enumeran y explican las diferentes bases de datos, las diferentes herramientas bioinformáticas y aplicaciones que se han utilizado de forma cronológica para poder llegar a realizar el proyecto punto por punto.

#### Resultados

Se exponen tanto el listado actualizado de proteínas que cumplen la condición de ser al mismo tiempo moonlighting y NOGD, como el árbol filogenético obtenido a partir de ellas.

#### Discusión

Se analizan los resultados obtenidos, estudiando la relación evolutiva de las proteínas que existen entre ambos fenómenos, el de multifuncionalidad y el NOGD.

#### Conclusiones

Nos encontramos con una enumeración de las diferentes ideas finales del proyecto, junto con alguna reflexión final y posibles salidas para seguir investigando sobre el tema en cuestión.

## Glosario

En el encontramos las diversas palabras, junto con sus abreviaturas y lo que significan que vamos encontrando en el trabajo.

## Bibliografía

En este apartado nos encontramos con los diversos textos científicos que se han ido utilizando para poder introducir y llegar a entender el tema del que se fundamenta el trabajo del final de máster.

## Estado del arte

La investigación sobre las proteínas moonlighting es valiosa debido a que se han ido encontrando cientos de proteínas multifuncionales en los tres dominios biológicos y el número de las que se han ido encontrando va creciendo exponencialmente. La reevaluación de cuantos genes codificantes de proteínas presenta el genoma humano, el cual ronda los 19000, se cree que puede estar relacionado con este tipo de proteínas, ya que con pocas proteínas codificadas por el genoma pueden realizar gran cantidad de funciones por cada una de ellas, ampliando así el rango del proteoma funcional humano. El lado negativo de estas proteínas es que pueden estar involucradas en diversas enfermedades humanas como el cáncer o la fibrosis quística y que la regulación de las diferentes funciones puede traer conflictos evolutivos.<sup>10</sup>

Por tanto, el trabajo es de importancia ya que empieza a labrar el camino de progreso con estas proteínas, llevando así a avanzar en los campos de la biología y la medicina dilucidando preguntas como cual puede ser la biología estructural de estas proteínas o como se pueden utilizar para que sean objetivos más concretos de medicamentos.

## Metodología

### Bases de datos empleadas

#### Multitasking / Moonlighting Proteins DataBase

Corresponde a la base de datos creada por el grupo del Dr. Franco, la cual se ha actualizado por última vez en el año 2018. Esta base de datos contiene la información de 694 proteínas, donde se puede encontrar información de gran interés de estas a simple vista. En la tabla se puede detallar los códigos UniProt,<sup>11</sup> el nombre de la proteína, la función canónica, la función moonlight, los códigos GO de ambas funciones y referencias entre otras características.<sup>12</sup> Todo

ello, con sus distintos links a las diferentes páginas webs para realizar un análisis de forma más rápida.<sup>9</sup>

La página web de la base de datos es: [Multitasking Proteins DataBase](#)<sup>13</sup>

Lista actualizada de enzimas análogas

Corresponde a la actualización del 2010 sobre la base de datos creada por Koonin en 1998. Esta base de datos contiene la información sobre diversas proteínas NOGD, donde se puede encontrar el nombre de la enzima, los códigos tanto de EC como NCBI, UniProt o PDB,<sup>11,14,15</sup> las referencias bibliográficas o las familias y superfamilias a las que pertenecen estas proteínas entre otras características.<sup>2</sup>

La página web de la base de datos es: [An updated list of analogous enzymes](#)<sup>16</sup>

UniProt

Base de datos la cual proporciona una información completa sobre cualquier dato relacionado con las secuencias de las proteínas conocidas, debido a que recopila, estandariza y organiza el conocimiento que se tiene sobre las proteínas y sus funciones.<sup>17</sup> Esta base de datos ha sido utilizada para obtener la secuencia de aminoácidos de las proteínas de las cuales se han realizado los alineamientos.

La página web de la base de datos es: [UniProt](#)<sup>11</sup>

ClustalW y Clustal Omega

Ambas son herramientas bioinformáticas que realizan alineamientos entre secuencias de proteínas. Hemos empleado ClustalW para obtener la comparación bioinformática entre secuencias fasta en formato de texto. Mientras que la herramienta de Clustal Omega la hemos empleado para obtener de los alineamientos la matriz de con el porcentaje de similaridad entre ambas secuencias.

Las páginas web de las bases de datos son: [ClustalW](#) y [Clustal Omega](#)<sup>18,19</sup>

PSI-BLAST (Position-Specific Iterative Basic Local Alignment Search Tool)

El PSI-BLAST deriva de la técnica de BLAST la cual es un método para encontrar similitudes entre secuencias, tanto estas secuencias de nucleótidos como de proteínas se comparan con otras para llegar a identificar regiones en alineamientos locales similares y según como sea este alineamiento, darles una puntuación. Por tanto, el PSI-BLAST utiliza una matriz de alineaciones múltiples de secuencias (*MSA, multiple sequence alignment*) a partir de la secuencia introducida. Este MSA da una matriz de secuencias coincidentes con la introducida y a partir de ahí se pueden llegar a utilizar para intentar encontrar relaciones con secuencias posteriormente que son más distantes.<sup>20</sup>



Para realizar el PSI-BLAST se dejaron todos los ajustes por defecto con el algoritmo del PSI-BLAST y la página web que se utilizó es: [PSI-BLAST](#)<sup>21</sup>

#### NGPhylogeny

Es una herramienta diseñada para ejecutar flujos de trabajo de árboles filogenéticos mediante la utilización de flujos de trabajo preestablecidos que se basan en los siguientes pasos: Primero, buscar secuencias basadas en BLAST, luego realizar un MSA, posteriormente, correcciones del alineamiento, cuarto, deducir la filogenia de los árboles y finalmente visualizar el árbol.

Esta herramienta presenta diversas ventajas como la de tener un alcance general adaptándose a las diferentes necesidades; ser flexible, debido a que se puede actualizar la herramienta; y es escalable, ya que es capaz de soportar gran cantidad de cálculos y análisis en paralelo mediante clusters.<sup>22</sup>

Para realizar los análisis, se ha empleado el modo “Un click” en la opción de “PhyML+SMS” debido a que es la que se emplea para conjuntos de datos entre pequeños y medianos.<sup>23</sup> La que la opción de “FastTree” se emplea para cuando se tienen más de 5000 secuencias.<sup>24</sup> Y las opciones de “FastMe” y “PhyML” para conjunto de datos entre las anteriores.<sup>25,26</sup>

La página web de la herramienta es: [NGPhylogeny](#)<sup>27</sup>

#### Interactive Tree Of Life (iTOL)

ITOL es una herramienta online a la cual somos redirigidos una vez NGPhylogeny ha terminado de analizar los datos que le hemos introducido. ITOL permite visualizar y manipular y exportar los árboles filogenéticos resultantes en diferentes formatos.<sup>28</sup>

La página web de la herramienta es: [Interactive Tree of Life](#)<sup>29</sup>

#### Actualizar el inventario de proteínas que son moonlighting y NOGD

Lo primero que se realizó fue acceder a las bases de datos de proteínas multifuncionales del grupo del Dr. Franco y de enzimas análogas actualizadas para poder llegar a estudiar que proteínas se encontraban en ambas tablas.

Ambas tablas comparten datos característicos para poder buscar similitudes entre proteínas. Entre estos datos comunes encontramos los códigos UniProt, los códigos PDB, los códigos EC y nombres de proteínas. Con todo esto se realizó una búsqueda manual de proteínas que estuvieran en ambas bases de datos, para esto se utilizó el código EC debido a que este define cual es su función sin tener en cuenta cual es su estructura u organismo, mientras que las otras características que se encuentran en las dos tablas como los otros códigos o el nombre son únicos dependiendo de la estructura y por tanto de esta forma no se encontrarían proteínas que presentan la misma función aún con diferente

estructura que pudiesen llegar a ser NOGD. El primer problema que se presentó al realizar la búsqueda es que hay proteínas en la base de datos de las proteínas multifuncionales que no presentan el código EC y por tanto en esos momentos se utilizó el nombre de las proteínas para intentar encontrar el mayor número de posibles coincidencias posible.

Gracias a esto se obtuvo un documento actualizado con información de las proteínas que presentaban función moonlight en un organismo y NOGD en otros.

