Universitat Oberta de Catalunya

uoc.edu

Búsqueda y Comparación de Epítopos en Proteínas Multitarea

Josep Rivas Prieto

Máster universitario en Bioinformática y bioestadística UOC-UB

TFM – Bioinformática y Bioestadística Área 2 Aula 1

Consultor: Luis Franco Serrano

Profesor/a responsable de la asignatura: Carles Ventura Royo

23 de diciembre de 2021

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Búsqueda y Comparación de Epítopos en Proteínas Multitarea*
Nombre del autor:	*José Rivas Prieto*
Nombre del consultor/a:	*Luis Franco Serrano*
Nombre del PRA:	*Carles Ventura Royo*
Fecha de entrega (mm/aaaa):	12/2021
Titulación:	*Máster universitario en Bioinformática y bioestadística UOC-UB*
Área del Trabajo Final:	*TFM – Bioinformática y Bioestadística Área 2 Aula 1*
Idioma del trabajo:	*Castellano*
Palabras clave	• *Proteínas multitarea* • *Epítopos* • *Enfermedades autoinmunes*

Resumen del Trabajo:

El TFM presente se ha centrado en la consecución de dos objetivos: la creación de una base de datos de epítopos de las proteínas multitarea presentes en la base datos MultitaskProtDB-II y la obtención de los mimotopos de las proteínas multitarea con estructura proteica análoga a 1iok_A con el fin de comprobar si presentan mimetismo con proteínas humanas asociadas a patologías.

La predicción de epítopos se ha realizado con Ellipro, en el caso que se conociera su código PDB, y con Bepipred-2.0, en caso contrario, previa obtención de la secuencia en formato FASTA de Uniprot.

En una primera fase, se ha realizado una alineación de secuencias con Uniprot Align (mediante Clustal Omega), identificándose con exactitud los epítopos análogos de las seis proteínas que, a su vez, han sido analizados en Tomtom, obteniéndose 19 mimotopos consenso.

En una segunda fase, se ha realizado un análisis comparativo de los mimotopos en FASTM, obteniéndose una lista de cuarenta proteínas humanas con un alto grado de similitud, a partir de la cual se ha creado una base de datos de las patologías asociadas, extraídas de Open Targets y de DisGeNet, de cada proteína de la lista.

Finalmente, a partir de la lista de motivos de ELM análogos a cada mimotopo que ha proporcionado Tomtom, se ha creado una base de datos de las proteínas con motivos análogos a los 3 primeros mimotopos y se ha generado una base de datos patologías asociadas a las proteínas con motivos análogos al primer mimotopo.

Abstract:

The present Master's Final Work has focused on the achievement of two objectives: the creation of a database of epitopes of the multitasking proteins present in the MultitaskProtDB-II database and the obtaining of mimotopes of the multitasking proteins with a protein structure analogous to 1iok_A, in order to check if they present mimicry with human proteins associated with pathologies.

Epitope prediction was performed with Ellipro, if its PDB code was known and with Bepipred-2.0, if not, after obtaining the sequence in FASTA format from Uniprot.

In a first phase, a sequence alignment was performed with Uniprot Align (using Clustal Omega), accurately identifying the analogous epitopes of the six proteins which, in turn, have been analyzed in Tomtom, obtaining 19 consensus mimotopes.

In a second phase, a comparative analysis of the mimotopes in FASTM has been carried out, obtaining a list of forty human proteins with a high degree of similarity, from which has been created a database of associated pathologies, extracted from Open Targets and DisGeNet, for each protein on the list.

Finally, from the list of ELM motifs analogous to each mimotope provided by Tomtom, a database of proteins with motifs analogous to the first 3 mimotopes has been created and, too, a database with the associated pathologies of the proteins with motifs analogous to the first mimotope.

Índice

Lista de Figuras

Lista de Tablas

1. Introducción

1.1 Contexto y Justificación

Las proteínas multitarea **[1]** son proteínas, descubiertas recientemente que presentan funciones alternativas a la canónica, realizadas por una sola cadena polipeptídica, como consecuencia de diferencias en el estado, condiciones y/o localización en la que se puedan encontrar. **[2, 3, 4]**.

Esta capacidad de cumplir diversas funciones es un factor importante a nivel experimental, ya que, al participar en diferentes funciones, puede perturbar el resultado de un experimento, el 48% de las proteínas multitarea humanas son objetivo de los fármacos actuales **[5]**, complicando tanto la interpretación de los ensayos como aumentando la posibilidad de generar efectos secundarios de los fármacos objeto de estudio.

Por otro lado, recientes investigaciones indican que un elevado porcentaje de las mismas pueden estar implicadas en diversas enfermedades humanas, participando en la invasión de tejidos en cáncer, en infecciones de microorganismos patógenos, el 25% de las proteínas de la base de datos corresponden a funciones de pluriempleo relacionadas con la actividad de virulencia de patógenos **[6]**, o actuando de desencadenantes de enfermedades autoinmunes.

La mayoría de las proteínas multitarea presentan una secuencia de aminoácidos altamente conservados al pertenecer habitualmente al metabolismo central (glucólisis, ciclo de Krebs…) **[7]**.

El hecho que, por la importancia crítica del metabolismo central, este tipo de proteínas tengan una baja tasa de mutación implica que se hayan conservado mucho evolutivamente, lo que, según las últimas hipótesis, podría haber provocado que ciertos organismos patógenos también hayan aprovechado esa situación, por selección positiva, para mantener una alta conservación de las proteínas implicadas en el proceso de infección del huésped y evitar inducir una respuesta del sistema inmune del hospedador durante la infección **[8]**.

Al existir un riesgo importante que se genere una respuesta autoinmune sobre proteínas con funciones canónicas fundamentales para la supervivencia de la célula, se cree que el organismo hospedador no genera anticuerpos contra las proteínas patógenas con las que comparte epítopos.

Una prueba que refuerza esta hipótesis es que actualmente no existe una vacuna exitosa basada en una proteína multitarea.

Por otro lado, el mismo hecho de compartir mimetismo de epítopos podría provocar, en ciertas circunstancias, un error en el funcionamiento del sistema inmune del hospedador que diera como consecuencia la aparición de enfermedades autoinmunes**.**

Cabe destacar que el descubrimiento de las proteínas multitarea es muy reciente **[9]** y, pese a que su investigación está creciendo exponencialmente en los últimos años, todavía presenta muchas lagunas de conocimiento.

En primer lugar, hay que tener en cuenta que las proteínas del pluriempleo actualmente se revelan experimentalmente por casualidad y que los métodos actuales para detectarlas todavía están en fases tempranas de evolución por lo que posiblemente aún falten muchas por descubrir.

Actualmente existen sólo tres bases de datos actualizadas de proteínas multitarea: *MultitaskProtDB-II* (694 proteínas), *MoonProt* **[10]** (512 proteínas) y *MoonDB* **[11]** (345 proteínas).

Además del número limitado de proteínas multitarea que se conocen, hay varios mecanismos intrínsecos de las mismas que también se desconocen. Por ejemplo, el mecanismo de secreción de las proteínas citoplasmáticas del pluriempleo aún no se comprende bien.

De esta forma, pese a que las proteínas patógenas son intracelulares y carecen de un péptido señal canónico o motivos de secreción, se desconoce como éstas son capaces de salir de la bacteria y adherirse se a la membrana celular, consiguiendo interactuar y con la matriz extracelular del huésped (MEC) [12].

Igualmente, el conocimiento profundo de la estructura intrínseca de las proteínas multitarea es limitado, siendo inexistente una base de datos de epítopos que permita realizar comparaciones entre los epítopos de diferentes especies y que permita realizar de forma más automatizada la búsqueda de patrones comunes entre los epítopos de proteínas patógenas y de hospedadores.

Todos estos factores dan una idea de la importancia de obtener herramientas que faciliten la predicción e identificación de estas proteínas, en especial para el sector sanitario y farmacéutico, puesto que una mejor comprensión de las mismas podría permitir la obtención de información sobre la base molecular de las enfermedades de base genética, mejorar el diseño de fármacos y comprender mejor el mecanismo de infección de los patógenos.

Este TFM pretende profundizar en el conocimiento de las proteínas multitarea a partir del el análisis y comparación de los epítopos de diferentes especies patógenas y la búsqueda de relaciones entre la aparición de enfermedades autoinmunes y neoplasmas en el ser humano y la coinfección por bacterias o virus, de forma que pueda facilitar avances significativos en la lucha contra las enfermedades de base genética.

Entrando en más detalle, la idea de crear específicamente una base de datos de epítopos viene condicionada porqué, hasta el momento, no existe actualmente ninguna base de datos de este tipo y no es posible realizar comparaciones entre epítopos de diferentes especies de forma rápida y automatizada, siendo necesario realizar cualquier tipo de comparación entre epítopos de especies patógenas y hospedadoras de forma individual y manual.

La intención del TFM, por tanto, es crear una base de datos que permita poder seleccionar varias especies patógenas y realizar tanto análisis comparativos de epítopos, con programas como *Tomtom*, como análisis estadísticos de los datos obtenidos, puesto que dentro del proyecto se prevé convertir la base de datos a *R*, lo que permitirá realizar los análisis de forma más sencilla e, incluso, crear algún tipo de aplicación con Shiny con la que poder realizar los análisis de forma automatizada.

1.2 Objetivos

El objetivo principal del TFM es la creación de una base de datos de epítopos de proteínas multitarea, englobado dentro del proyecto de ampliación de la base de datos *MultitaskProtDB-II*, promovido por el Institut de Biotecnologia i Biomedicina de la Universitat Autònoma de Barcelona.

Como complemento a la base de datos generada, y como prueba de la utilidad de la misma para futuras investigaciones, se ha previsto realizar un análisis comparativo de los epítopos de proteínas con la misma función canónica de diferentes especies patógenas, identificar si presentan mimetismo de epítopos, buscar proteínas humanas con motivos análogos a estos mimotopos que tengan asociadas algún tipo de patología e intentar determinar posibles relaciones causales entre ambos factores.

De esta forma, se puede considerar que el TFM consta de dos objetivos generales que a la vez se subdividen en varios objetivos específicos.

1.2.1 Objetivos generales

Los objetivos generales del proyecto son:

1. Crear una base de datos de epítopos de las proteínas multitarea (moonlighting) de organismos patógenos que se englobará dentro del proyecto ampliación de la base de datos *MultitaskProtDB-II*, promovido por el Institut de Biotecnologia i Biomedicina de la Universitat Autònoma de Barcelona.

2. Identificar mimotopos de proteínas multitarea implicadas en enfermedades y determinar posibles relaciones causales entre ambos factores.

1.2.2 Objetivos específicos

Dentro del primer objetivo general, el de la creación de la base de datos de epítopos, se engloban cuatro objetivos específicos:

1.1 Buscar y seleccionar las herramientas informáticas más adecuadas para la creación de la base de datos de epítopos.

1.2 Obtener los archivos de datos de epítopos de las proteínas multitarea de organismos patógenos y adaptar los mismos a un formato adecuado para su posterior inclusión en la base de datos.

1.3 Crear la base de datos de epítopos en formato Excel y modificar la misma para su duplicación en formato R.

1.4 Analizar estadísticamente con R la base de datos generada, cuantitativa y cualitativamente.

La segunda parte del trabajo, la identificación de mimotopos de proteínas multitarea implicadas en enfermedades autoinmunes y la determinación de posibles relaciones causales entre ambos factores se subdivide en cuatro objetivos específicos:

2.1 Buscar y seleccionar las herramientas informáticas más adecuadas para la identificación de mimotopos.

2.2 Comparar los epítopos de proteínas ortólogas de diferentes especies y detectar patrones comunes y mimótopos.

2.3 Comparar los mimótopos obtenidos con los epítopos de sus correspondientes proteínas ortólogas humanas y determinar posibles relaciones causales entre ambos factores.

1.3 Enfoque y Metodología

El Trabajo de Final de Máster se compone de dos objetivos muy bien delimitados, siendo el primer objetivo eminentemente práctico y el segundo de tipo más analítico y deductivo.

Tanto el primer objetivo del trabajo, la creación de una base de datos de epítopos de proteínas multitarea de organismos patógenos como el segundo objetivo, la comparación de los epítopos de diversos organismos patógenos y la obtención de proteínas humanas con epítopos ortólogos implicadas en enfermedades humanas han requerido de un proceso de investigación previo para obtener una base científica y bibliográfica suficiente para poder planificar y cumplir con los objetivos.

De forma complementaria a este proceso de investigación, también ha sido necesario investigar, analizar y seleccionar las herramientas informáticas más adecuadas para poder realizar los diferentes análisis y completar todas las fases del trabajo.

La primera fase de investigación ha consistido en la selección de la herramienta más adecuada para la predicción de epítopos de las proteínas multitarea. En el proceso de selección se ha priorizado que ésta fuera polivalente y sirviera para predecir cualquier tipo de proteína y que el output de salida fuera sencillo y fácil de incluir en la base de datos.

Las herramientas de predicción de epítopos analizadas han sido: *EpiSearch* **[13]**, *Pepitope* **[14]**, BEPITOPE **[15]**, *Rankpep* **[16, 17, 18]**, *BCPREDS* **[19, 20, 21]**, *BepiPred-2.*0 **[22, 23]**, *Discotope* **[24, 25]** y *ElliPro* **[26]**.

Pese a que lo ideal hubiera sido obtener los epítopos de todas las proteínas multitarea de *MultitaskProtDB-II* (http://wallace.uab.es/multitaskII/proteins_list.php), se ha optado por realizar únicamente la predicción de **253 proteínas** multitarea implicadas en procesos de virulencia, incluidas en un archivo denominado *Moon.xlsx*, incluyendo todos los epítopos predichos en una única base de datos en formato *excel* con el objetivo de limitar el número de proteínas analizadas y tener más tiempo para completar el resto de objetivos del trabajo.

Por otro lado, aunque la idea inicial era utilizar una única herramienta para la predicción de los epítopos, finalmente se ha optado por combinar *ElliPro* en las proteínas de las que se conoce su código *PDB* y *BepiPred-2.0* en las proteínas en las que únicamente se tiene el código *Uniprot* **[27]**.

El primer paso ha consistido, por tanto, en la revisión de los códigos *PDB* mediante la base de datos *RCSB PDB* (https://www.rcsb.org/) **[28]** y la comprobación si éstos son específicos de la proteína anotada o, por el contrario, son simplemente estructuras análogas.

La revisión de códigos ha dado como resultado que sólo se conocía el código PDB de **32 proteínas,** por lo que, aunque *ElliPro* (http://tools.iedb.org/ellipro/) presenta el output más completo, al indicar la cadena, la puntuación del PI (Protrusion Index) y prever los epítopos discontinuos, la herramienta principal de predicción de epítopos ha sido finalmente *Bepipred-2.0*, al utilizarse para realizar la predicción de epítopos de **221 proteínas**.

Cabe indicar, que la predicción de epítopos con *Bepipred-2.0* (http://tools.iedb.org/bcell/) requiere la estructura primaria de la secuencia que se quiere analizar, en formato FASTA, por lo que ha sido necesario obtenerlas previamente de *Uniprot* (https://www.uniprot.org/).

Una vez completada la base de datos de epítopos, denominada *Epítopos.xlsx*, ésta ha sido pre-procesada y cargada en *R* **[29]**. R es un programa que permite tratar los datos de forma mucho más flexible que *excel*, tanto para la selección de datos como para la creación de subtablas, además de permitir la aplicación de paquetes gráficos y estadísticos para realizar análisis cuantitativos y/o cualitativos de las bases de datos cargados.

El último paso del primer objetivo específico ha consistido en estudiar los datos extraídos del análisis de la base de datos de epítopos y seleccionar el grupo de proteínas sobre el cual se ha realizado la segunda fase del trabajo.

El grupo de proteínas escogido ha sido el de las proteínas con estructura análoga a la estructura PDB *1iok_A*, un grupo de formado por 6 proteínas de seis organismos patógenos diferentes del tipo *chaperonas 60kDa*.

Con la finalización del análisis estadístico de la base de datos se ha concluido el primer objetivo del TFM por lo que la siguiente fase del trabajo ha consistido en realizar un análisis y selección de las herramientas más adecuadas para la comparación de epítopos de las proteínas multitarea de organismos patógenos y la búsqueda de secuencias similares de proteínas humanas con patologías asociadas.

Las herramientas de alineación de secuencias múltiples y de búsqueda de similitudes analizadas han sido: *Uniprot Align*, *BLAST* **[30, 31, 32]**, *PSI-BLAST* **[33]**, *Clustal Omega* **[34]**, *FASTA* **[35, 36]**, *FASTM* **[37]**, *T-Coffe Expresso* **[38]** y *Tomtom* **[39, 40, 41]**.

Por medio de dos herramientas de alineación múltiple de secuencias, Uniprot Align (https://www.uniprot.org/align/), que utiliza Clustal Omega como herramienta *T_Coffe Expresso* de alineación de secuencias, y *T_Coffe Expresso* (http://tcoffee.crg.cat/apps/tcoffee/do:expresso), se han alineado las 6 proteínas análogas a la estructura *1iok_A* y se han detectado las regiones peptídicas comunes sobre las que obtener los mimotopos de las proteínas analizadas.

A partir de las regiones comunes de epítopos seleccionadas, se ha utilizado la herramienta de comparación de motivos *Tomtom* (https://meme-suite.org/meme/tools/tomtom) para obtener una lista compuesta por **19 epítopos consenso (mimotopos)** que han sido introducidos en *FASTM* (https://www.ebi.ac.uk/Tools/sss/fastm/), una herramienta para la búsqueda de similitudes de péptidos como los epítopos, con la que se ha obtenido una lista de **40 proteínas con un alto grado de similitud** respecto a la totalidad de mimotopos de 1iok_A.

A partir de la lista de proteínas extraída de *FASTM* (https://www.ebi.ac.uk/Tools/sss/fastm/), por medio de *Uniprot* se ha accedido a dos bases de datos de detección de asociaciones entre proteínas y patologías, *DisGeNet* **[42, 43, 44, 45, 46]** y *Open Targets* **[47]** y se ha creado una nueva base de datos, denominada *1iok_A.xlsx*, compuesta de **41 anotaciones compuestas por 12 variables**, en la que se relacionan las proteínas humanas análogas y el número de referencias de patologías asociadas a las mismas.

Como proceso final de esta fase del trabajo, se ha creado una base de datos, denominada *1iok_A_Disease.xlsx*, con **437 anotaciones compuestas por 4 variables** con la lista de patologías asociadas a las proteínas humanas análogas a 1iok_A extraídas de *DisGeNet* (https://www.disgenet.org/) y *Open Targets (https://platform.opentargets.org/)*.

Una vez analizada la secuencia total de mimótopos obtenidos, se ha decidido realizar un análisis específico de los mimotopos de forma individualizada, ya que por medio de Tomtom se ha obtenido una lista de **139 coincidencias de motivos proteicos** análogos a alguno de los mimotopos de 1iok_A

Estos motivos proteicos se encuentran descritos en *ELM* (*The Eukaryotic Linear Motif Resource*) **[48]** y cada uno de ellos lleva asociado una lista de proteínas tanto humanas como de otros organismos con funciones similares.

El planteamiento base, cuando se obtuvo la lista de mimotopos era analizar las proteínas de los motivos de los 19 mimotopos, pero la gran cantidad de proteínas asociadas a algunos de los motivos extraídos de *ELM* (http://elm.eu.org/) han provocado que, finalmente, solo se analizaran tres mimotopos, de forma que existiera una variedad de datos y motivos que hiciera más representativo el análisis pero que, a la vez, fuera posible completar el resto de objetivos específicos.

De esta forma, se han analizado las proteínas incluidas en la lista de motivos de los tres primeros mimotopos y se ha generado una base de datos, denominada *Mimotopos.xlsx*, con las referencias de patologías asociadas extraídas de DisGeNet y de Open Targets, con **595 anotaciones compuestas por 15 variables**.

Como se puede observar, la lista de anotaciones obtenida ha sido muy grande, y cada una de las anotaciones presenta un número de patologías asociadas que en muchos casos es muy grande, por lo que, se ha optado por crear la lista de patologías asociadas de un único mimotopo, tanto por una cuestión de temporización como por el hecho que esta segunda fase del TFM se ha enfocado finalmente como un ejemplo de protocolo de análisis de epítopos.

Así, como paso final de esta fase, y de igual manera que se ha hecho con el análisis de la secuencia total de mimotopos, se ha creado una base de datos, denominada *IKFXZB_Disease.xlsx*, con **6918 anotaciones compuestas por 4 variables** con la lista de patologías asociadas a las proteínas con motivos análogos al primer de los mimotopos de 1iok_A.

Cabe indicar que en el plan de trabajo del TFM también estaba previsto encontrar secuencias similares en organismos humanos con *PSI-BLAST* (https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE=Proteins&PROGRAM=blastp&RUN_PSIBLAST=on), al ser una buena herramienta para detectar proteínas humanas ortólogas pero, pese a haberse realizado el análisis con *PSI-BLAST*, no se ha realizado ningún análisis posterior con los datos obtenidos, al considerarse más significativo centrarse en el análisis de las secuencias proteicas obtenidas con FASTM y de los motivos proteicos obtenidos con Tomtom.

Finalmente, como fase final de este segundo objetivo general, se ha realizado un análisis estadístico, cuantitativo y cualitativo, de todas las bases de datos generadas de forma que se puedan extraer conclusiones objetivas, y respaldadas por datos, del trabajo realizado.

1.4 Planificación

El TFM ha sido dividido en varias tareas e hitos diferentes, algunos de las cuales se han realizado simultáneamente, en función de si se correspondían a uno u otro objetivo principal.

1.4.1 Tareas

Las tareas previstas se han dividió en función de si pertenecen a uno u otro objetivo o si forman parte del proceso de creación de la memoria final.

1.4.1.1 Objetivo General 1

1.1 Investigación de antecedentes previos.
1.2 Búsqueda y selección de las herramientas informáticas más adecuadas para la creación de la base de datos de epítopos de proteínas multitarea de organismos patógenos.
1.3 Obtención de los archivos de datos de epítopos de las proteínas multitarea de organismos patógenos y adaptación de los mismos a un formato adecuado para su posterior inclusión en la base de datos.
1.4 Creación de la base de datos de epítopos de las proteínas multitarea de organismos patógenos.
1.5 Análisis estadístico de la base de datos de epítopos con R, cuantitativa y cualitativamente.

1.4.1.2 Objetivo General 2

2.1 Búsqueda y selección de las herramientas informáticas más adecuadas para la identificación de mimotopos.

2.2 Comparación de los epítopos de proteínas multitarea ortólogas de diferentes especies mediante alineamientos múltiples y detección de patrones comunes.
2.3 Creación de una secuencia consenso a partir de los alineamientos múltiples realizados.
2.4 Búsqueda de secuencias proteicas humanas similares al epítopo consenso
2.5 Análisis de las funciones de las proteínas identificadas y de su relación con enfermedades.

1.4.1.3 Elaboración de la memoria

3.1 Elaboración de la Propuesta del TFM.
3.2 Elaboración del Plan de Trabajo.
3.3 Elaboración de la Memoria del TFM.
3.4 Elaboración de la Presentación para la Defensa Pública.
3.5 Preparación de la Defensa Pública.

1.4.2 Hitos

Pese a la cantidad y diversidad de tareas previstas, estas se pueden agrupar en cuatro hitos:

A Obtención de los epítopos de todas las proteínas multitarea implicadas en procesos de virulencia.

B Creación de la base de datos de epítopos.

C Obtención de los epítopos consenso de la familia de proteínas multitarea seleccionada.

D Obtención de las proteínas humanas ortólogas a los epítopos consenso y análisis de su relación con enfermedades o procesos de respuesta autoinmune.

1.4.3 Calendario

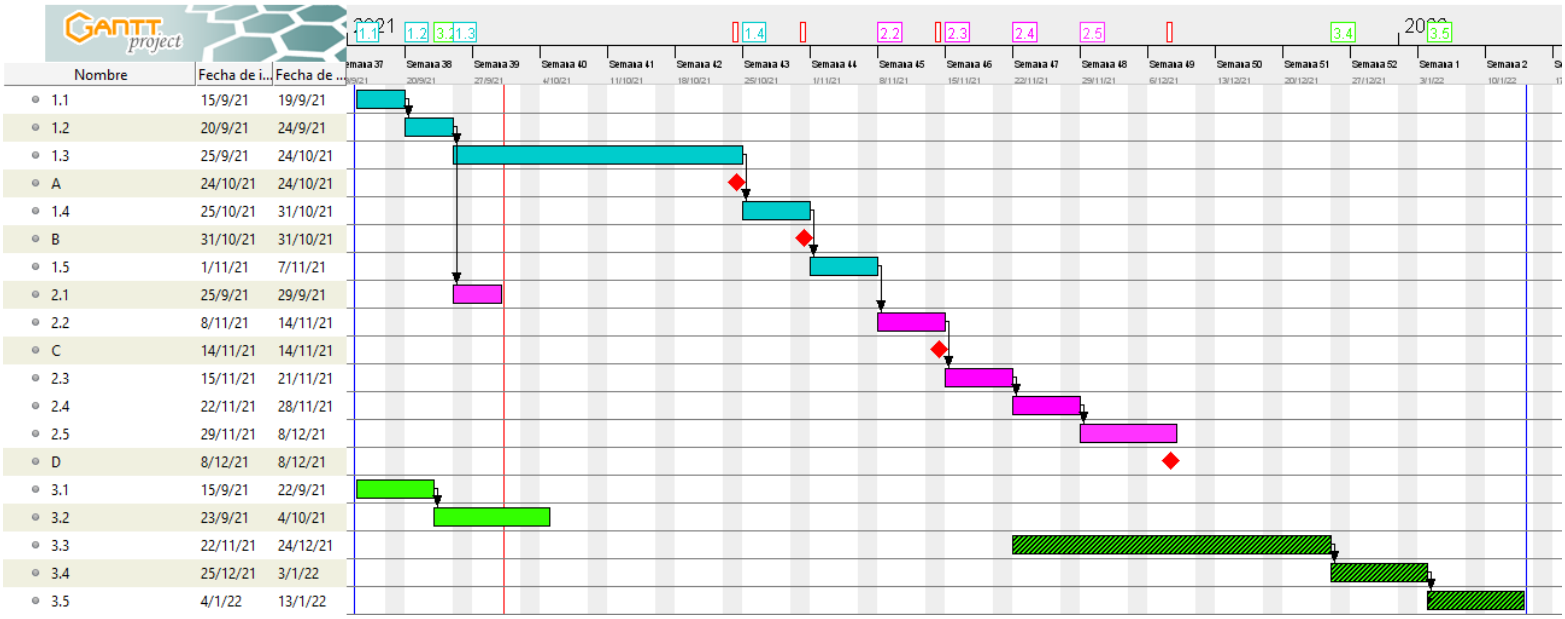
El calendario se ha realizado con la herramienta libre Gantt Project **[49]**.

A la izquierda se presenta la lista de tareas e hitos explicados en los apartados 4.1 y 4.2, mientras que a la derecha se encuentra un calendario en el que cada tarea se encuentra representada en forma de barra de mayor o menor longitud en función de la duración de la misma.

Los colores de cada barra son un indicativo de los objetivos generales a los que pertenecen:

- Azul: Creación de la base de datos de epítopos
- Violeta: Búsqueda de proteínas humanas ortólogas a las proteínas multitarea de la base de datos relacionadas con algún tipo de enfermedad.
- Verde: Pruebas de Evaluación Continuada.

En el calendario también se pueden visualizar los hitos, que se representan mediante un rombo rojo.



1.5 Sumario de Productos Obtenidos

Este TFM es de tipo analítico, por lo que todos los productos generados se corresponden a archivos y bases de datos virtuales.

No existe un producto físico final, pero sí una serie de archivos extraídos durante el proceso de realización del TFM y una serie de bases de datos generadas de novo que son el foco central de los objetivos generales del TFM.

Todos los archivos generados durante el proceso de creación del TFM se han guardado en una carpeta perteneciente al drive de la cuenta de gmail: jrivaspr@uoc.edu, denominada *TFM* que se puede visualizar por medio del siguiente link.

La carpeta presenta 4 subcarpetas con los siguientes archivos:

- Bases de Datos:

  Ésta es la carpeta principal, puesto que incluye todas las bases de datos generadas:

  - *Moon.xlsx*: Base de datos de las proteínas multitarea analizadas.
  - *Epítopos.xlsx*: Base de datos con todos los epítopos de las proteínas multitarea predichos.
  - *PDB.xlsx*: Base de datos con la lista de las proteínas análogas a las proteínas multitarea a las que se han predicho sus epítopos.
  - *1iok_A.xlsx*: Base de datos con las anotaciones de proteínas humanas similares a la secuencia de mimotopos predicha.