### Alineamiento de secuencias entre las proteínas moonlighting y NOGD

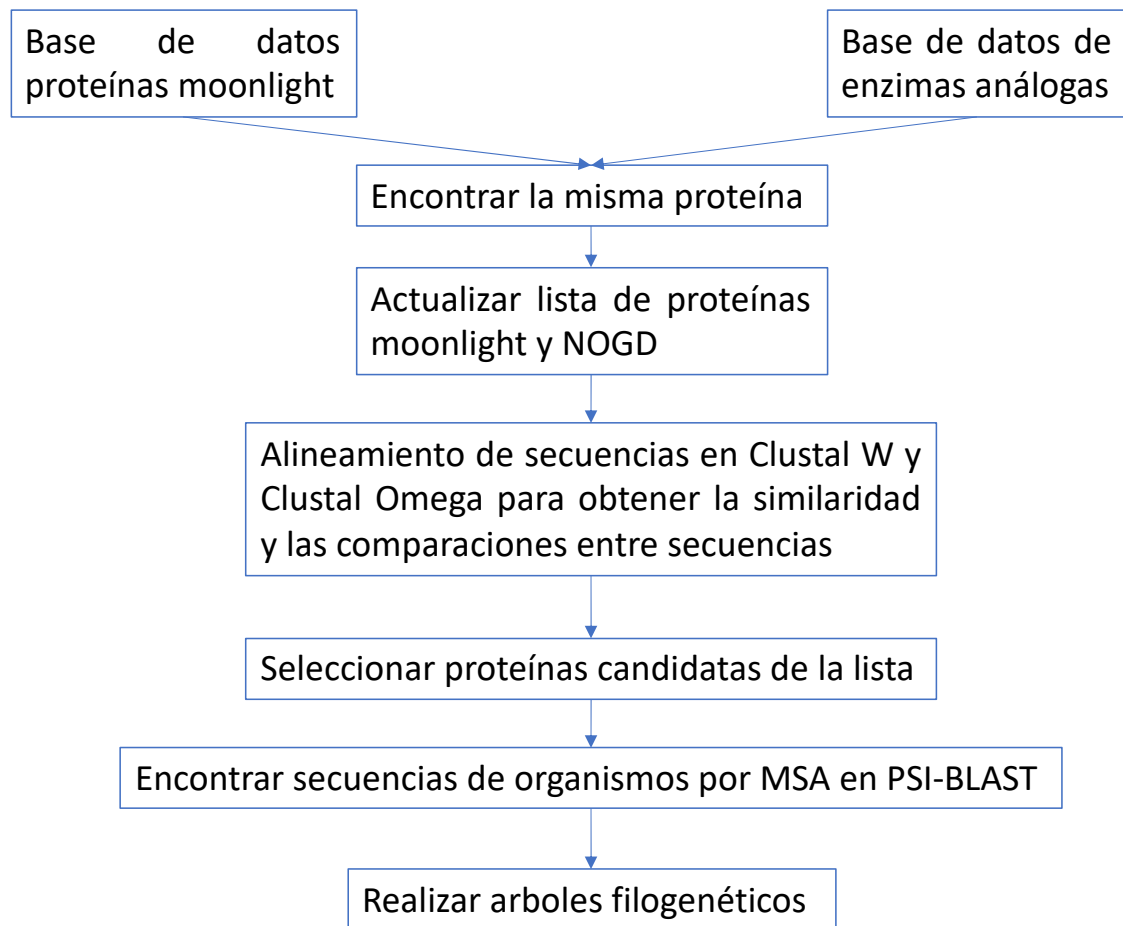
A partir de aquí se utilizó los códigos UniProt obtenidos entre ambas bases de datos. De estas entradas en UniProt se obtuvo de cada proteína la secuencia fasta que fue utilizada para realizar alineamientos múltiples tanto en ClustalW como en Clustal Omega.

De la herramienta Clustal Omega, mediante los alineamientos se obtuvieron las matrices con los porcentajes de homología entre las proteínas. Mientras que de la herramienta ClustalW se obtuvo de cada alineamiento un documento de texto simple con el alineamiento completo entre las secuencias de aminoácidos, donde en cada posición del alineamiento se obtiene información de conservación específica gracias a los símbolos “\*”, “:”, “.” y “”, los cuales significan que los residuos en esa misma posición son idénticos, residuos que conservan propiedades altamente similares, residuos que presentan baja similitud entre ellos y residuos que no presentan similitud entre ellos respectivamente.<sup>30</sup>

### MSA y análisis filogenético entre las proteínas moonlight y NOGD

La realización del análisis de homología remota entre proteínas, para encontrar la línea evolutiva de cada proteína y así observar de que rama podría proceder la proteína multifuncional se realizó mediante PSI-BLAST la cual, es una de las herramientas más utilizadas para este cometido.<sup>31</sup> Para empezar, se introdujeron las diferentes secuencias a analizar (tanto las moonlight como las NOGD) en el buscador con los ajustes por defecto y se fueron seleccionando de cada búsqueda, las proteínas que tuviesen una puntuación alta, con un E-valor lo más bajo posible y con índices de similitud entre el 40% y el 100%.

Una vez obtenidas las secuencias de organismos que cumplían estos parámetros, se introdujeron todas las secuencias a analizar en NGPhylogeny, el cual unos minutos más tarde nos daba información sobre los árboles filogenéticos y un botón para poder entrar en iTOL y poder obtener una imagen de estos.



**Figura 2:** Proceso seguido para llegar a realizar los árboles filogenéticos de las proteínas seleccionadas.

## Resultados

### Actualizar el inventario de proteínas que son moonlighting y NOGD

Gracias a la búsqueda realizada mediante toda la información disponible que se disponía se llegaron a encontrar hasta 52 funciones diferentes de proteínas NOGD que eran realizadas por hasta 98 proteínas multifuncionales, sin tener en cuenta la cantidad de organismos. Al mirar más en detalle en se puede llegar a observar que el número de proteínas diferentes que realizan la misma función en diferentes organismos es muy variado, siendo lo más simple y sencillo con un número mínimo de 2 proteínas distintas hasta incluso 5 proteínas diferentes en distintos organismos y funciones.

En el total de todas las combinaciones se han podido encontrar particularidades tanto que una misma función de una proteína NOGD era realizada por diferentes proteínas multifuncionales, como es la función de N-acetylmuramoyl-L-alanina amidasa que es producida por N-acetylmuramoyl-L-alanina amidasa SLE1, la amidasa, las proteínas de superficie PSPA y PSCP y la LytA. Como, al contrario, donde una proteína moonlighting puede realizar diversas funciones de proteínas NOGD como la peroxirredoxina 6 que realiza las funciones NOGD de peroxirredoxina, fosfolipasada 2 y lisofosfolipasa.

## Alineamiento múltiple de secuencias entre las proteínas moonlighting y NOGD

Una vez obtenido refrescado todo el inventario de proteínas que al mismo tiempo son multifuncionales y NOGD se pasó a realizar los alineamientos múltiples de secuencias, donde se llegaron a realizar hasta 321 alineamientos tanto en ClustalW como en Clustal Omega, para llegar a obtener tanto los porcentajes de similitud entre secuencias como los formatos fasta de los alineamientos.

De todos los análisis entre las diversas proteínas se llegaron a seleccionar las que cumplían los siguientes requisitos:

- En el porcentaje de similaridad entre la proteína multifuncional y las NOGD, existe una con un porcentaje superior al 40%, mientras que las otras presentaban unos porcentajes de similitud muy bajos.
- Los organismos en los cuales se produce el fenómeno NOGD son muy diversos entre ellos.
- Las proteínas que presentan el fenómeno NOGD son de organismos muy cercanos evolutivamente hablando o incluso llegan a ser del mismo organismo que el de la proteína moonlight.

Teniendo en cuenta las diferentes características se llegaron a seleccionar 10 proteínas multifuncionales.

**Tabla 2:** *En esta tabla encontramos la selección de proteínas analizadas para realizar la homología reversa y los arboles filogenéticos por las características comentadas anteriormente.*

NOGD protein	Moonlighting Protein	Moonlighting in	NOGD in	NOGD UniProt	Similarity (%)
Nitric-oxide synthase	Endothelial Nitric Oxide Synthase	<i>Homo sapiens</i>	<i>Rattus norvegicus</i>	NOS1_RAT	57,56
			<i>Arabidopsis thaliana</i>	NOS1_ARATH	21,08
Protein-tyrosine-phosphatase	Developmental protein eyes absent	<i>Drosophila melanogaster</i>	<i>Streptomyces coelicolor</i>	PTPA_STRCO	21,79
			<i>Homo sapiens</i>	PTPRD_HUMAN	20,06
			<i>Homo sapiens</i>	MPIP3_HUMAN	19,89
			<i>Bacillus subtilis</i>	YWQE_BACSU	19,52
			<i>Mus musculus</i>	EYA3_MOUSE	43,71
			<i>Homo sapiens</i>	PUR9_HUMAN	38,26
IMP cyclohydrolase	Phosphoribosylaminoimidazolecarboxamide formyltransferase-IMP cyclohydrolase (AICARFT-IMPCHase)	<i>Arabidopsis thaliana</i>	<i>Methanotheroxothrix thermoacetophila</i>	PURO_METTP	24,74
Protein-glutamine gamma-glutamyltransferase	Protein-glutamine gamma-glutamyltransferase 2	<i>Homo sapiens</i>	<i>Homo sapiens</i>	F13A_HUMAN	37,39
			<i>Streptomyces cmobaraensis</i>	TGAS_STRMB	18,96
			<i>Bacillus subtilis</i>	TGL_BACSU	19,37
			<i>Homo sapiens</i>	LDHA_HUMAN	92,47
L-lactate dehydrogenase y D-lactate dehydrogenase	Upsilon-crystallin	<i>Ornithorhynchus anatinus</i>	<i>Cupriavidus necator</i>	LDH_RALEH	19,06
			<i>Escherichia coli</i>	LDHD_ECOLI	17,02
			<i>Escherichia coli</i>	DLD_ECOLI	24,1
			<i>Homo sapiens</i>	LDHA_HUMAN	72,59
		<i>Anas platyrhynchos</i>	<i>Cupriavidus necator</i>	LDH_RALEH	20,23
			<i>Escherichia coli</i>	LDHD_ECOLI	19,59
			<i>Escherichia coli</i>	DLD_ECOLI	25,53
	L-lactate dehydrogenase A chain	<i>Rattus norvegicus</i>	<i>Homo sapiens</i>	LDHA_HUMAN	94,28
			<i>Cupriavidus necator</i>	LDH_RALEH	18,35
			<i>Escherichia coli</i>	LDHD_ECOLI	17,55
			<i>Escherichia coli</i>	DLD_ECOLI	24,5
			<i>Homo sapiens</i>	G6PI_HUMAN	88,89
			<i>Thermococcus litoralis</i>	G6PI_THELI	23,12
			<i>Homo sapiens</i>	G6PI_HUMAN	88,98
Glucose-6-phosphate isomerase	Glucose-6-phosphate isomerase	<i>Mus musculus</i>	<i>Thermococcus litoralis</i>	G6PI_THELI	21,82
	Bifunctional phosphoglucose/phosphomannose isomerase	<i>Mustela putorius furo</i>	<i>Aeropyrum pemix</i>	TDXH_AERPE	35,81
Peroxiredoxin	<i>Peroxiredoxin-6 EC:1.11.1.15</i>	<i>Homo sapiens</i>	<i>Mycobacterium tuberculosis</i>	AHPD_MYCTU	20,12
			<i>Escherichia coli</i>	OSMC_ECOLI	21,19
			<i>Mus musculus</i>	PRDX6_MOUSE	89,73

## Homología remota y árboles filogenéticos

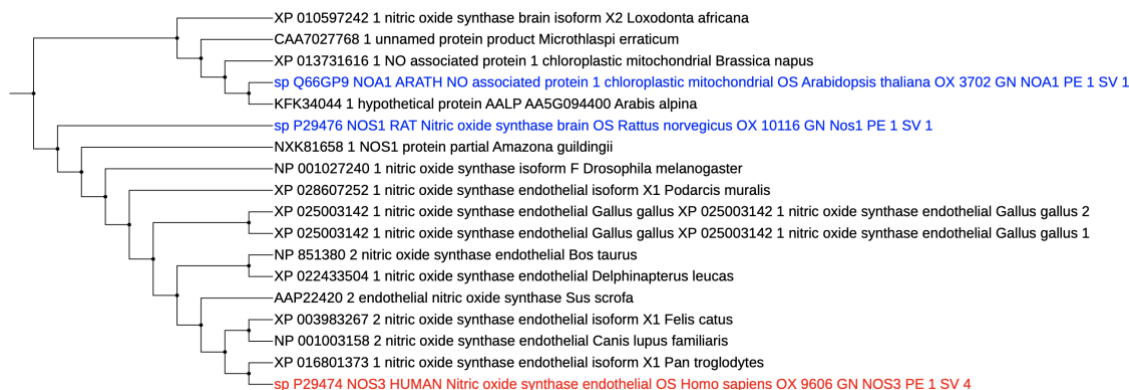
Lo siguiente que se ha realizado es el estudio de la homología remota mediante la página web de PSI-BLAST, en la que íbamos introduciendo las distintas proteínas a analizar. De estas proteínas poco a poco se iban seleccionando secuencias que estuviesen más alejadas tanto filogenicamente hablando como en la similitud que se presentaban para ir haciendo nuevas búsquedas e ir introduciendo nuevas secuencias cada vez más alejadas y poder hacer así líneas evolutivas.

### Oxido nítrico sintasa

Enzima que produce óxido nítrico, el cual interviene en la angiogénesis inducida por el factor de crecimiento endotelial vascular (VEGF, *vascular endothelial growth factor*) en los vasos coronarios en su forma multifuncional en *Homo sapiens*.<sup>32</sup> Esta enzima presenta diversas isoformas según en el organismo en el que se encuentre, ya que se presenta con el mecanismo NOGD en *Rattus norvegicus* con una isoforma cerebral, donde la producción de óxido nítrico se utiliza como neurotransmisor y en *Arabidopsis thaliana* con isoforma mitocondrial/cloroplastica donde la proteína presenta actividad en la respuesta inmune de la planta.<sup>33,34</sup>

Gracias a las secuencias obtenidas mediante el PSI-BLAST se llegó a realizar un árbol filogenético, el cual se puede observar en la *figura 3*, donde se puede analizar que existe una relación evolutiva entre las isoformas endotelial y

cerebral, mientras que la citoplasmática/mitocondrial constituye una línea evolutiva diferente, es decir, que proviene de una secuencia ancestrales diferente. Esto en un principio puede parecer contradictorio, debido a que los 3 organismos de los cuales miramos la enzima son del dominio Eukaryota, pero lo que hace que provengan de líneas evolutivas diferentes es que en *Arabidopsis thaliana*, el enzima es subcelular a nivel de cloroplastos y mitocondrias, que es sabido gracias a la teoría endosimbiontica que son del dominio Bacteria.



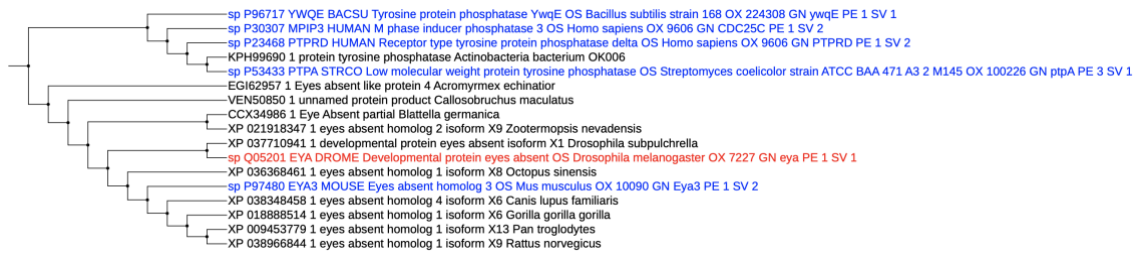
**Figura 3:** Árbol filogenético de la enzima óxido nítrico sintasa, donde se puede observar en rojo la enzima objetivo de homo sapiens y en azul las diversas secuencias analizadas. El resultado existente de 2 líneas evolutivas distintas donde las isoformas cerebrales y endoteliales van por una rama y la isoforma cloroplastica/mitocondrial por otra.

Debido a los resultados obtenidos entre la similaridad de secuencias y el árbol filogenético se deduce que la proteína multifuncional es la que se encuentra en *Rattus norvegicus*.

### Tirosina fosfatasa

Se cree que la tirosina fosfatasa en la forma moonlighting en *Drosophila melanogaster* tiene un papel regulador de la transcripción durante la organogénesis en el desarrollo ocular.<sup>35</sup> Esta función también la encontramos en el organismo *Mus musculus* y es por tanto que de todas las secuencias que analizamos la que se encuentra en la misma rama del árbol filogenético y presenta mayor similaridad como podemos encontrar en la *figura 4*, es por esto que se considera la secuencia multifuncional.<sup>36</sup>

El resto de las secuencias que analizamos con la homología remota se pueden observar que, al tener funciones diversas, como la desfosforilación de proteínas en *Bacillus subtilis* o como la del control de la mitosis o envuelta en la diferenciación presináptica en *homo sapiens*, se encuentran en otra rama filogénica distinta, además de presentar unos porcentajes de similaridad de secuencia muy inferiores a la de *Mus musculus*.<sup>37-39</sup>

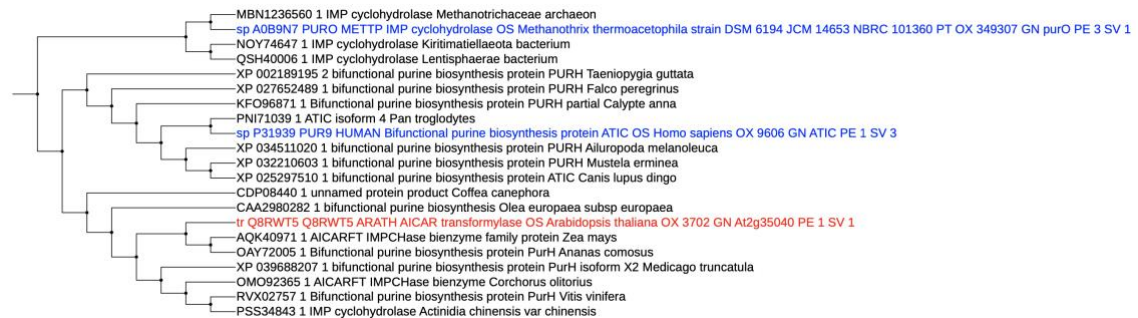


**Figura 4:** Árbol filogenético de la enzima tirosina fosfatasa, donde en rojo se encuentra la secuencia objetivo de *Drosophila melanogaster*, mientras que en azul las diversas secuencias analizadas. Se puede ver a simple vista que existen 2 ramas en el árbol que indican sus linajes evolutivos distintos.