- *1iok_A_Disease.xlsx*: Base de datos con la lista de patologías asociadas a las proteínas de la base de datos *1iok_A.xlsx*.
- *Mimotopos.xlsx*: Base de datos con las anotaciones de proteínas humanas con motivos similares a los tres mimotopos analizados individualmente:
  - *IKFXZB.*
  - *G*
  - *GXPX*
- *IKFXZB_Disease.xlsx*: Base de datos con la lista de patologías asociadas a las proteínas con motivos análogos al mimotopo *IKFXZB.*

- Archivos:

  Carpeta con todos los archivos extraídos o generados en el proceso de realización del TFM

  - Alineaciones: Carpeta con los archivos generados para poder realizar la base de datos *Mimotopos.xlsx*. La carpeta consta de 4 archivos:
    - *1iok_A FASTA.txt*: Archivos con los códigos *FASTA* de las seis proteínas multitarea análogas a 1iok_A.
    - *1iok_A FASTM.txt*: Archivo con los resultados del análisis de secuencias análogas a 1iok_A realizado con *FASTM*.
    - *Mimotop_1iok_A.txt*: Archivo con la secuencia de mimotopos obtenida a partir de Tomtom.
    - *Motivos.docx*: Archivo con las secuencias de epítopos con los cuales se han generado los mimotopos y la lista de motivos proteicos análogos a cada uno de los mimotopos.
  - Bepipred: Carpeta con todos los archivos de epítopos generados con *bepipred-2.0* (**228 archivos**), guardados en formato tabulado.
  - Ellipro: Carpeta con todos los archivos de patologías asociadas a las proteínas humanas análogas a los mimotopos analizados. Dentro de la carpeta hay dos tipos de archivos:
    - Archivos con formato xlsx: Archivos de patologías asociadas extraídos de *DisGeNet* (**131 archivos**).
    - Archivos con formato tsv: Archivos de patologías asociadas extraídos de Open Targets (**139 archivos**).
  - FASTA: Carpeta con todos los archivos de estructuras primarias extraídos en formato *FASTA* (**228 archivos**), guardados en formato de texto.
  - Disease: Carpeta con todos los archivos de epítopos generados con Ellipro, guardados en formato tabulado. La carpeta consta de dos subcarpetas:
    - Lineales: Guarda los archivos de epítopos lineales generados (**76 archivos**).
    - Discontinuos: Guarda los archivos de epítopos discontinuos generados (**67 archivos**).

- Planificación:

  Carpeta con la planificación del TFM. Consta de dos archivos:
  - *Planificacion.gan*: Archivo de planificación del TFM.
  - *Planificacion.png*: Archivo en formato de imagen de la planificación del TFM.

- R:

  En esta carpeta se incluyen los archivos generados a partir del análisis de la base de datos realizado con *R*. Consta de cinco archivos y una subcarpeta.
  - *TFM.Rmd*: Archivo en formato *Rmarkdown* a partir del cual se ha realizado el análisis estadístico de la base de datos de epítopos.
  - *TFM.R*: Archivo en formato *R* con el código en bruto utilizado para realizar el análisis de la base de datos de epítopos.
  - *TFM.html*: Archivo de resultados del análisis con *R* en formato web.
  - *TFM.pdf*: Archivo de resultados del análisis con *R* en formato pdf.
  - *TFM.docx*: Archivo de resultados del análisis con *R* en formato de documento de texto.
  - *Imágenes:* Carpeta con las imágenes utilizadas con *Rmarkdown* (**4 imágenes**) para realizar el análisis de la base de datos de epítopos.
  - *Data*: Carpeta con las bases de datos utilizadas en el análisis estadístico con *R*.

1.6 Descripción de los Capítulos de la Memoria

La memoria consta de nueve capítulos, siendo el primer capítulo una introducción a los objetivos y planificación del TFM.

Entre los capítulos dos y seis se explica el trabajo realizado a lo largo del semestre mientras que los tres últimos capítulos se dedicarán a aspectos formales de la memoria.

El capítulo dos engloba todos los procedimientos y tareas realizados para completar el primer objetivo general, que ha consistido en la creación y análisis de una base de datos de epítopos de proteínas multitarea implicadas en procesos de patogenia. Éste consta de dos subapartados principales.

El primer subapartado del capítulo dos se centra en el proceso de creación de la base de datos de epítopos, explicando los diferentes procedimientos realizados para obtener la estructura primaria de las proteínas, o su código PDB cuando era posible, para la predicción de epítopos con las dos herramientas seleccionadas, *Ellipro* y *bepipred-2.0* y para la anotación de los mismos en la base de datos de epítopos.

La segunda parte de este capítulo realiza un análisis estadístico de la base de datos de epítopos tanto a nivel general como específico, centrándose posteriormente en el análisis de las proteínas presentes en la base de datos de *Bacillus anthracis* y de las proteínas del tipo *chaperona de 60 kDa*

El capítulo tres, se centra en el proceso de identificación de mimotopos de proteínas multitarea implicadas en enfermedades, dando, por tanto, inicio al proceso de cumplimiento del segundo objetivo del TFM.

La tipología de proteínas escogida para realizar el análisis ha sido la de las *chaperonas de 60kDA* con estructura análoga a la proteína con código PDB *1iok_A*

De esta forma, la primera parte de este capítulo ha consistido en realizar un análisis estadístico específico en la base de datos de epítopos de las *chaperonas de 60kDA con estructura análoga 1iok_A*.

Una vez realizado el análisis esstadístico, la segunda parte de este capítulo muestra el proceso de alineación de las secuencias de este grupo de proteínas con *Align* y *T-Coffe Expresso* y la posterior obtención de los mimotopos con *Tomtom*.

En el capítulo cuatro se describe el proceso obtención de proteínas con secuencias análogas al conjunto de mimotopos generados en el capítulo anterior y a la creación de una base de datos con la lista de proteínas análogas y a otra con las patologías asociadas a esas proteínas.

Así, el primer apartado del capítulo explica el proceso de realización del análisis comparativo de secuencias con FASTM para obtener proteínas humanas con una secuencia análoga al conjunto de mimotopos.

El segundo apartado describe el proceso de creación de la base de datos de proteínas análogas y el análisis estadístico de la misma mientras que la última sección del capítulos realiza un proceso de descripción y análisis similar al apartado anterior pero sobre la base de datos de patologías asociadas a las proteínas de la base de datos del apartado anterior.

El último capítulo del TFM dedicado al análisis de datos es el capítulo cinco, que presenta una estructura similar al capítulo cuatro pero que, en este caso, se basa en el análisis y comparación de los mimotopos generados pero a nivel individual.

Por tanto, en el primer apartado del capítulo también se obtendrá una lista de proteínas análogas a los mimotopos, aunque en este caso la lista generada será de motivos proteicos análogos, extraídos de *ELM*, y la base de datos correspondiente guardará únicamente las proteínas de tres mimotopos, *IKFXZB*, *G* y *GXPX*.

Como en el capítulo anterior, también se realizará un análisis estadístico de la base de datos generada, en la sección dos del capítulo, y se creará una nueva base de datos con la lista de patologías asociadas aunque, en este caso, ésta se ceñirá exclusivamente a las proteínas con motivos análogos al mimotopo *IKFXZB*.

Con los datos generados y analizados, el capítulo seis servirá para extraer conclusiones sobre los datos generados y analizar qué cambios se podrían realizar en el procedimiento realizado o que nueva líneas de investigación podrían derivarse como ampliación del trabajo.

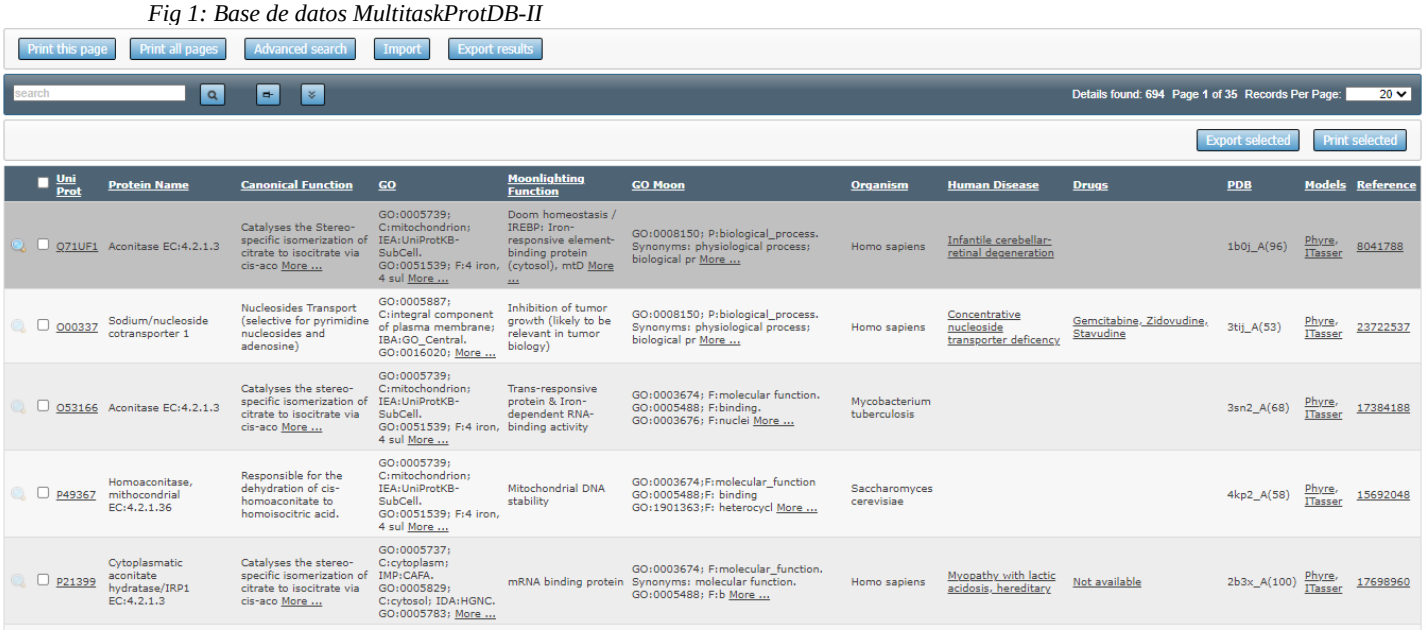
Para concluir, el capítulo siete proporcionará un glosario de términos científicos y acrónimos utilizados en el TFM, el capítulo 8 listará todas las referencias bibliográficas utilizadas y el capítulo nueve incluirá los documentos o enlaces anexos con información suplementaria a la presentada durante el TFM.

2. Análisis de la Bases de Datos de Epítopos

2.1 Realización de la Base de Datos de Epítopos

Antes de iniciar el proceso de creación de la base de datos de epítopos, se ha analizado la información contenida en la base de datos de proteínas multitarea *MultitaskProtDB-II* (http://wallace.uab.es/multitaskII/proteins_list.php).

Fig 1: Base de datos MultitaskProtDB-II



Uni Prot	Protein Name	Canonical Function	GO	Moonlighting Function	GO Moon	Organism	Human Disease	Drugs	PDB	Models	Reference
Q71UF1	Aconitase EC:4.2.1.3	Catalyses the Stereo-specific isomerization of citrate to isocitrate via cis-aco More ...	GO:0005739; C:mitochondrion; IEA:UniProtKB-SubCell. GO:0051539; F:4 iron, 4 sul More ...	Doom homeostasis / IREBP: Iron-responsive element-binding protein (cytosol), mtD More ...	GO:0008150; P:biological_process. Synonyms: physiological process; biological pr More ...	Homo sapiens	Infantile cerebellar-retinal degeneration		1b0j_A(96)	Phyre, ITasser	8041788
O00337	Sodium/nucleoside cotransporter 1	Nucleosides Transport (selective for pyrimidine nucleosides and adenosine)	GO:0005887; C:integral component of plasma membrane; IBA:GO_Central. GO:0016020; More ...	Inhibition of tumor growth (likely to be relevant in tumor biology)	GO:0008150; P:biological_process. Synonyms: physiological process; biological pr More ...	Homo sapiens	Concentrative nucleoside transporter deficency	Gemcitabine, Zidovudine, Stavudine	3tij_A(53)	Phyre, ITasser	23722537
O53166	Aconitase EC:4.2.1.3	Catalyses the stereo-specific isomerization of citrate to isocitrate via cis-aco More ...	GO:0005739; C:mitochondrion; IEA:UniProtKB-SubCell. GO:0051539; F:4 iron, 4 sul More ...	Trans-responsive protein & Iron-dependent RNA-binding activity	GO:0003674; F:molecular function. GO:0005488; F:binding. GO:0003676; F:nucleI More ...	Mycobacterium tuberculosis			3sn2_A(68)	Phyre, ITasser	17384188
P49367	Homoaconitase, mithocondrial EC:4.2.1.36	Responsible for the dehydration of cis-homoaconitate to homoisocitric acid.	GO:0005739; C:mitochondrion; IEA:UniProtKB-SubCell. GO:0051539; F:4 iron, 4 sul More ...	Mitochondrial DNA stability	GO:0003674;F:molecular_function GO:0005488;F: binding GO:1901363;F: heterocycl More ...	Saccharomyces cerevisiae			4kp2_A(58)	Phyre, ITasser	15692048
P21399	Cytoplasmatic aconitate hydratase/IRP1 EC:4.2.1.3	Catalyses the stereo-specific isomerization of citrate to isocitrate via cis-aco More ...	GO:0005737; C:cytoplasm; IMP:CAFA. GO:0005829; C:cytosol; IDA:HGNC. GO:0005783; More ...	mRNA binding protein	GO:0003674; F:molecular_function. Synonyms: molecular function. GO:0005488; F:b More ...	Homo sapiens	Myopathy with lactic acidosis, hereditary	Not available	2b3x_A(100)	Phyre, ITasser	17698960

Pese a que lo ideal hubiera sido obtener los epítopos de todas las proteínas multitarea de *MultitaskProtDB-II* (http://wallace.uab.es/multitaskII/proteins_list.php), se ha optado por realizar únicamente la predicción de **253 proteínas** multitarea implicadas en procesos de virulencia, incluidas en un archivo denominado *Moon.xlsx*, incluyendo todos los epítopos predichos en una única base de datos en formato *excel* con el objetivo de limitar el número de proteínas analizadas y tener más tiempo para completar el resto de objetivos del trabajo.