### IMP ciclohidrolasa

Enzima que esta involucrada en la biosíntesis de novo del IMP en la proteína moonlighting en *Arabidopsis thaliana*.<sup>40</sup> Esta misma función se realiza en los organismos NOGD *homo sapiens* y *Methanotrix thermoacetophila*.<sup>41,42</sup>

En este árbol filogenético que podemos observar en la figura 5 encontramos especies muy diferentes debido al resultado obtenido por la homología remota mediante el PSI-BLAST, donde la proteína a analizar más cercana es la correspondiente a *homo sapiens*. Aunque se observan en ramas diferentes, al ser la más cercana que tenemos actualmente a la proteína objetivo, junto al porcentaje de similaridad y el dominio es el Eukarya, se deduce que provienen de un mismo ancestro común y es la proteína moonlighting. Mientras que la del microorganismo presenta un porcentaje de similaridad bajo, el organismo es del dominio Archaea y se encuentra en otra rama con otros microorganismos.

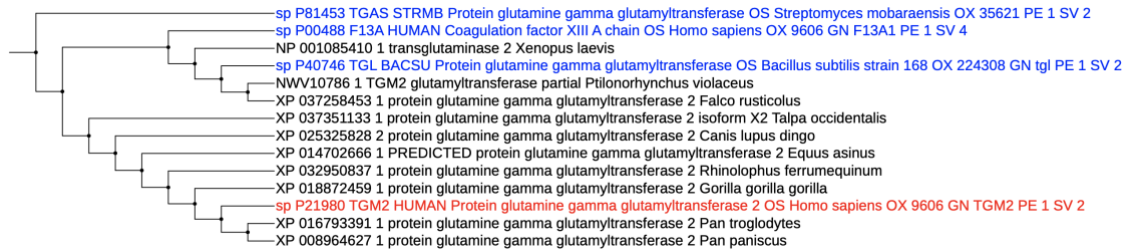


**Figura 5:** Árbol filogenético de la enzima IMP ciclohidrolasa que se encuentra en la vía de síntesis de novo de la IMP. Se observa en rojo la secuencia objetivo, mientras que en azul encontramos las que se han analizado mediante homología remota.

### Proteína-glutamina gamma-glutamyltransferasa 2

Enzima dependiente de calcio, que cataliza la formación de enlaces covalentes entre la glutamina y péptidos, que presentan grupos de amino primarios. Esta función es realizada tanto por la proteína objetivo en *Homo sapiens* como en las proteínas de otros organismos con las que se comparan.<sup>43-46</sup>

En el árbol filogenético que de la *figura 6* se puede observar que la proteína objetivo en el ser humano, aunque esta muy alejada de las proteínas analizadas, la más cercana que tiene es la del factor de coagulación en el mismo organismo, por tanto, es muy probable que ambas proteínas desciendan de una común y que cada una posteriormente tuviese cambios que las hiciesen diferentes pero que aún así estuviesen ligadas funcionalmente por el mecanismo NOGD, además de que por similitud es la que tiene un porcentaje mayor.



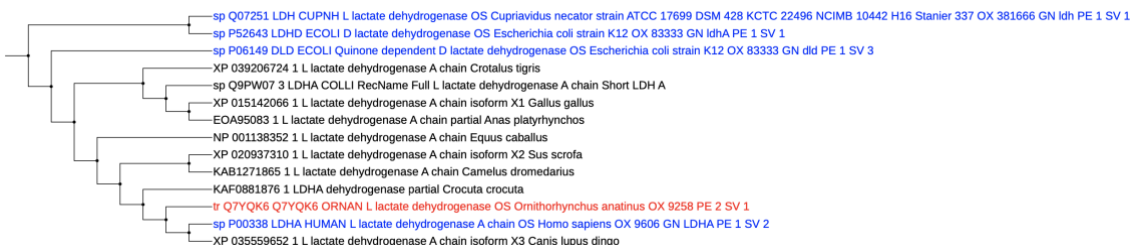
**Figura 6:** Árbol filogenético de la enzima Proteína-glutamina gamma-glutamiltransferasa 2, donde encontramos que la proteína objetivo se encuentra en rojo y bastante distanciada de las diversas proteínas analizadas que se resaltan en azul.

Lactato deshidrogenasa

*Upsilon*-cristalino

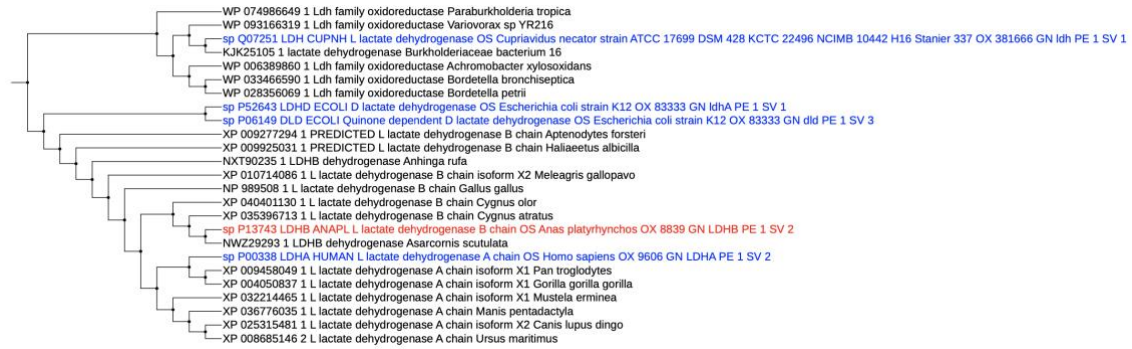
Enzima que esta envuelta en el subproceso de sintetizar lactato a partir de piruvato por un proceso de fermentación en la proteína objetivo de los organismos *Ornithorhynchus anatinus* y *Anas platyrhynchos* en el cristalino del ojo.<sup>47–49</sup> Esta función es realizada de igual forma en las proteínas que se han analizado en *Homo sapiens*, *Escherichia coli* y *Cupriavidus necator*, la única diferencia es que una de las secuencias de *Escherichia coli* en lugar de realizar la oxidación con NADH lo lleva a cabo con un quinol.<sup>50–53</sup>

Como se observa en los árboles filogenéticos de las *figuras 7 y 8*, la secuencia que menos distancia evolutiva tiene con la proteína objetivo es la que se encontró en *Homo sapiens*, esto es plausible debido a que, las otras secuencias analizadas son del dominio de las Bacterias y están en otra rama diferente a la del dominio Eukaryota. Además, que la secuencia de *Homo sapiens* es la que presenta mayor porcentaje de similitud con las secuencias objetivo.



**Figura 7:** Árbol filogenético de la enzima *Upsilon*-cristalino, la cual es una lactato deshidrogenasa que se encuentra en el cristalino, en este caso del organismo *Ornithorhynchus anatinus*. Se pueden observar diferentes ramas en el árbol donde encontramos resaltado en rojo la secuencia objetivo y en azul las secuencias analizadas.



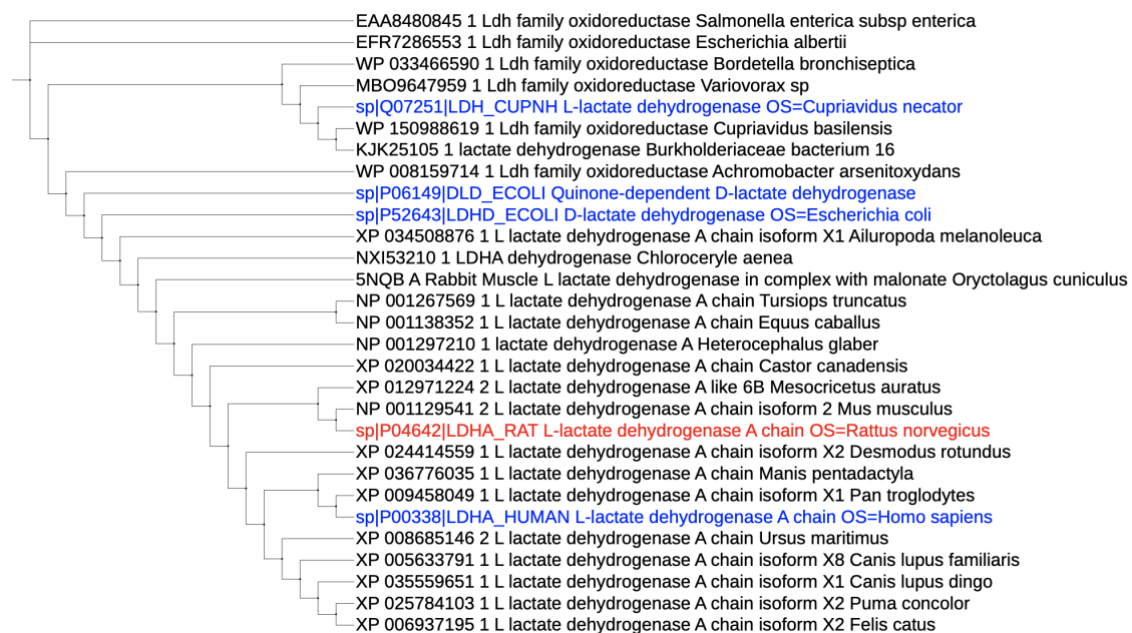


**Figura 8:** Árbol filogenético de la enzima epsilon-cristalina, la cual es una lactato deshidrogenasa que se encuentra en el cristalino, en este caso del organismo *Anas platyrhynchos*. Se pueden observar diferentes ramas en el árbol donde encontramos resaltado en rojo la secuencia objetivo y en azul las secuencias analizadas.