Una de las variables incluida en la base de datos se corresponde al código *PDB* de la estructura tridimensional análoga a la proteína multitarea indicada, pero en la fase inicial de análisis se creyó que se correspondía a la estructura exacta de la proteína representada, siendo una situación que se da en pocos casos.

Este error de comprensión provocó que se tuvieran que revisar todos los archivos de epítopos generados con *Ellipro* mediante la base de datos *RCSB PDB* (https://www.rcsb.org/) y que se descartaran para la base de datos los que no se correspondían con la estructura real de las proteínas analizadas.

Como consecuencia del error de comprensión inicial con los códigos PDB, se realizó la predicción de epítopos con Ellipro de varias proteínas que no eran objeto del trabajo pero, al representar estructuras análogas de las proteínas multitarea se han guardado los archivos generados con la intención de aprovecharlos en futuros análisis y se ha creado una nueva base de datos, *PDB.xlsx*, en la que se han incluido las características principales de éstas.

La base de datos *PDB.xlsx* consta de **76** anotaciones con las siguientes variables:

- Organisme: Variable factorial que indica el organismo al cual pertenece el epítopo.
- Proteína: Variable factorial que indica la proteína a la cual pertenece el epítopo.
- Uniprot: Variable factorial que indica el código Uniprot de la proteína a la que pertenece el epítopo.
- PDB: Variable factorial que indica el código PDB de la estructura a la que es análoga la proteína.
- Tool: herramienta con la que se ha realizad la predicción de epítopos:
  e = *Ellipro*
  b = *Bepipred-2.0*

Fig 2: Base de datos PDB.xlsx

	A	B	C	D	E
1	Organismo	Proteína	UniProt	PDB	BD
2	*Staphylococcus aureus*	Triosephosphate isomerase	P68823	3m9y_A	b
3	*Escherichia coli*	fructose-1, 6-bisphosphate aldolase	6ZZQ6	1b57_A	b
4	*Trichomonas vaginalis*	Glyceraldehyde-3-phosphate dehydrogenase		3gnq_A	b
5	*Francisella tularensis* subsp. *tularensis*	60 kDa chaperonin		1aon_O	b
6	*Streptococcus suis*	Enolase		4ewl_A	b
7	*Histoplasma capsulatum*	Peptidyl Prolyl cis,trans-Isomerase		1h0p_A	b
8	*Pseudomonas aeruginosa*	Translation elongation factor Tu	A7L763	4zv4_A	e
9	*Streptococcus gordonii*	Elongation factor G		2bv3_A	b
10	*Streptococcus gordonii*	Protein translocase subunit SecA		1tf5_A	b
11	*Streptococcus gordonii*	Elongation factor Tu		1efc_A	b
12	*Streptococcus gordonii*	Enolase		1w6t_A	b
13	*Mycoplasma fermentans*	Enolase		4a3r_A	b
14	*Neisseria meningitidis serogroup C (strain 053442)*	Chaperone protein DnaK		1dkg_D	b
15	*Mycobacterium avium* subsp. *avium*	Superoxide dismutase	B1A031	1gn3_A	e
16	*Streptococcus pneumoniae*	CbpA	B1N649	2pms_C	e
17	*Streptococcus pneumoniae*	Choline binding protein G	B2ILT3	2v04_A	e
18	*Streptococcus pneumoniae*	Choline binding protein D	B2INL9	2ww5_A	e

Una vez completada la revisión de los códigos *PDB* y la comprobación de si éstos son específicos de la proteína anotada o, por el contrario, son simplemente estructuras análogas, se han eliminado un par de entradas duplicadas y se ha renombrado el código Uniprot de algunas de las proteínas multitarea, adaptándolas a los nuevos códigos que ha implementado Uniprot.

Los cambios realizados a la base de datos inicial han sido las siguientes:

- A1KFR2 -> El código estaba duplicado y se ha eliminado la entrada incorrecta: *Mycobacterium tuberculosis* 60 kDa chaperonin.
- C2GUHO -> El código estaba duplicado y se ha eliminado la entrada incorrecta: *Bifidobacterium longum*.
- Q96UH7 -> El código estaba duplicado y se ha eliminado la entrada incorrecta: *Paracoccidioides brasiliensis* Aldolase.
- Q97QS2 -> El código estaba duplicado y se ha eliminado la entrada incorrecta: *Streptococcus pneumoniae* Glyceraldehyde-3-phosphate dehydrogenase.
- A0A2X1W3A9 -> Se ha modificado el código de *meningitidis Peroxiredoxin* (Código original: C9WZY8).
- A0A0B7LRW6 -> Se ha modificado el código de *Streptococcus pneumoniae HtrA* (Código original: D3R759).
- W8S2K4 -> Se ha modificado el código de *Erysipelothrix rhusiopathiae* Glyceraldehyde-3-phosphate dehydrogenase (Código original: F5CUQ4).
- A0A0L0LUA6 -> Se ha modificado el código de *Bifidobacterium breve Enolase* (Código original: F9XY86).
- A0A0L0LUA6 -> Se ha modificado el código de *Streptococcus suis* 6-phosphogluconate dehydrogenase (Código original: G5L1C5).

- A0A126UNJ8 -> Se ha modificado el código de *Streptococcus pneumoniae* Glyceraldehyde-3-phosphate dehydrogenase (Código original: G6LM38).
- A0A1D2ITN5S -> Se ha modificado el código de *Listeria monocytogenes* Glyceraldehyde-3-phosphate dehydrogenase (Código original: L8DW58).
- E9KNS1 -> Se ha modificado el código de *Staphylococcus epidermidis* Ornithine carbamoyltransferase (Código original: M1XHM3).
- A0A113S8Y8 -> Se ha modificado el código de *Plasmodium berghei* Enolase (Código original: Q4YQJ5)

Finalmente, se ha procedido a la anotación de los datos revisados y corregidos en una base de datos denominada *Moon.xlsx*. La base de datos consta **de 254 anotaciones** con las siguientes variables por anotación:

- Organisme: Variable factorial que indica el organismo al cual pertenece el epítopo.
- Proteína: Variable factorial que indica la proteína a la cual pertenece el epítopo.
- Patogen*: Variable factorial que indica el tipo de patogenia de la proteína a la que pertenece el epítopo. La variable cuenta con cuatro factores representados de forma simplificada:

  P = Organismo Patógeno

  N = Organismo No Patógeno

  C = Organismo Comensal/Simbionte

  O = Organismo Patógeno Oportunista
- Uniprot: Variable factorial que indica el código Uniprot de la proteína a la que pertenece el epítopo.
- Moon: Variable factorial que indica si la proteína a la que pertenece el epítopo es una proteína multitarea.
- Review: Variable factorial que indica si la proteína que presenta el epítopo ha sido revisada y se corresponde con el código indicado.
- Estructura: Variable factorial que indica si se conoce alguna variante estructural análoga a la proteína que presenta el epítopo.
- PDB: Variable factorial que indica el código PDB de la estructura a la que es análoga la proteína.
- Tool: herramienta con la que se ha realizad la predicción de epítopos:
  e = *Ellipro*
  b = *Bepipred-2.0*

Fig 3: Base de datos Moon.xlsx

	A	B	C	D	E	F	G	H	I
1	Organisme	Proteïna	Patogen*	UniProt	Moon	Review	Estructura	PDB	Tool
2	*Bifidobacterium bifidum*	Glutamine Synthetase	C	A0A0H2Q1X0	S	N	N	1eae0_A	b
3	*Lactobacillus casei*	Class A sortase	C	A0A0K1MVR4	S	N	N		b
4	*Staphylococcus lugdunensis*	Glyceraldehyde-3-phosphate dehydrogenase	P	A0A133Q850	N	N	N		b
5	*Clostridium perfringens*	Aldehyde-alcohol dehydrogenase	P	A0A174GBI5	N	N	N		b
6	*Streptococcus anginosus*	Phosphoglycerate kinase	O	A0A1S1FAM2	S	N	N		b
7	*Polytolypa hystricis*	Isocitrate lyase	N	A0A2B7X647	N	N	N		b
8	*Polytolypa hystricis*	Triosephosphate isomerase	N	A0A2B7Z228	N	N	N		b
9	*Orenia marismortui*	Glyceraldehyde-3-phosphate dehydrogenase	N	A0A4R8GI02	N	N	N		b
10	*Macrococcus canis*	Glyceraldehyde-3-phosphate dehydrogenase	P	A0A6G7EP66	N	N	N		b
11	*Aeromonas hydrophila subsp. hydrophil*	Phosphoglycerate kinase	P	A0KGD3	S	S	N	1zmr_A	b
12	*Mycobacterium smegmatis*	60 kDa chaperonin 1	C	A0QQU5	S	S	N	3rtk_A	b
13	*Mycobacterium bovis*	60 kDa chaperonin 1	P	A1KFR2	S	S	S	3rtk_B	b
14	*Trichomonas vaginalis*	Glyceraldehyde-3-phosphate dehydrogenase	P	A2DHT2	S	N	N	3gnq_A	b
15	*Francisella tularensis subsp. tularensis*	60 kDa chaperonin	P	A4IWC5	S	S	N	1aon_O	b
16	*Streptococcus suis*	Enolase	P	A4W2T1	S	S	N	4ewj_A	b
17	*Histoplasma capsulatum*	Peptidyl Prolyl cis,trans-Isomerase	P	A6R3K7	S	N	N	1h0p_A	b
18	*Saccharomyces cerevisiae*	Aldolase	O	A6ZZQ6	S	N	N	1b57_A	b
19	*Streptococcus pneumoniae*	Ancillary pilus subunit	P	A7KT66	S	N	N		b
20	*Pseudomonas aeruginosa*	Translation elongation factor Tu	P	A7L763	S	N	N	4zv4_A	e

La idea inicial era utilizar como herramienta para la predicción de los epítopos *ElliPro*, que es más sencillo en el proceso de predicción de epítopos y presenta un output de salida más completo y *BepiPred-2.0* en las proteínas en las que no se tuviera el código *PDB* de la proteína, pero la revisión ha dejado patente que sólo se conocía el código *PDB* de **32 proteínas,** por lo que, aunque *ElliPro* (http://tools.iedb.org/ellipro/) presenta el output más completo, al indicar la cadena, la puntuación del PI (Protrusion Index) y prever los epítopos discontinuos, la herramienta principal de predicción de epítopos ha terminado siendo *Bepipred-2.0*, que se ha utilizado al para realizar la predicción de epítopos de **222 proteínas**.

En *Ellipro* sólo es necesario incluir el código *PDB* en la pestaña de selección y el código de cadena, si hay varias, en la ventana posterior para poder realizar la predicción de epítopos.

Fig. 4: Pantalla de selección de predicción de epítopos de Ellipro de la proteína con código PDB 4zv4

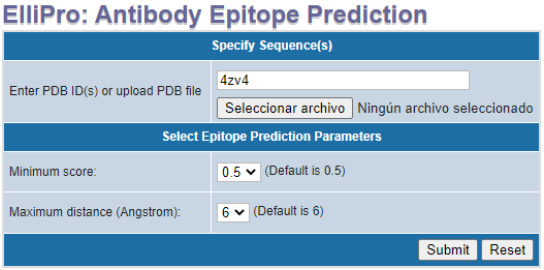


Fig. 5: Pantalla de selección de cadena de Ellipro

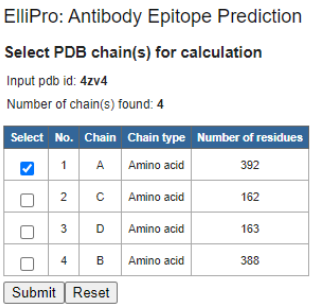
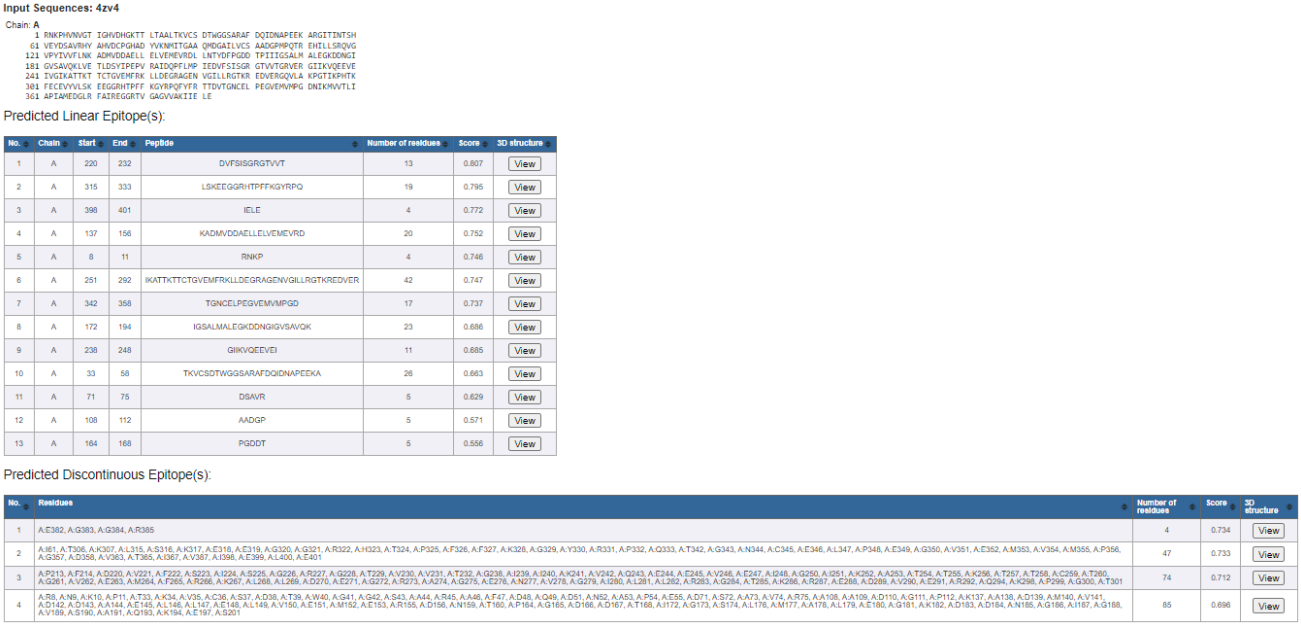
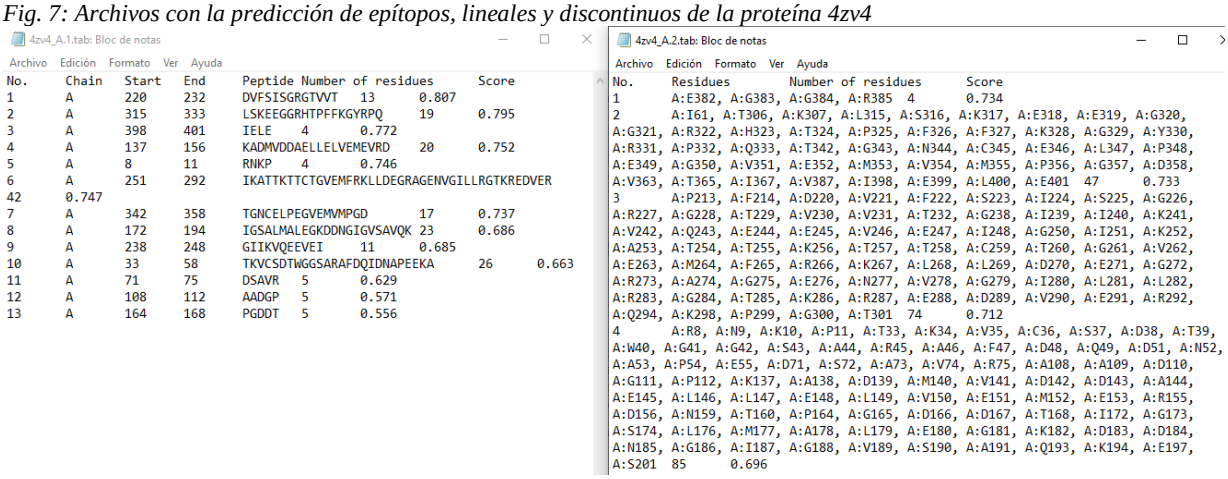


Fig. 6: Predicción de epítopos de Ellipro de la proteína 4zv4



Como output de salida, *Ellipro* realiza la predicción de los epítopos lineales y discontinuos. Ambos outputs de salida se han copiado y anotado en dos archivos, uno para los epítopos lineales, denominado n*ombre_de_la_proteína.1.tab* y el otro para los epítopos discontinuos, denominado n*ombre_de_la_proteína.2.tab*.

Fig. 7: Archivos con la predicción de epítopos, lineales y discontinuos de la proteína 4zv4



4zv4_A.1.tab: Bloc de notas

Archivo Edición Formato Ver Ayuda

No.	Chain	Start	End	Peptide	Number of residues	Score
1	A	220	232	DVFSISGRGTVVT	13	0.807
2	A	315	333	LSKEEGGRHTPFFKGYRPQ	19	0.795
3	A	398	401	IELE	4	0.772
4	A	137	156	KADMVDDAELLELVEMEVRD	20	0.752
5	A	8	11	RNKP	4	0.746
6	A	251	292	IKATTKTTCTGVEMFRKLLDEGRAGENVGILLRGTKREDVER		
42	0.747					
7	A	342	358	TGNCELPEGVEMVMPGD	17	0.737
8	A	172	194	IGSALMALEGKDDNGIGVSAVQK	23	0.686
9	A	238	248	GIIKVQEEVEI	11	0.685
10	A	33	58	TKVCSDTWGGSARAFDQIDNAPEEKA	26	0.663
11	A	71	75	DSAVR	5	0.629
12	A	108	112	AADGP	5	0.571
13	A	164	168	PGDDT	5	0.556

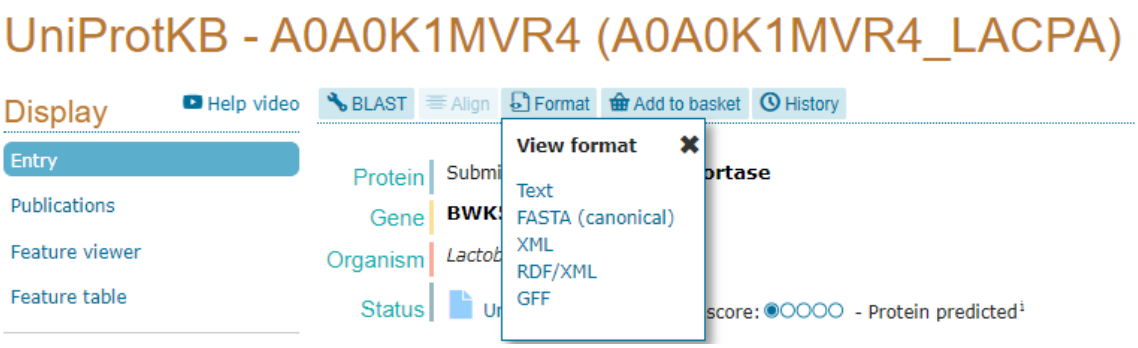
4zv4_A.2.tab: Bloc de notas

Archivo Edición Formato Ver Ayuda

No.	Residues	Number of residues	Score
1	A:E382, A:G383, A:G384, A:R385	4	0.734
2	A:I61, A:T306, A:K307, A:L315, A:S316, A:K317, A:E318, A:E319, A:G320, A:G321, A:R322, A:H323, A:T324, A:P325, A:F326, A:F327, A:K328, A:G329, A:Y330, A:R331, A:P332, A:Q333, A:T342, A:G343, A:N344, A:C345, A:E346, A:L347, A:P348, A:E349, A:G350, A:V351, A:E352, A:M353, A:V354, A:M355, A:P356, A:G357, A:D358, A:V363, A:T365, A:I367, A:V387, A:I398, A:E399, A:L400, A:E401	47	0.733
3	A:P213, A:F214, A:D220, A:V221, A:F222, A:S223, A:I224, A:S225, A:G226, A:R227, A:G228, A:T229, A:V230, A:V231, A:T232, A:G238, A:I239, A:I240, A:K241, A:V242, A:Q243, A:E244, A:E245, A:V246, A:E247, A:I248, A:G250, A:I251, A:K252, A:A253, A:T254, A:T255, A:K256, A:T257, A:T258, A:C259, A:T260, A:G261, A:V262, A:E263, A:M264, A:F265, A:R266, A:K267, A:L268, A:L269, A:D270, A:E271, A:G272, A:R273, A:A274, A:G275, A:E276, A:N277, A:V278, A:G279, A:I280, A:L281, A:L282, A:R283, A:G284, A:T285, A:K286, A:R287, A:E288, A:D289, A:V290, A:E291, A:R292, A:Q294, A:K298, A:P299, A:G300, A:T301	74	0.712
4	A:R8, A:N9, A:K10, A:P11, A:T33, A:K34, A:V35, A:C36, A:S37, A:D38, A:T39, A:W40, A:G41, A:G42, A:S43, A:A44, A:R45, A:A46, A:F47, A:D48, A:Q49, A:D51, A:N52, A:A53, A:P54, A:E55, A:D71, A:S72, A:A73, A:V74, A:R75, A:A108, A:A109, A:D110, A:G111, A:P112, A:K137, A:A138, A:D139, A:M140, A:V141, A:D142, A:D143, A:A144, A:E145, A:L146, A:L147, A:E148, A:L149, A:V150, A:E151, A:M152, A:E153, A:R155, A:D156, A:N159, A:T160, A:P164, A:G165, A:D166, A:D167, A:T168, A:I172, A:G173, A:S174, A:L176, A:M177, A:A178, A:L179, A:E180, A:G181, A:K182, A:D183, A:D184, A:N185, A:G186, A:I187, A:G188, A:V189, A:S190, A:A191, A:Q193, A:K194, A:E197, A:S201	85	0.696

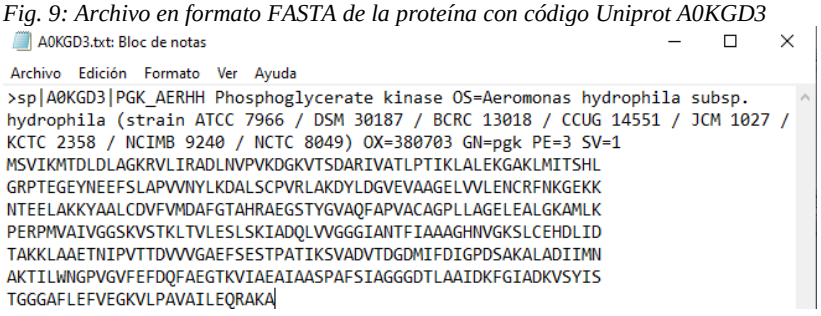
Cabe indicar, que la predicción de epítopos con *Bepipred-2.0* (http://tools.iedb.org/bcell/) requiere la estructura primaria de la secuencia que se quiere analizar, en formato FASTA, por lo que ha sido necesario obtenerlas previamente de *Uniprot* (https://www.uniprot.org/).

Fig. 8: Ventana de selección del archivo en formato FASTA en Uniprot



Todas las secuencias proteicas descargadas en formato FASTA han sido guardadas en un archivo tabulado denominado: n*ombre_de_la_proteína.tab*.

Fig. 9: Archivo en formato FASTA de la proteína con código Uniprot A0KGD3



A0KGD3.txt: Bloc de notas

Archivo Edición Formato Ver Ayuda

```
>sp|A0KGD3|PGK_AERHH Phosphoglycerate kinase OS=Aeromonas hydrophila subsp.
hydrophila (strain ATCC 7966 / DSM 30187 / BCRC 13018 / CCUG 14551 / JCM 1027 /
KCTC 2358 / NCIMB 9240 / NCTC 8049) OX=380703 GN=pgk PE=3 SV=1
MSVIKMTDLDLAGKRVLIRADLNVPVKDGKVTSDARIVATLPTIKLALEKGAKLMITSHL
GRPTEGEYNEEFSLAPVVNYLKDALSCPVRLAKDYLDGVEVAAGELVVLENCRFNKGEKK
NTEELAKKYAALCDVFVMDAFGTAHRAEGSTYGVAQFAPVACAGPLLAGELEALGKAMLK
PERPMVAIVGGSKVSTKLTVLESLSKIADQLVVGGGIANTFIAAAGHNVGKSLCEHDLID
TAKKLAAETNIPVTTDVVVGAEFSESTPATIKSVADVTDGDMIFDIGPDSAKALADIIMN
AKTILWNGPVGVFEFDQFAEGTKVIAEAIAASPAFSIAGGGDTLAAIDKFGIADKVSYIS
TGGGAFLEFVEGKVLPAVAILEQRAKA
```

Una vez obtenida la secuencia primaria de la proteína se ha procedido a realizar la predicción de los epítopos de predicción de epítopos de células B en *IEDB* (http://tools.iedb.org/bcell/)

Fig. 10: Pantalla de predicción de epítopos para Bepipred-2.0 con el código FASTA de la proteína A0KGD3

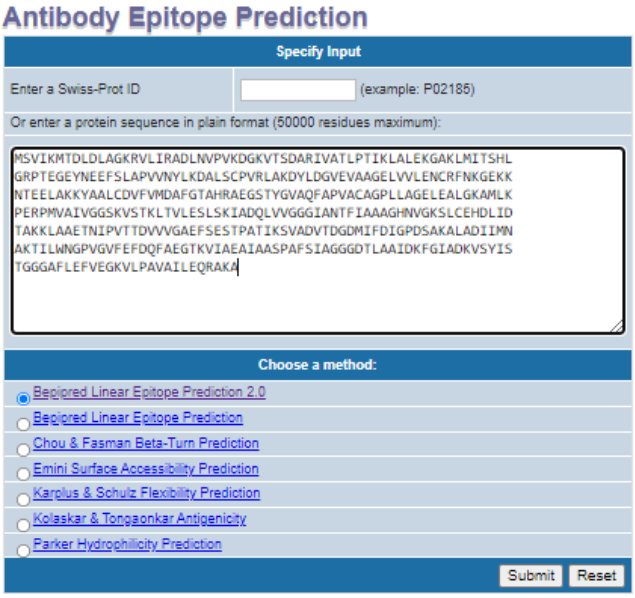
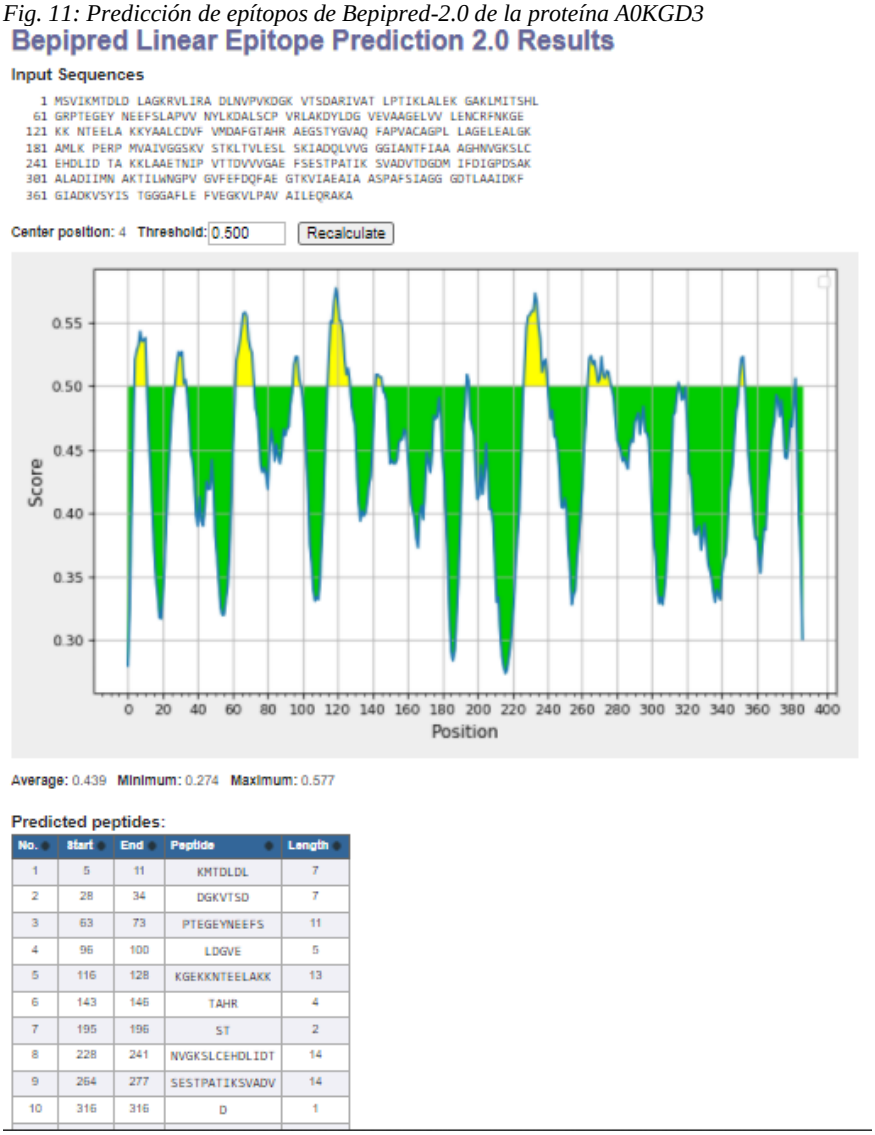


Fig. 11: Predicción de epítopos de Bepipred-2.0 de la proteína A0KGD3



No.	Start	End	Peptide	Length
1	5	11	KMTDLDL	7
2	28	34	DGKVTSD	7
3	63	73	PTEGEYNEEFS	11
4	96	100	LDGVE	5
5	116	128	KGEKKNTEELAKK	13
6	143	146	TAHR	4
7	195	196	ST	2
8	228	241	NVGKSLCEHDLIDT	14
9	264	277	SESTPATIKSVADV	14
10	316	316	D	1

De igual manera que con *Ellipro*, todas las predicciones de epítopos han sido copiadas, y guardadas en un archivo tabulado denominado: n*ombre_de_la_proteína.tab*.