### Lactato deshidrogenasa cadena A

Enzima que esta envuelta en el subproceso de sintetizar lactato a partir de piruvato por un proceso de fermentación en la proteína objetivo del organismo *Rattus norvegicus* en el musculo.<sup>54</sup> Esta función es realizada de igual forma en las proteínas que se han analizado en *Homo sapiens*, *Escherichia coli* y *cupriavidus necator*, la única diferencia es que hay una de las secuencias de *Escherichia coli* que en lugar de realizar la oxidación con NADH lo lleva a cabo con un quinol.<sup>50-53</sup>

Como se observa en el árbol filogénico de la *figura 9*, la secuencia que menos distancia evolutivamente tiene con la proteína objetivo, junto con mayor porcentaje de similaridad, es la que se encontró en *Homo sapiens*, esto es plausible debido a que, las otras secuencias analizadas son del dominio de las Bacterias y están en otra rama diferente a la del dominio Eucaryota.

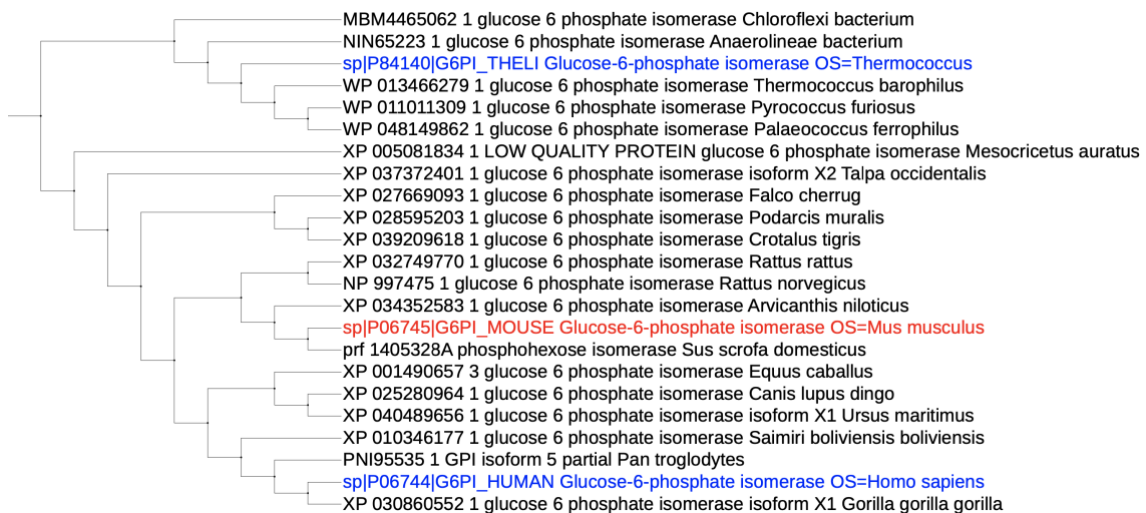


**Figura 9:** Árbol filogenético de una lactato deshidrogenasa muscular, donde se pueden observar diferentes ramas en las que encontramos resaltado en rojo la secuencia objetivo y en azul las secuencias analizadas.

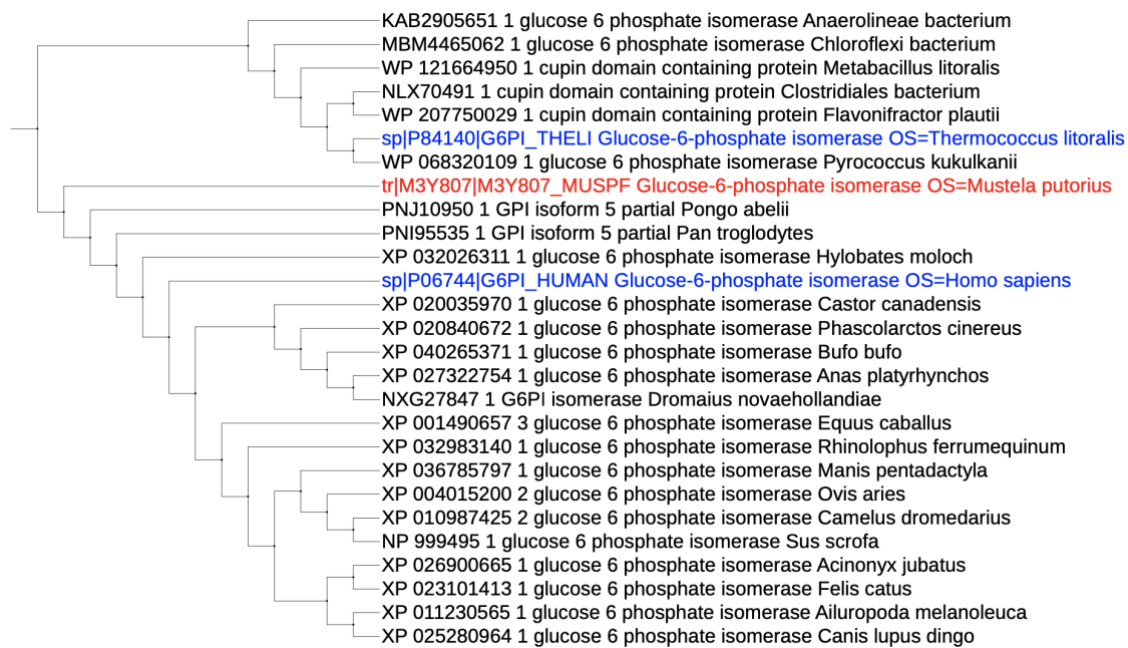
#### Glucosa-6-fosfato isomerasas

Enzima que está envuelta en el subproceso de síntesis del gliceraldehído-3-fosfato y dihidroxiacetona fosfato a partir de glucosa en el citoplasma. Esta función es realizada tanto por las secuencias objetivo de *Mus musculus* y *Mustela putorius furo*, como en las secuencias analizadas de *Homo sapiens* y *Thermococcus litoralis*.<sup>55-58</sup>

En los árboles filogenéticos de *Mus musculus* y *Mustela putorius furo* que son las figuras 10 y 11 respectivamente, se puede observar que la secuencia que se encuentra en una misma rama evolutiva es la de *Homo sapiens*, resultado razonable ya que es del dominio Eukaryota como las secuencias objetivo, mientras que *Thermococcus litoralis* es del dominio Archaea. Además de que en ambos árboles la secuencia en *Homo sapiens* es la que presenta mayor porcentaje de similitud con la secuencia objetivo.



**Figura 10:** Árbol filogenético de la enzima glucosa-6-fosfato isomerasa centrada en la secuencia del organismo *Mus musculus* que encontramos resaltada en rojo, mientras que las secuencias NOGD se encuentran resaltadas en azul.

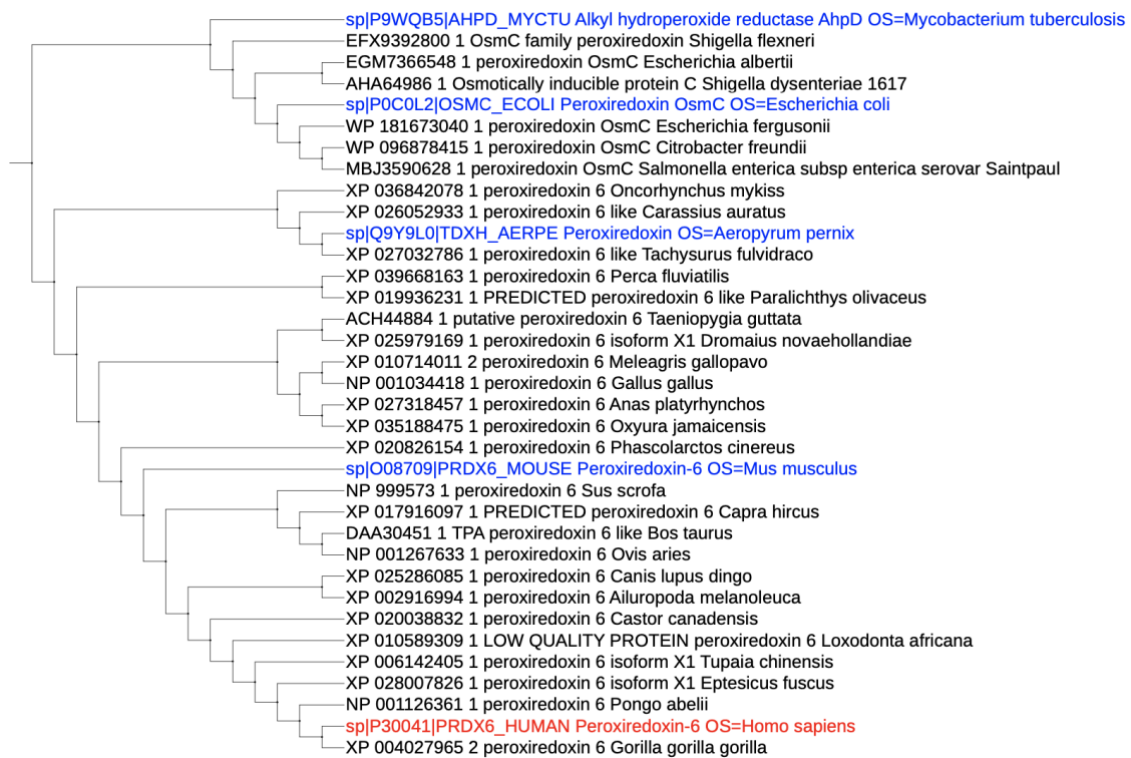


**Figura 11:** Árbol filogenético de la enzima glucosa-6-fosfato isomerasa centrada en la secuencia del organismo *Mustela putorius furo* que encontramos resaltada en rojo, mientras que las secuencias analizadas se encuentran resaltadas en azul.

## Peroxisredoxina

Enzima que cataliza la reducción del peróxido de hidrogeno e hidroperóxidos orgánicos en agua y alcohol respectivamente, tanto en la secuencia objetivo en *homo sapiens* como en los organismos analizados *Aeropyrum pernix*, *Mycobacterium tuberculosis*, *Escherichia coli* y *Mus musculus*.<sup>59-63</sup>

En el árbol filogenético que se enseña en la *figura 12* se puede observar que la secuencia analizada de *Mus musculus* es la más cercana a la secuencia objetivo, esto puede deberse a que tanto, *Homo sapiens* como *Mus musculus* son del dominio Eukaryota, mientras que *Aeropyrum pernix*, *Mycobacterium tuberculosis* y *Escherichia coli* son de los dominios *Archaea* y *Bacteria*. También encontramos que la secuencia de *Homo sapiens* es la que presenta un mayor porcentaje de similitud con la secuencia objetivo.



**Figura 12:** *Árbol filogenético de la enzima peroxirredoxina en la que la secuencia objetivo de Homo sapiens se encuentra resaltada en rojo, mientras que las secuencias analizadas se encuentran resaltadas en azul.*

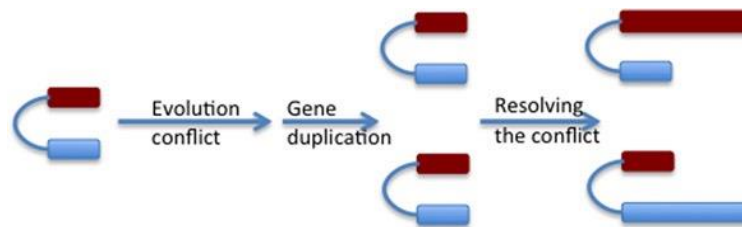
## Discusión

En las últimas décadas han ido apareciendo diversas conexiones entre muchísimos procesos biológicos que nos han hecho plantearnos como científicos hasta que punto nos va a sorprender la evolución. Es por esto, que al pensar en el fenómeno NOGD, en el cual cuando una proteína desaparece de una ruta metabólica, encontramos otra, que realiza esa función en algunos casos conservando la que tenía anteriormente. Esto nos hace replantearnos la forma en la que se realiza este intercambio de funciones o cuando se empezó a emplear sin el fenómeno del splicing alternativo. Gracias a las herramientas de predicción bioinformáticas, actualmente se pueden realizar gran cantidad de estudios como de interacción entre proteínas, de similaridad, de dominios conservados o como el que ocupa a este trabajo sobre el origen evolutivo de cada secuencia.

Actualmente cuando se estudia evolutivamente una proteína, es muy complicado analizar cuando se pudo llegar a producir esta adaptación evolutiva al incluir una nueva función, ya que no hay mucha información sobre la conexión entre la multifuncionalidad y el fenómeno NOGD. Aunque se ha llegado a estimar mediante algoritmos que por la información ofrecida por UniProt, entre el 8 – 24% del proteoma humano podrían ser proteínas moonlighting, mientras que las predicciones de las interacciones entre proteínas sugiere que el 3% del interactoma puede llegar a estar formado por proteínas multifuncionales.<sup>64</sup>

Las vías evolutivas que conducen a las proteínas a adquirir esta función extra aún se desconocen, aunque existen hipótesis en las que se cree que la duplicación génica tiene algo que ver, creando así por duplicación otro gen al que posteriormente se le aplican “retoques” dando así genes parálogos. Estos genes poco a poco van modificando su secuencia hasta obtener una función diferente cada una y que es la conocida, posteriormente al producirse el fenómeno NOGD son estas las que suplen a la otra proteína para realizar su función.<sup>65</sup> Este es un método inteligente y económico por parte de la naturaleza que permite aprovechar los recursos, debido a que simplemente reutiliza algo que ya ha sido ensayado con anterioridad y ha funcionado.

La duplicación además resuelve el conflicto inicial de que una proteína moonlight presente funciones alternativas. Ya que mediante esta duplicación génica, cada una de las copias se especializa en una de las funciones y si pasase un fenómeno de NOGD en alguna de ellas se llegarían a suplir, como se puede observar en la *figura 13*.<sup>66</sup>



**Figura 13:** Se observa como mediante la duplicación génica se puede llegar a resolver el conflicto en el que una proteína moonlight presente diversas funciones alternativas.<sup>66</sup>

Existen otras hipótesis, como es la de la neutral teoría de la evolución de Kimura, en la que las mutaciones que se producen constantemente y que se van fijando en el reloj molecular son selectivamente neutras y es en esta neutralidad donde se van desarrollando zonas moonlighting.<sup>10,67</sup>

En este proyecto se ha podido estudiar mediante la creación de árboles filogenéticos la relación entre proteínas multifuncionales y el fenómeno NOGD donde no hay que olvidar que las proteínas objetivo que hemos analizado, el nombre que presentan es el de la función principal por las que se debieron descubrir y no el de la función moonlight. Es, por tanto, que se ha intentado buscar su origen evolutivo contrastándolas con otras secuencias y donde se ha podido observar que siempre existe alguna secuencia NOGD que presenta este origen común con la secuencia multifuncional gracias a la homología remota.

Una vez creado todos los arboles filogenéticos, se ha podido comprobar que en todos los experimentos, la proteína NOGD que se encontraba más cercana evolutivamente era la que provenía de un organismo que se encontraba en el mismo dominio, reino y filo. Con estos resultados podemos llegar a pensar que, aunque la especiación hasta llegar a un ancestro común entre los organismos más coincidentes es muy lejana, el metabolismo no ha evolucionado tanto, puesto que es muy probable que estos ancestros comunes ya tuviesen un

metabolismo suficientemente adaptado a las condiciones ambientales que incluso tenemos hoy en día. Es por eso que aunque cuando comparamos las secuencias no son totalmente idénticas si que presentan una similaridad tan alta entre ellas, tanto en longitud de las secuencias como en los aminoácidos que presentaba cada una, ya que el hecho de que dos secuencias estuviesen muy emparentadas evolutivamente se podría haber debido a algún fenómeno explicado anteriormente y es por eso que la secuencia objetivo con algunas de las NOGD presentase un porcentaje de similitud normalmente mínimo del 40% mientras que con las que simplemente ha coincidido la función los porcentajes de similaridad en la mayoría de los casos no supera el 20% de similitud.

El estudio de la evolución de las proteínas multifuncionales da un abanico de oportunidades en el futuro en conocer como se reguló su expresión y actividad durante la evolución para poder llegar a emplearlo en el diseño de proteínas con nuevas funciones bioquímicas a través de la utilización de una proteína andamio y luego añadirle mediante cambios en la estructura, nuevos sitios funcionales.<sup>68</sup>

Otro motivo de estudio de las proteínas moonlighting es que, se ha descubierto que el número de estas que están involucradas en enfermedades es muy elevado. Un ejemplo es la fosfoglucosa isomerasa, enzima de la glucólisis, pero que también es una citoquina que presenta un papel importante en la metástasis en el cáncer de pecho.<sup>69</sup> Otro ejemplo es el de la enzima alcohol acetaldehído deshidrogenasa que también presenta la función de proteína de adhesión de *Listeria* (LAP, *Listeria* adhesión protein), que permite la unión de la *Listeria monocytogenes* a las células epiteliales intestinales y favorecer la infección.<sup>70</sup> Es por esto que muchas proteínas moonlighting pueden estar involucradas en la infección, respuesta inmune o virulencia, dando así una perspectiva de futuro optimista en el desarrollo de nuevas terapias siendo dianas farmacológicas.<sup>68</sup>

Además, algunas proteínas multifuncionales están involucradas en vías metabólicas actuando como nexo de unión y ayudando así a las células a responder a los cambios en el entorno (como de sustratos, pH o concentraciones de iones entre otros). Suelen ser reguladoras de la transcripción o traducción o como proteínas que producen algún tipo de retroalimentación celular.<sup>68</sup> Un ejemplo es la timidilato sintasa cuya función es la síntesis de timidilato, pero que también se encuentra como proteína de unión al RNA.<sup>71</sup>

Entre estas aplicaciones, también podemos encontrar que las proteínas moonlighting pueden servir para mejorar los algoritmos en la predicción de funciones proteicas. Esto es debido a que el análisis por homología puede no ser útil ya que hay secuencias que pueden llegar tener funciones que no cuadren con el análisis. También pueden servir para dilucidar los mecanismos por los cuales se realiza una función u otra ya que incluso se han hecho estudios en los cuales los dominios de una proteína colocados de formas diferentes pueden llegar a tener efectos opuestos en vías metabólicas.<sup>68</sup>

Debido a todo lo anterior, es que la llegada de las proteínas moonlighting son un nicho de estudio el cual esta empezando a florecer pero que seguramente en un futuro no muy lejano lleguen a ser empleadas en diversos ámbitos científicos,

farmacológicos o médicos mejorando drásticamente la calidad científica y de vida mundial.