Fig. 12: Archivos con la predicción de epítopos lineales de la proteína con código Uniprot A0KGD3

A0KGD3.tab: Bloc de notas

Archivo Edición Formato Ver Ayuda

No.	Start	End	Peptide	Number of residues
1	5	11	KMTDLDL	7
2	28	34	DGKVTSD	7
3	63	73	PTEGEYNEEFS	11
4	96	100	LDGVE	5
5	116	128	KGEKKNTEELAKK	13
6	143	146	TAHR	4
7	195	196	ST	2
8	228	241	NVGKSLCEHDLIDT	14
9	264	277	SESTPATIKSVADV	14
10	316	316	D	1
11	351	354	GIAD	4
12	383	383	Q	1

Una vez realizada la predicción de los epítopos de todas las proteínas de la base de datos *Moon.xlsx*, los epítopos lineales han sido anotados en una base de datos conjunta denominada *Epítopos.xlsx*.

Este proceso es más complejo que los anteriores, ya que en las anotaciones de cada epítopo se ha incluido toda la información de la proteína presente en la base de datos *Moon.xlsx*, por lo que se han tenido que copiar los datos de la proteína de origen en cada anotación de epítopo de forma manual.

Por otra parte, el hecho que los archivos de Ellipro tengan dos columnas de datos más, colocadas en un orden diferente a los archivos generados con *Bepipred 2.0*, ha obligado a reordenar los datos en la propia base de datos *Epítopos.xlsx* cada vez que se incluían epítopos generados con Ellipro.

Fig 13: Base de datos Epítopos.xlsx

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Organisme	Proteïna	Patogen*	UniProt	Moon	Review	Estructura	PDB	Chain	Nº	Start	End	Peptide	Nº residuos	Score
2	*Lactobacillus casei*	Class A sortase	C	A0A0K1MVR4	S	N	N			1	4	17	SNKKQPPRKHRWLW	14	
3	*Lactobacillus casei*	Class A sortase	C	A0A0K1MVR4	S	N	N			2	45	90	LSQWTMAQVDRDNVNKNEKRKATFDFKGVKALDINTVGNAALNRDL	46	
4	*Lactobacillus casei*	Class A sortase	C	A0A0K1MVR4	S	N	N			3	101	103	VGL	3	
5	*Lactobacillus casei*	Class A sortase	C	A0A0K1MVR4	S	N	N			4	105	105	L	1	
6	*Lactobacillus casei*	Class A sortase	C	A0A0K1MVR4	S	N	N			5	108	117	FKGLSNENLS	10	
7	*Lactobacillus casei*	Class A sortase	C	A0A0K1MVR4	S	N	N			6	125	131	ADQKMGE	7	
8	*Lactobacillus casei*	Class A sortase	C	A0A0K1MVR4	S	N	N			7	141	149	MTNQGILFS	9	
9	*Lactobacillus casei*	Class A sortase	C	A0A0K1MVR4	S	N	N			8	152	155	KNVK	4	
10	*Lactobacillus casei*	Class A sortase	C	A0A0K1MVR4	S	N	N			9	178	192	YVNETQVQWIDDVAG	15	
11	*Lactobacillus casei*	Class A sortase	C	A0A0K1MVR4	S	N	N			10	204	209	PTEGEV	6	
12	*Lactobacillus casei*	Class A sortase	C	A0A0K1MVR4	S	N	N			11	219	229	QSVKKATKQHL	11	
13	*Mycobacterium smegmatis*	60 kDa chaperonin 1	C	A0QQU5	S	S	N	3rtk_A	A	1	5	11	IAYDEEA	7	
14	*Mycobacterium smegmatis*	61 kDa chaperonin 1	C	A0QQU5	S	S	N	3rtk_A	A	2	30	32	LGP	3	
15	*Mycobacterium smegmatis*	62 kDa chaperonin 1	C	A0QQU5	S	S	N	3rtk_A	A	3	42	48	KWGAPTI	7	
16	*Mycobacterium smegmatis*	63 kDa chaperonin 1	C	A0QQU5	S	S	N	3rtk_A	A	4	61	66	LEDPYE	6	
17	*Mycobacterium smegmatis*	64 kDa chaperonin 1	C	A0QQU5	S	S	N	3rtk_A	A	5	81	86	DDVAGD	6	
18	*Mycobacterium smegmatis*	65 kDa chaperonin 1	C	A0QQU5	S	S	N	3rtk_A	A	6	124	125	EK	2	
19	*Mycobacterium smegmatis*	66 kDa chaperonin 1	C	A0QQU5	S	S	N	3rtk_A	A	7	127	128	TE	2	
20	*Mycobacterium smegmatis*	67 kDa chaperonin 1	C	A0QQU5	S	S	N	3rtk_A	A	8	131	157	LKSAKEVETKEQIAATAGISAGDQSIG	27	
21	*Mycobacterium smegmatis*	68 kDa chaperonin 1	C	A0QQU5	S	S	N	3rtk_A	A	9	169	169	N	1	
22	*Mycobacterium smegmatis*	60 kDa chaperonin 1	C	A0QQU5	S	S	N	3rtk_A	A	10	173	185	ITVEESNTFGLQL	13	
23	*Mycobacterium smegmatis*	61 kDa chaperonin 1	C	A0QQU5	S	S	N	3rtk_A	A	11	188	188	T	1	
24	*Mycobacterium smegmatis*	62 kDa chaperonin 1	C	A0QQU5	S	S	N	3rtk_A	A	12	191	196	MRFDKG	6	
25	*Mycobacterium smegmatis*	63 kDa chaperonin 1	C	A0QQU5	S	S	N	3rtk_A	A	13	206	214	AERQEAVLE	9	
26	*Mycobacterium smegmatis*	64 kDa chaperonin 1	C	A0QQU5	S	S	N	3rtk_A	A	14	225	245	VSTVKDLLPLLEKVIQSGKPL	21	
27	*Mycobacterium smegmatis*	65 kDa chaperonin 1	C	A0QQU5	S	S	N	3rtk_A	A	15	249	257	AEDVEGEAL	9	
28	*Mycobacterium smegmatis*	60 kDa chaperonin 1	C	A0QQU5	S	S	N	3rtk_A	A	16	278	287	GFGDRRKAML	10	
29	*Mycobacterium smegmatis*	61 kDa chaperonin 1	C	A0QQU5	S	S	N	3rtk_A	A	17	299	313	ISEEYGLSLETADVS	15	
30	*Mycobacterium smegmatis*	62 kDa chaperonin 1	C	A0QQU5	S	S	N	3rtk_A	A	18	315	315	L	1	
31	*Mycobacterium smegmatis*	63 kDa chaperonin 1	C	A0QQU5	S	S	N	3rtk_A	A	19	323	342	VTKDETTIVEGAGDAEAIQG	20	
32	*Mycobacterium smegmatis*	64 kDa chaperonin 1	C	A0QQU5	S	S	N	3rtk_A	A	20	351	371	IENSDSDYDREKLQERLAKLA	21	
33	*Mycobacterium smegmatis*	65 kDa chaperonin 1	C	A0QQU5	S	S	N	3rtk_A	A	21	384	392	EVELKERKH	9	
34	*Mycobacterium smegmatis*	66 kDa chaperonin 1	C	A0QQU5	S	S	N	3rtk_A	A	22	426	433	ELSLTGDE	8	
35	*Mycobacterium smegmatis*	67 kDa chaperonin 1	C	A0QQU5	S	S	N	3rtk_A	A	23	481	486	EYEDLL	6	
36	*Mycobacterium smegmatis*	68 kDa chaperonin 1	C	A0QQU5	S	S	N	3rtk_A	A	24	520	537	DKPEKAAAPAGDPTGGMG	18	
37	Bifidobacterium bifidum	Glutamine Synthetase	C	A0A0H2Q1X0	S	N	N	1eae0_A	A	1	5	23	APLDLSIHKSQGLYDPSSE	19	
38	Bifidobacterium bifidum	Glutamine Synthetase	C	A0A0H2Q1X0	S	N	N	1eae0_A	A	2	38	40	RPE	3	
39	Bifidobacterium bifidum	Glutamine Synthetase	C	A0A0H2Q1X0	S	N	N	1eae0_A	A	3	59	67	AVGAEENTG	9	
40	Bifidobacterium bifidum	Glutamine Synthetase	C	A0A0H2Q1X0	S	N	N	1eae0_A	A	4	86	92	GVELPEA	7	
41	Bifidobacterium bifidum	Glutamine Synthetase	C	A0A0H2Q1X0	S	N	N	1eae0_A	A	5	105	113	DIATSGQQE	9	
42	Bifidobacterium bifidum	Glutamine Synthetase	C	A0A0H2Q1X0	S	N	N	1eae0_A	A	6	132	139	VVPTNPDG	8	
43	Bifidobacterium bifidum	Glutamine Synthetase	C	A0A0H2Q1X0	S	N	N	1eae0_A	A	7	162	164	KLA	3	
44	Bifidobacterium bifidum	Glutamine Synthetase	C	A0A0H2Q1X0	S	N	N	1eae0_A	A	8	205	205	Q	1	
45	Bifidobacterium bifidum	Glutamine Synthetase	C	A0A0H2Q1X0	S	N	N	1eae0_A	A	9	207	216	TPFFPDLSDS	10	

2.2 Análisis de la Base de Datos de Epítopos

2.2.1 Estructura y Resumen de la Base de Datos de Epítopos

El primer paso del análisis estadístico de la base de datos de epítopos, *Epítopos.xlsx*, ha sido su pre-procesamiento, de forma que esta tenga un formato adecuado para poder realizar los análisis posteriores.

El pre-procesamiento ha consistido en trasformar la variable *Score* en variable numérica y el resto de variables, excepto *Nº*, *Start*, *End* y *Nº residuos* en factores.

La base de datos se estructura en **3719 observaciones**, una por cada epítopo detectado, compuesta cada una de ellas de **15 variables**. Las variables incluidas en cada una de las observaciones son:

- Organisme: Variable factorial que indica el organismo al cual pertenece el epítopo.
- Proteína: Variable factorial que indica la proteína a la cual pertenece el epítopo.
- Patogen*: Variable factorial que indica el tipo de patogenia de la proteína a la que pertenece el epítopo. La variable cuenta con cuatro factores representados de forma simplificada:
  P = Organismo Patógeno
  N = Organismo No Patógeno
  C = Organismo Comensal/Simbionte
  O = Organismo Patógeno Oportunista
- Uniprot: Variable factorial que indica el código Uniprot de la proteína a la que pertenece el epítopo.
- Moon: Variable factorial que indica si la proteína a la que pertenece el epítopo es una proteína multitarea.
- Review: Variable factorial que indica si la proteína que presenta el epítopo ha sido revisada y se corresponde con el código indicado.
- Estructura: Variable factorial que indica si se conoce alguna variante estructural análoga a la proteína que presenta el epítopo.
- PDB: Variable factorial que indica el código PDB de la proteína a la que pertenece el epítopo.
- Chain: Variable factorial que indica la cadena a la que pertenece el epítopo.
- Nº: Variable numérica que enumera los epítopos de cada proteína.
- Start: Variable numérica que indica el inicio del epítopo.
- End: Variable numérica que indica el final del epítopo.
- Peptide: Variable factorial que indica el código PDB de la estructura análoga a la proteína a la que pertenece el epítopo.
- Nº residuos: Variable numérica que indica el número de residuos del epítopo.
- Score: Variable numérica que indica el índice de protrusión del epítopo.

El análisis de la estructura de la base de datos muestra que se han predicho los epítopos de **253 proteínas**, de **85 tipos proteicos diferentes**, pertenecientes a **120 organismos**, de los que se conoce el código *PDB* de **128 estructuras análogas**.

Tab 1: Sumario de 1iok_A

```
'data.frame':   3719 obs. of  15 variables
 Organisme: Factor w/ 120 levels
 Proteïna   : Factor w/ 85 levels
 Patogen*   : Factor w/ 4
 UniProt    : Factor w/ 253 levels
 Moon       : Factor w/ 2 levels
 Review     : Factor w/ 2 levels
 Estructura : Factor w/ 2 levels
 PDB        : Factor w/ 128 levels
 Chain      : Factor w/ 6 levels
 Peptide    : Factor w/ 2910

       Nº              Start                End                      Peptide
 Min.   : 1.000   Min.   :   0.0   Min.   :   5.0   E      :  69
 1st Qu.: 4.000   1st Qu.: 102.0   1st Qu.: 111.0   G      :  46
 Median : 8.000   Median : 215.0   Median : 222.0   D      :  43
 Mean   : 9.343   Mean   : 241.5   Mean   : 250.8   A      :  38
 3rd Qu.:13.000   3rd Qu.: 331.0   3rd Qu.: 343.0   K      :  38
 Max.   :49.000   Max.   :1499.0   Max.   :1513.0   N      :  27
                                                    (Other):3458

  Nº residuos             Score
 1st Qu.:  4.00   1st Qu.:0.624
 Min.   :  1.00   Min.   :0.501
 Median :  7.00   Median :0.692
 3rd Qu.: 13.00   3rd Qu.:0.749
 Mean   : 10.25   Mean   :0.687
 Max.   :224.00   Max.   :0.873
                  NA's   :3417
```

Por su parte, el sumario muestra la frecuencia absoluta de los seis factores mayoritarios de las variables factoriales y los cuartiles, mínimos, máximos, medias y medianas de las variables numéricas.