## Conclusiones

### Conclusiones

Mediante la realización de este TFM se ha podido comprobar que:

1. Existe una relación clara entre el fenómeno NOGD y las proteínas multifuncionales gracias a las relaciones evolutivas que se han podido comprobar tanto por homología de secuencias como mediante árboles filogenéticos.
2. De todas las proteínas analizadas solamente se han estudiado un pequeño grupo para la realización de esta tesis. Por tanto, es posible que con un mayor número de proteínas analizadas y se podrían obtener resultados mucho más amplios.
3. A partir de resultados bioinformáticos como estos, se debería realizar experimentos para poder confirmar la posible multifuncionalidad de las proteínas.
4. El ámbito de las proteínas moonlighting se encuentra muy poco estudiado y con el tiempo puede llegar a dar muchos avances en campos como la medicina o la farmacología.
5. Habitualmente la proteína moonlighting pertenece a solo una de las múltiples formas NOGD de una proteína. Demostrando así que solamente con esa NOGD presenta el mismo origen evolutivo, mientras que las otras proteínas NOGD pertenecen a otros linajes. Por tanto, como consecuencia encontramos que hay proteínas que se denominan igual por la función que realizan, pero no por ser la misma proteína desde el punto de vista evolutivo.

### Líneas de futuro

Debido al tiempo limitado que se presenta en la realización de trabajos como este se han dejado de explorar líneas de futuro como el estudio de similitud y comparación de plegado tridimensional de las proteínas por desviación media cuadrática (RMSD, *Root-mean-square deviation*). También se podría haber explorado el analizar la conservación de los dominios de las proteínas analizadas por PSI-BLAST.

### Seguimiento de la planificación

A lo largo de la realización de este TFM se ha seguido la planificación como la metodología impuesta perfectamente en cada PEC, donde en los momentos en

los que se necesitaba una búsqueda de herramientas o aprender a manejarlas ajustaba un poco más el tiempo para poder llegar a hacer el análisis. Además de que al ser la comunicación tan buena con el director del trabajo, no se han tenido que introducir cambios debido a que ya desde antemano se sabía que era lo que mínimamente se tenía que hacer y si todo el proceso fuese a mejor donde sobrase incluso tiempo se hablaba de ir ampliando el trabajo.

## Glosario

NOGD = Non-Orthologous Gene Displacement

HGT = Horizontal Gene Transfer

GAPDH = Glyceraldehyde 3-phosphate dehydrogenase

PSI-BLAST = Position-Specific Iterative Basic Local Alignment Search Tool

MSA = Multiple Sequence Alignment

iTOL = Interactive Tree Of Life

VEGF = Vascular Endothelial Growth Factor

LAP = Listeria Adhesión Protein

RMSD = Root-Mean-Square Deviation

## Bibliografía

1. Galperin MY, Koonin E V. Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement and operon disruption. *In Silico Biol.* 1998;1(1):55-67.
2. Omelchenko M V., Galperin MY, Wolf YI, Koonin E V. Non-homologous isofunctional enzymes: A systematic analysis of alternative solutions in enzyme evolution. *Biol Direct.* 2010;5:31. doi:10.1186/1745-6150-5-31
3. Martínez-Carranza E, Barajas H, Alcaraz LD, Servín-González L, Ponce-Soto GY, Soberón-Chávez G. Variability of bacterial essential genes among closely related bacteria: The case of *Escherichia coli*. *Front Microbiol.* 2018;9(MAY):1-7. doi:10.3389/fmicb.2018.01059
4. Jeffery CJ. Protein moonlighting: What is it, and why is it important? *Philos Trans R Soc B Biol Sci.* 2018;373(1738). doi:10.1098/rstb.2016.0523
5. Jeffery CJ. Moonlighting proteins - An update. *Mol Biosyst.* 2009;5(4):345-350. doi:10.1039/b900658n
6. Huberts DHEW, van der Klei IJ. Moonlighting proteins: An intriguing mode of multitasking. *Biochim Biophys Acta - Mol Cell Res.* 2010;1803(4):520-525. doi:10.1016/j.bbamcr.2010.01.022
7. Copley SD. Enzymes with extra talents: Moonlighting functions and catalytic promiscuity. *Curr Opin Chem Biol.* 2003;7(2):265-272. doi:10.1016/S1367-5931(03)00032-2
8. Jeffery CJ. An introduction to protein moonlighting. *Biochem Soc Trans.* 2014;42(6):1679-1683. doi:10.1042/BST20140226
9. Franco-Serrano L, Hernández S, Calvo A, et al. MultitaskProtDB-II: An update of a database of multitasking/moonlighting proteins. *Nucleic Acids Res.*



- 2018;46(D1):D645-D648. doi:10.1093/nar/gkx1066
10. Henderson B, Martin ACR. The biological and biomedical consequences of protein moonlighting. *Biochem Soc Trans.* 2014;42(6):1671-1678. doi:10.1042/BST20140273
  11. UniProt. Accessed May 7, 2021. <https://www.uniprot.org/>
  12. Gene Ontology Resource. Accessed May 7, 2021. <http://geneontology.org/>
  13. Proteins. Accessed May 7, 2021. [http://wallace.uab.es/multitaskII/proteins\\_list.php](http://wallace.uab.es/multitaskII/proteins_list.php)
  14. National Center for Biotechnology Information. Accessed May 7, 2021. <https://www.ncbi.nlm.nih.gov/>
  15. RCSB PDB: Homepage. Accessed May 7, 2021. <https://www.rcsb.org/>
  16. Analogous Enzymes 2010 update. Accessed May 7, 2021. [https://www.ncbi.nlm.nih.gov/Complete\\_Genomes/AnalEnzymes.html](https://www.ncbi.nlm.nih.gov/Complete_Genomes/AnalEnzymes.html)
  17. Bateman A, Martin MJ, Orchard S, et al. UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Res.* 2021;49(D1):D480-D489. doi:10.1093/nar/gkaa1100
  18. Multiple Sequence Alignment - CLUSTALW. Accessed May 7, 2021. <https://www.genome.jp/tools-bin/clustalw>
  19. Clustal Omega < Multiple Sequence Alignment < EMBL-EBI. Accessed May 7, 2021. <https://www.ebi.ac.uk/Tools/msa/clustalo/>
  20. Bhagwat M, Aravind L. PSI-BLAST tutorial. *Methods Mol Biol.* 2007;395:177-186. doi:10.1385/1-59745-514-8:177
  21. Protein BLAST: search protein databases using a protein query. Accessed May 7, 2021. [https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE\\_TYPE=BlastSearch&PROGRAM=blastp&BLAST\\_PROGRAMS=psiBlast](https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch&PROGRAM=blastp&BLAST_PROGRAMS=psiBlast)
  22. Lemoine F, Correia D, Lefort V, et al. NGPhylogeny.fr: New generation phylogenetic services for non-specialists. *Nucleic Acids Res.* 2019;47(W1):W260-W265. doi:10.1093/nar/gkz303
  23. Lefort V, Longueville JE, Gascuel O. SMS: Smart Model Selection in PhyML. *Mol Biol Evol.* 2017;34(9):2422-2424. doi:10.1093/molbev/msx149
  24. Price MN, Dehal PS, Arkin AP. FastTree 2 - Approximately maximum-likelihood trees for large alignments. *PLoS One.* 2010;5(3). doi:10.1371/journal.pone.0009490
  25. Lefort V, Desper R, Gascuel O. FastME 2.0: A comprehensive, accurate, and fast distance-based phylogeny inference program. *Mol Biol Evol.* 2015;32(10):2798-2800. doi:10.1093/molbev/msv150
  26. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Syst Biol.* 2010;59(3):307-321. doi:10.1093/sysbio/syq010
  27. NGPhylogeny.fr. Accessed May 8, 2021. <https://ngphylogeny.fr/>
  28. Letunic I, Bork P. Interactive Tree of Life (iTOL) v4: Recent updates and new developments. *Nucleic Acids Res.* 2019;47(W1):W256. doi:10.1093/nar/gkz239
  29. iTOL: Interactive Tree Of Life. Accessed May 8, 2021. <https://itol.embl.de/>
  30. Atas H, Tuncbag N, Doğan T. Phylogenetic and other conservation-based approaches to predict protein functional sites. *Methods Mol Biol.* 2018;1762:51-