El sumario proporciona una serie de resúmenes con datos significativos como:

- El organismo con mayor número de epítopos predichos es *Streptococus pneumoniae*.
- El tipo de proteínas con mayor número de epítopos predichos es el grupo de las *Enolasas*.
- La proteína con mayor número de epítopos predichos es *A0A0H2Q1X0*.
- El número medio de residuos en los epítopos es de **10.25** y la mayor parte de los mismos se sitúan entre los residuos **102** (inicio) y los residuos **343** (final).

2.2.2 Análisis General de la Base de Datos de Epítopos

2.2.2.1 Resumen de la Base de Datos de Epítopos

Para poder realizar también consultas sobre la base de datos original, se ha creado una base de datos, denominada *ListBase*. Esta base de datos imita la base de datos a partir de la cual se han seleccionado las proteínas sobre las que se ha realizado la predicción de los epítopos.

Tab 2: Sumario de ListBase

```
Organisme
 Streptococcus pneumoniae     : 17
 Mycobacterium tuberculosis   : 12
 Candida albicans             :  8
 Listeria monocytogenes       :  8
 Paracoccidioides brasiliensis:  7
 Lactobacillus plantarum      :  6
 (Other)                      :195

Proteïna                                      Patogen*       UniProt
 Enolase                                 : 42  C: 34          A0A0B7LRW6:  1
 Glyceraldehyde-3-phosphate dehydrogenase: 36  N: 19          A0A0H2Q1X0:  1
 60 kDa chaperonin                       : 22  O: 39          A0A0K1MVR4:  1
 Chaperone protein DnaK                  : 17  P:161          A0A0L0LUA6:  1
 Phosphoglycerate kinase                 : 13                 A0A0T9H7T7:  1
 Elongation factor Tu                    : 10                 A0A113S8Y8:  1
 (Other)                                 :113                 (Other)   :247

Moon           Review         Estructura          PDB
 N   : 38      N   : 95       N   :207          4a3r_A :  7
 S   :214      S   :148       S   : 36          1iok_A :  6
 NA's:  1      NA's: 10       NA's: 10         4qx6_A :  6
                                               1dkg_D :  5
                                               1efc_A :  4
                                               (Other):154
                                               NA's   : 71
```

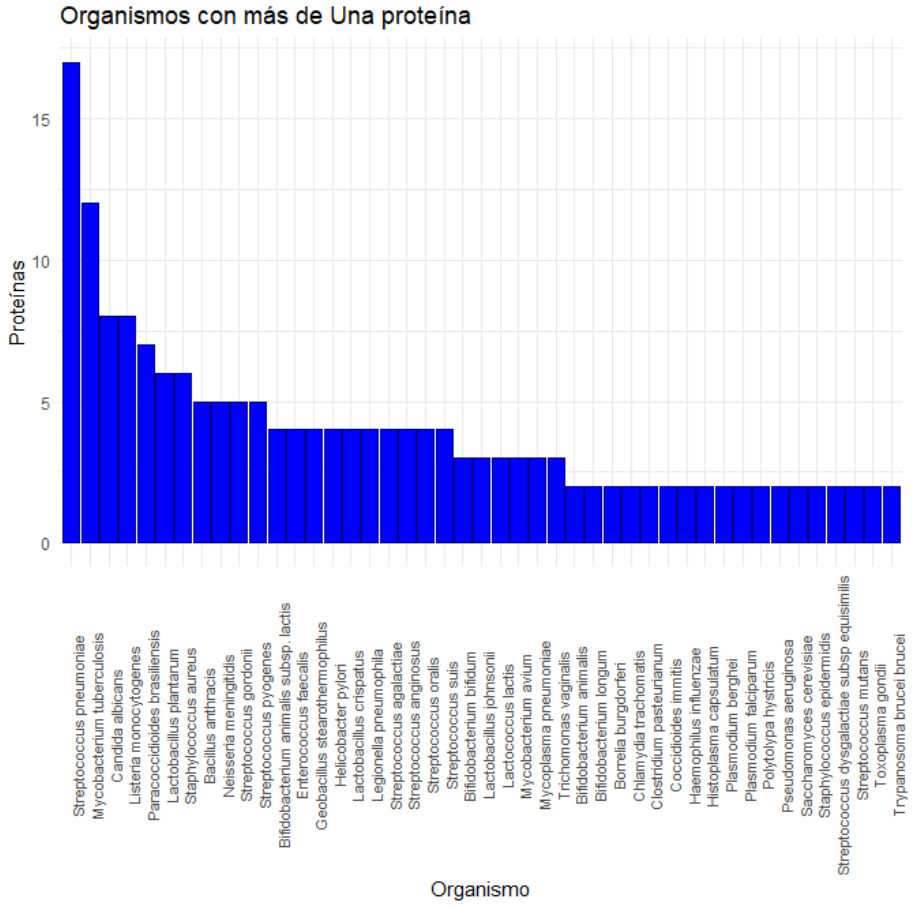
La base de datos *ListBase* también aporta datos significativos:

- El organismo con mayor número de proteínas estudiadas es *Streptococus pneumoniae*.
- El tipo de proteínas con mayor número de proteínas estudiadas es el de las *Enolasas*.
- Las estructuras con mayor número de proteínas análogas, **7**, es la correspondiente al código PDB *4a3r_A*.
- La mayor parte de proteínas estudiada son de organismos patógenos, **161**, representando la suma de organismos patógenos y oportunistas un **79.05%** del total de proteínas estudiadas.
- La mayor parte de proteínas estudiadas son multitarea, **214**, representando un **84.92%** del total de proteínas estudiadas.
- El porcentaje de proteínas revisadas es más bajo que en los casos anteriores, representando únicamente el **58.5%** del total de proteínas estudiadas.
- Finalmente, se observa que el número de proteínas con estructura conocida es bastante bajo, únicamente se conoce la estructura de **36** proteínas, el equivalente al **14.81%** del total.

2.2.2.2 Distribución de Proteínas por Organismo

A nivel de distribución de proteínas, **45** organismos cuentan con más de una proteína en la lista, **11** de los cuales presentan 5 o más proteínas en la lista, dominando la misma *Streptococus pneumoniae* con **17 proteínas** y *Mycobacterium tuberculosis* con **12**.

Fig. 14: Diagrama de barras de organismos con más de una proteína presente en Epítopos.xlsx



2.2.2.3 Distribución de Epítopos por Organismo

La distribución de epítopos por organismo se ve copada en los dos primeros lugares por los mismos organismos, Streptococus pneumoniae, con **244** epítopos y *Mycobacterium tuberculosis* con **127**, pero los siguientes puestos cambian en ambas listas, encontrándose *Streptococcus gordonii*, *Candida albicans* y *Paracoccidioides brasiliensis* en la lista de epítopos mientras que en la lista de proteínas *Candida albicans* se sitúa en la tercer posición, aparece *Listeria monocytogenes* en la cuarta y *Streptococcus gordonii* no aparece hasta la décima posición.

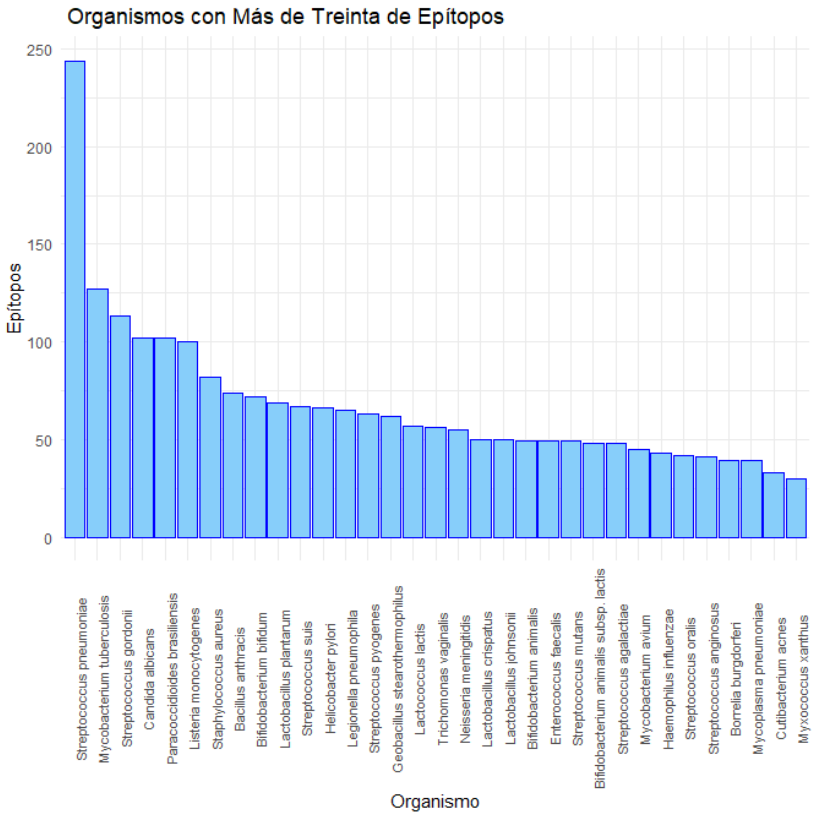
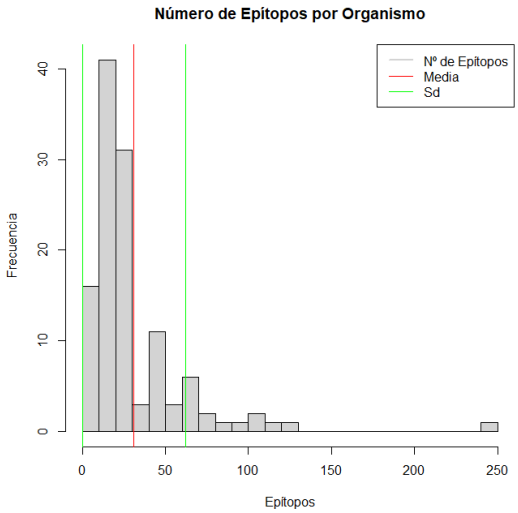


Fig. 15: Organismos con más de 30 epítopos

Por otro lado, se observa que la media de epítopos por organismo es de **30.99**, con una desviación típica de **31.08**, y que una proporción mayoritaria de los organismos tienen **entre 10 y 30 epítopos**.

Parámetros Estadísticos					
Min.	*1st Qu.*	*Median*	*Mean*	*3rd Qu.*	*Max.*
4.00	*14.00*	*22.00*	*30.99*	*39.00*	*244.00*

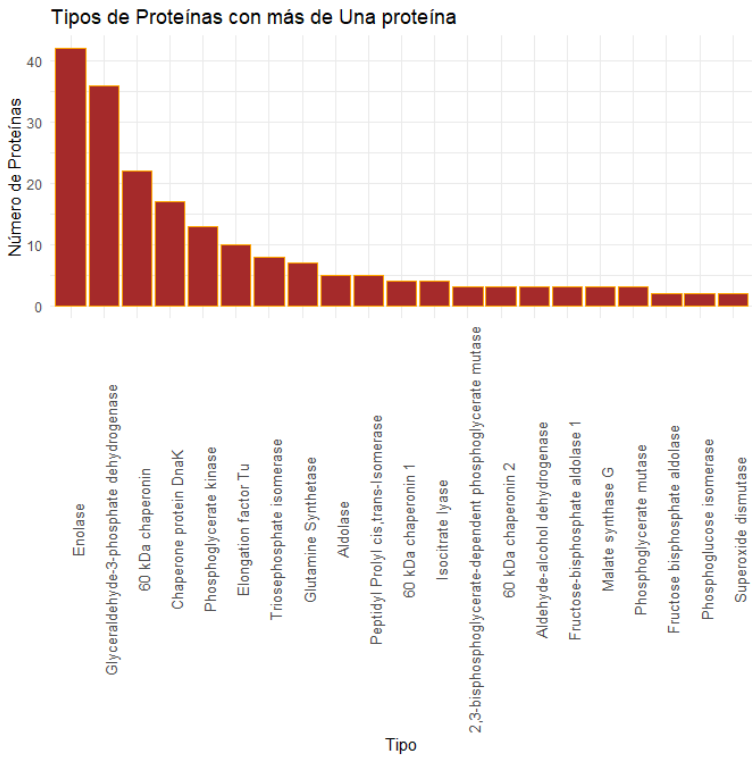
Fig. 16: Histograma de distribución del número de epítopos por organismo



2.2.2.4 Distribución de Proteínas por Tipo

A nivel de distribución de proteínas por grupo, únicamente **21** grupos cuentan con más de una proteína en la lista, concentrándose mayoritariamente en 6 grandes grupos, que presentan más de **10** proteínas cada uno, destacando especialmente el grupo de las *Enolasas*.