69. doi:10.1007/978-1-4939-7756-7\_4
31. Jin X, Liao Q, Wei H, Zhang J, Liu B. SMI-BLAST: a novel supervised search framework based on PSI-BLAST for protein remote homology detection. *Bioinformatics*. Published online 2020. doi:10.1093/bioinformatics/btaa772
  32. NOS3 - Nitric oxide synthase, endothelial - Homo sapiens (Human) - NOS3 gene & protein. Accessed April 29, 2021. <https://www.uniprot.org/uniprot/P29474#function>
  33. NOA1 - NO-associated protein 1, chloroplastic/mitochondrial precursor - Arabidopsis thaliana (Mouse-ear cress) - NOA1 gene & protein. Accessed April 30, 2021. <https://www.uniprot.org/uniprot/Q66GP9>
  34. Nos1 - Nitric oxide synthase, brain - Rattus norvegicus (Rat) - Nos1 gene & protein. Accessed April 30, 2021. <https://www.uniprot.org/uniprot/P29476>
  35. eya - Developmental protein eyes absent - Drosophila melanogaster (Fruit fly) - eya gene & protein. Accessed April 29, 2021. <https://www.uniprot.org/uniprot/Q05201>
  36. Eya3 - Eyes absent homolog 3 - Mus musculus (Mouse) - Eya3 gene & protein. Accessed April 30, 2021. <https://www.uniprot.org/uniprot/P97480>
  37. PTPRD - Receptor-type tyrosine-protein phosphatase delta precursor - Homo sapiens (Human) - PTPRD gene & protein. Accessed April 30, 2021. <https://www.uniprot.org/uniprot/P23468>
  38. CDC25C - M-phase inducer phosphatase 3 - Homo sapiens (Human) - CDC25C gene & protein. Accessed April 30, 2021. <https://www.uniprot.org/uniprot/P30307>
  39. ywqE - Tyrosine-protein phosphatase YwqE - Bacillus subtilis (strain 168) - ywqE gene & protein. Accessed April 30, 2021. <https://www.uniprot.org/uniprot/P96717>
  40. At2g35040 - AICAR transformylase - Arabidopsis thaliana (Mouse-ear cress) - At2g35040 gene & protein. Accessed April 29, 2021. <https://www.uniprot.org/uniprot/Q8RWT5>
  41. ATIC - Bifunctional purine biosynthesis protein ATIC - Homo sapiens (Human) - ATIC gene & protein. Accessed April 30, 2021. <https://www.uniprot.org/uniprot/P31939>
  42. purO - IMP cyclohydrolase - Methanotherix thermoacetophila (strain DSM 6194 / JCM 14653 / NBRC 101360 / PT) - purO gene & protein. Accessed April 30, 2021. <https://www.uniprot.org/uniprot/A0B9N7>
  43. TGM2 - Protein-glutamine gamma-glutamyltransferase 2 - Homo sapiens (Human) - TGM2 gene & protein. Accessed April 30, 2021. <https://www.uniprot.org/uniprot/P21980>
  44. F13A1 - Coagulation factor XIII A chain precursor - Homo sapiens (Human) - F13A1 gene & protein. Accessed April 30, 2021. <https://www.uniprot.org/uniprot/P00488>
  45. tgl - Protein-glutamine gamma-glutamyltransferase - Bacillus subtilis (strain 168) - tgl gene & protein. Accessed April 30, 2021. <https://www.uniprot.org/uniprot/P40746>
  46. Protein-glutamine gamma-glutamyltransferase precursor - Streptomyces mobaraensis. Accessed April 30, 2021. <https://www.uniprot.org/uniprot/P81453>
  47. L-lactate dehydrogenase - Ornithorhynchus anatinus (Duckbill platypus).

- Accessed April 30, 2021. <https://www.uniprot.org/uniprot/Q7YQK6>
48. Van Rheede T, Amons R, Stewart N, De Jong WW. Lactate dehydrogenase A as a highly abundant eye lens protein in platypus (*Ornithorhynchus anatinus*): Upsilon ( $\upsilon$ )-crystallin. *Mol Biol Evol.* 2003;20(6):994-998. doi:10.1093/molbev/msg116
  49. LDHB - L-lactate dehydrogenase B chain - *Anas platyrhynchos* (Mallard) - LDHB gene & protein. Accessed April 30, 2021. <https://www.uniprot.org/uniprot/P13743>
  50. LDHA - L-lactate dehydrogenase A chain - *Homo sapiens* (Human) - LDHA gene & protein. Accessed April 30, 2021. <https://www.uniprot.org/uniprot/P00338>
  51. Ldh - L-lactate dehydrogenase - *Cupriavidus necator* (strain ATCC 17699 / DSM 428 / KCTC 22496 / NCIMB 10442 / H16 / Stanier 337) - Ldh gene & protein. Accessed April 30, 2021. <https://www.uniprot.org/uniprot/Q07251>
  52. LdhA - D-lactate dehydrogenase - *Escherichia coli* (strain K12) - LdhA gene & protein. Accessed April 30, 2021. <https://www.uniprot.org/uniprot/P52643>
  53. dld - Quinone-dependent D-lactate dehydrogenase - *Escherichia coli* (strain K12) - dld gene & protein. Accessed April 30, 2021. <https://www.uniprot.org/uniprot/P06149>
  54. Ldha - L-lactate dehydrogenase A chain - *Rattus norvegicus* (Rat) - Ldha gene & protein. Accessed April 30, 2021. <https://www.uniprot.org/uniprot/P04642>
  55. Gpi - Glucose-6-phosphate isomerase - *Mus musculus* (Mouse) - Gpi gene & protein. Accessed April 30, 2021. <https://www.uniprot.org/uniprot/P06745>
  56. GPI - Glucose-6-phosphate isomerase - *Mustela putorius furo* (European domestic ferret) - GPI gene & protein. Accessed April 30, 2021. <https://www.uniprot.org/uniprot/M3Y807>
  57. GPI - Glucose-6-phosphate isomerase - *Homo sapiens* (Human) - GPI gene & protein. Accessed April 30, 2021. <https://www.uniprot.org/uniprot/P06744>
  58. pgiA - Glucose-6-phosphate isomerase - *Thermococcus litoralis* - pgiA gene & protein. Accessed April 30, 2021. <https://www.uniprot.org/uniprot/P84140>
  59. PRDX6 - Peroxiredoxin-6 - *Homo sapiens* (Human) - PRDX6 gene & protein. Accessed May 3, 2021. <https://www.uniprot.org/uniprot/P30041>
  60. APE\_2278 - Peroxiredoxin - *Aeropyrum pernix* (strain ATCC 700893 / DSM 11879 / JCM 9820 / NBRC 100138 / K1) - APE\_2278 gene & protein. Accessed May 3, 2021. <https://www.uniprot.org/uniprot/Q9Y9L0>
  61. ahpD - Alkyl hydroperoxide reductase AhpD - *Mycobacterium tuberculosis* (strain ATCC 25618 / H37Rv) - ahpD gene & protein. Accessed May 3, 2021. <https://www.uniprot.org/uniprot/P9WQB5>
  62. osmC - Peroxiredoxin OsmC - *Escherichia coli* (strain K12) - osmC gene & protein. Accessed May 3, 2021. <https://www.uniprot.org/uniprot/POC0L2>
  63. Prdx6 - Peroxiredoxin-6 - *Mus musculus* (Mouse) - Prdx6 gene & protein. Accessed May 3, 2021. <https://www.uniprot.org/uniprot/O08709>
  64. Singh N, Bhalla N. Moonlighting Proteins. *Annu Rev Genet.* 2020;54:265-285. doi:10.1146/annurev-genet-030620-102906
  65. Espinosa-Cantú A, Ascencio D, Barona-Gómez F, De Luna A. Gene duplication and the evolution of moonlighting proteins. *Front Genet.* 2015;6(JUL). doi:10.3389/fgene.2015.00227
  66. Fares MA. The evolution of protein moonlighting: Adaptive traps and

- promiscuity in the chaperonins. *Biochem Soc Trans.* 2014;42(6):1709-1714. doi:10.1042/BST20140225
67. Kimura M. The neutral theory of molecular evolution: A review of recent evidence. *Japanese J Genet.* 1991;66(4):367-386. doi:10.1266/jjg.66.367
  68. Jeffery CJ. Why study moonlighting proteins? *Front Genet.* 2015;6(JUN):211. doi:10.3389/fgene.2015.00211
  69. Watanabe H, Carmi P, Hogan V, et al. Purification of human tumor cell autocrine motility factor and molecular cloning of its receptor. *J Biol Chem.* 1991;266(20):13442-13448. doi:10.1016/s0021-9258(18)98859-9
  70. Jagadeesan B, Koo OK, Kim KP, et al. LAP, an alcohol acetaldehyde dehydrogenase enzyme in *Listeria*, promotes bacterial adhesion to enterocyte-like Caco-2 cells only in pathogenic species. *Microbiology.* 2010;156(9):2782-2795. doi:10.1099/mic.0.036509-0
  71. Chu E, Koeller DM, Casey JL, et al. Autoregulation of human thymidylate synthase messenger RNA translation by thymidylate synthase. *Proc Natl Acad Sci U S A.* 1991;88(20):8977-8981. doi:10.1073/pnas.88.20.8977

## Anexos

1. Tabla de proteínas NOGD y proteínas moonlighting: [Tabla moonlight y NOGD](#)
2. Carpeta con todos los alineamientos de secuencias: [Alineamientos de secuencias](#)
3. Carpeta con las secuencias utilizadas para realizar los árboles filogenéticos: [Secuencias de árboles](#)