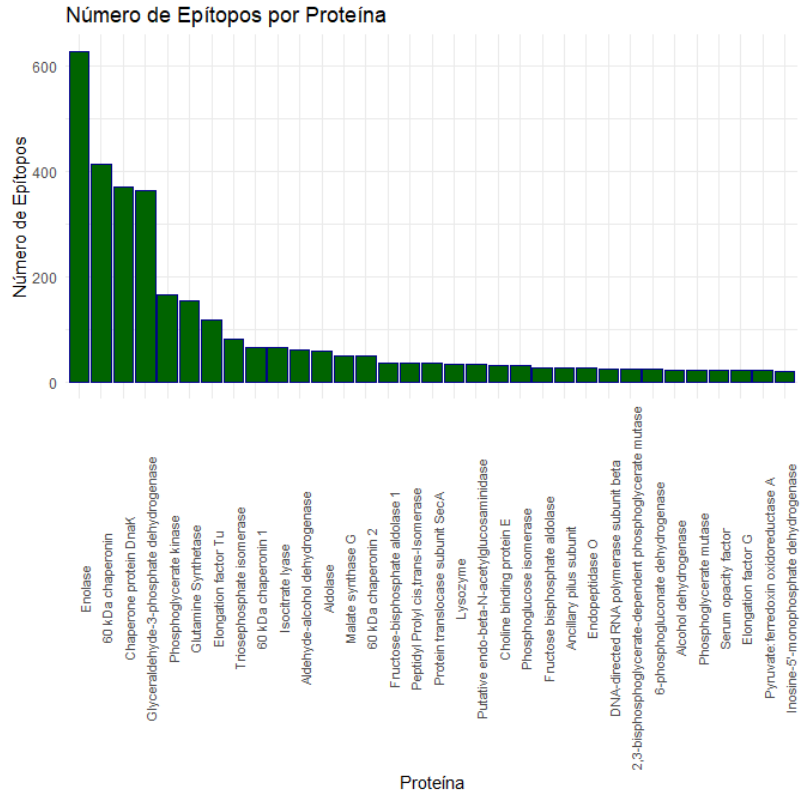
Fig. 17: Diagrama de barras de tipos de proteínas con más de una proteína presente en Epítopos.xlsx



2.2.2.5 Distribución de Epítopos por Tipo de Proteínas

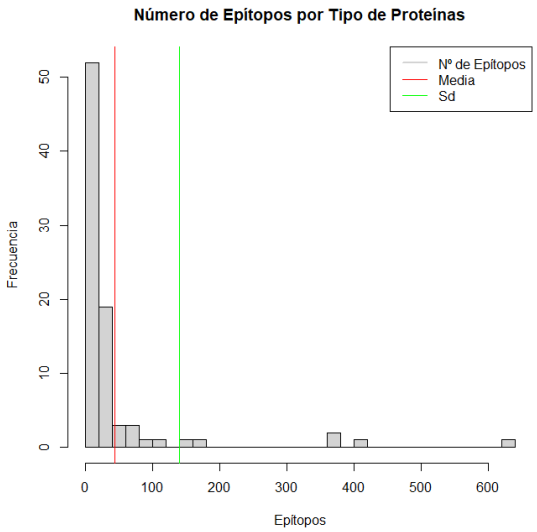
La distribución de epítopos por tipo de proteínas es similar a la proporción de proteínas de cada tipo, con la particularidad que hay una menor proporción de epítopos de *Glyceraldehyde-3-phosphate dehydrogenase*, que pasa de ser el segundo de la lista al cuarto, y que *Glutamine Synthetase* supera a *Elongation factor Tu* y se sitúa como el quinto grupo de proteínas con mayor número de epítopos.

Fig. 18: Diagrama de barras de tipos de proteínas con más de una proteína presente en Epítopos.xlsx



Aunque la media de epítopos por tipo de proteínas es de **43.75**, la mitad de los tipos de proteína tienen menos de **16 epítopos**, al estar compuestos de una única proteína. Esta diversidad en el número de epítopos por grupo queda patente en la elevada desviación típica, **97.02**, que es más del doble de la media de epítopos.

Fig. 19: Histograma de distribución del número de epítopos por tipo de proteína



2.2.3 Análisis Específico de la Base de Datos de Epítopos

Una vez obtenida una visión global de la base de datos generada, se ha realizado un análisis más específico de un grupo de proteínas, el de las proteínas análogas a la estructura *1iok_A* para, de esta forma, poder determinar sus características principales.

Hay que tener en cuenta que, antes de analizar al grupo *1iok_A*, se realizarán análisis previos de las proteínas de uno de los organismos que presenta proteínas con ese tipo de estructura, *Bacillus anthracis*, y del principal tipo de proteínas con esta estructura tridimensional, el grupo de las *60 kDa chaperonin*.

2.2.3.1 Análisis de las Proteínas de Bacillus anthracis

El organismo seleccionado para realizar el análisis específico de la base de datos es *Bacillus anthracis*, el cual fue descubierto en el siglo XIX tras determinarse como el agente causante de la enfermedad del carbunco. **[50, 51, 52]**

A principios de siglo se pudo demostrar la transmisibilidad de la enfermedad del carbunco mediante la inoculación de la sangre de un caballo muerto por carbunco a un caballo y una oveja. De esta forma, posteriormente, se pudo caracterizar el material biológico potencialmente infeccioso y comprobar que si se realizaba un filtraje de las muestras éstas perdían su poder de infección, lo que llevó a denominar la bacteria y a determinarla como la causante de la enfermedad.

Los experimentos de Koch con *Bacillus anthracis* desarrollaron el método de cultivo bacteriano y permitieron demostrar científicamente la patogenia de las bacterias y el fenómeno de esporulación. **[53]**

Bacillus anthracis es un tipo de bacteria inmóvil y capsulada de entre 1 y 6 μm de diámetro. Forma esporas muy resistentes a la temperatura y a los desinfectantes químicos, pero bastante sensibles a la penicilina, que se suelen encontrar en productos derivados de animales.
Estas esporas adoptan su forma vegetativa en medios favorables como la sangre y los tejidos biológicos siempre que éstos se encuentren en condiciones aerobias. **[54]**

El carbunco se desarrolla cuando las esporas penetran en los organismos por vías sanguínea, oral o respiratoria, y germinan en los ganglios linfáticos hasta alcanzar el torrente sanguíneo. El carbunco afecta tanto a humanos como a animales y, en función de la vía de transmisión, se clasifica en los siguientes tipos de enfermedad:

- Carbunco cutáneo.
- Carbunco pulmonar.
- Carbunco digestivo.

Bacillus anthracis presenta **89 cepas conocidas** con diferentes grados de virulencia, desde cepas altamente virulentas, utilizadas en la guerra biológica, a cepas con muy bajo grado de virulencia. El nivel de virulencia viene determinado por la presencia y actividad de varios genes que regulan la producción de antígenos y de toxinas.

En la base de datos existen cinco proteínas de *Bacillus anthracis*, (*C3LDI0, Q81UL6, Q81VE1, Q81X74, y Q81X75*) dos de las cuales son del tipo *Glyceraldehyde-3-phosphate dehydrogenase* y una de ellas es del tipo *60 kDa chaperonin*, que es la que se ha seleccionado como referencia para análisis posteriores.

Tab 3: Proteínas de *Bacillus anthracis* con estructura análoga a 1iok_A

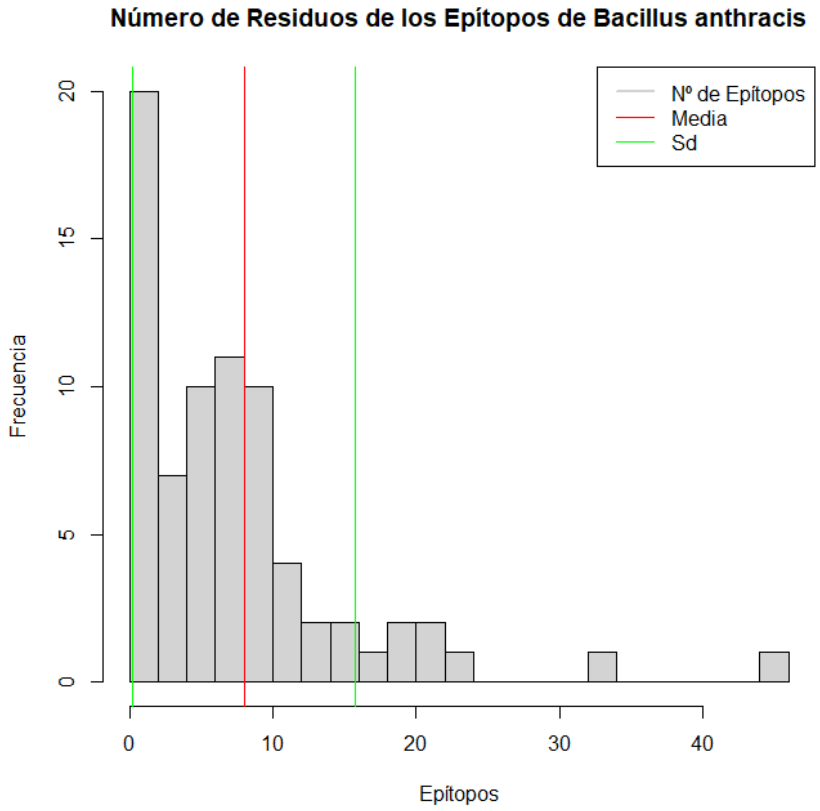
	Proteïna	UniProt	PDB
728	Enolase	C3LDI0	4a3r_A
2975	Glyceraldehyde-3-phosphate dehydrogenase	Q81UL6	1euh_A
2992	60 kDa chaperonin	Q81VE1	<NA>
3010	Glyceraldehyde-3-phosphate dehydrogenase	Q81X74	1gd1_O
3020	Phosphoglycerate kinase	Q81X75	<NA>

Número de residuos y distribución de los epítopos de las proteínas de *Bacillus anthracis*

El número medio de residuos en los epítopos de *Bacillus anthracis* es de **8**, con una desviación típica de **7,77**, por lo que se puede decir aproximadamente el 75% de los epítopos tiene menos de 16 residuos.

Parámetros Estadísticos					
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.0	2.0	6.5	8.0	9.0	46.0

Fig. 20: Histograma de distribución del número de residuos de los epítopos de Bacillus anthracis



Si se observa la distribución de epítopos de las diferentes proteínas de *Bacillus anthracis*, se encuentran diferencias importantes en la distribución de los epítopos de todas las proteínas, incluidas las dos pertenecientes al grupo de *Glyceraldehyde-3-phosphate dehydrogenase*, algo lógico teniendo en cuenta que cada una de las proteínas analizadas tienen funciones diferentes y, por tanto, estructuras diferentes.

Cabe destacar que la proteína correspondiente al grupo de las *60 kDa chaperonin* muestra una distribución más amplia de sus epítopos y un nº de residuos mayor que el resto de proteínas observadas.

Fig. 21: Distribución del inicio y final de los epítopos de las proteínas de Bacillus anthracis

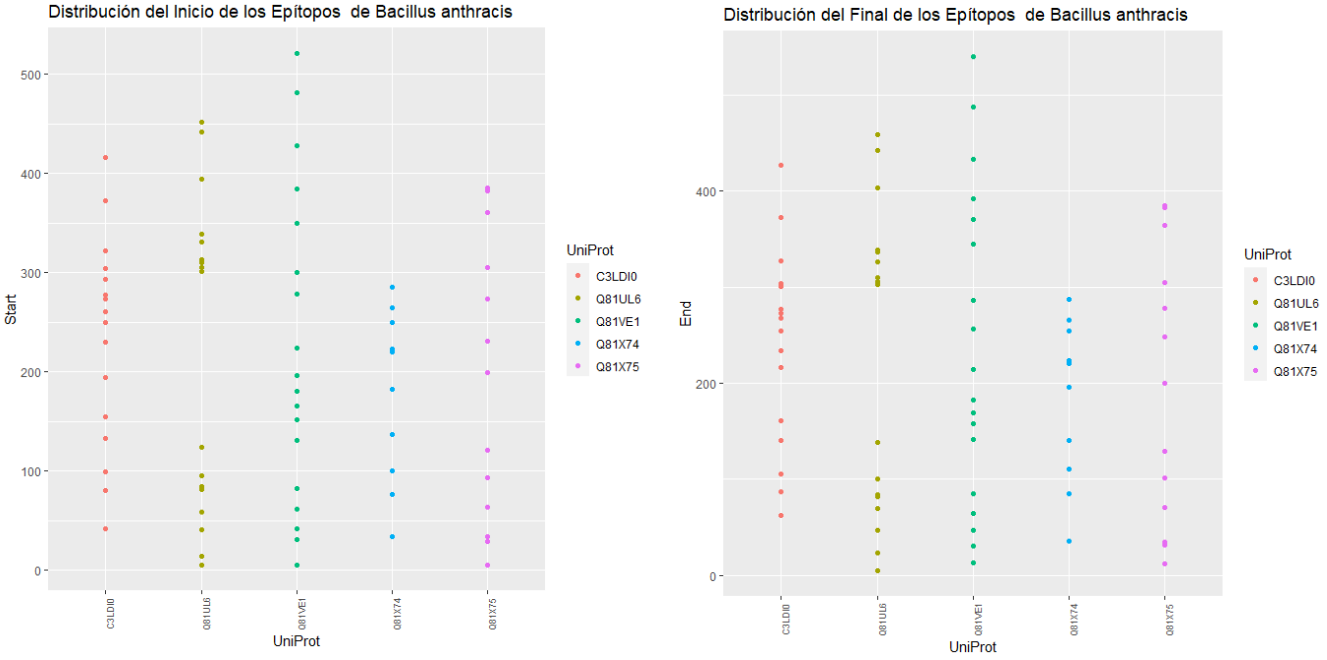
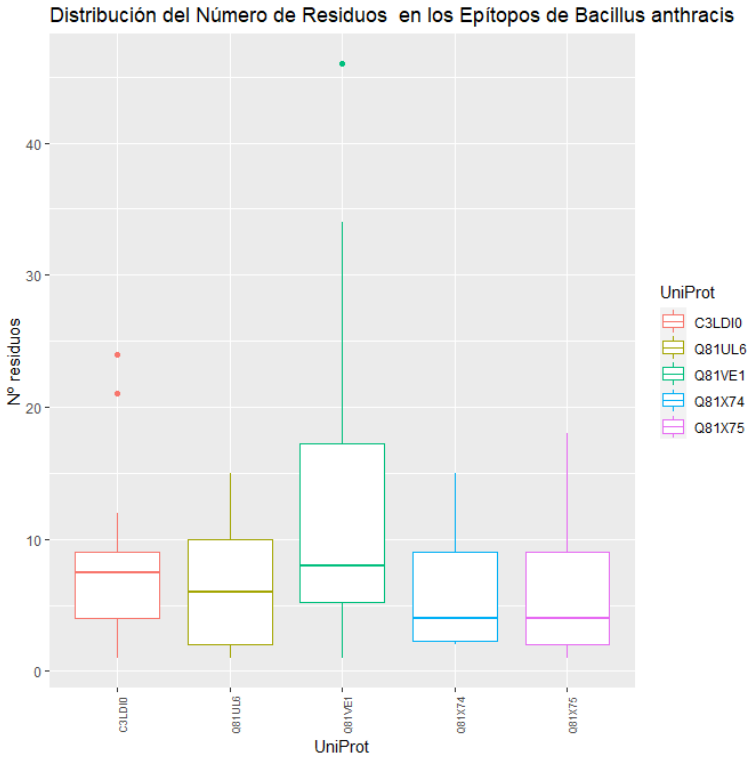


Fig. 21: Distribución del número residuos de los epítopos de las proteínas de Bacillus anthracis



2.2.3.2 Análisis de las Proteínas de Tipo 60 kDa chaperonin

El segundo objetivo primario del TFM es analizar algún tipo de proteína presente en la base de datos, buscar mimótopos con los que poder realizar comparaciones con estructuras análogas en el ser humano y buscar posibles relaciones entre estos mimótopos y enfermedades humanas.

Se ha decidido realizar el análisis comparativo de las proteínas del grupo de las *60 kDa chaperonin* ya que presentan un par de estructuras proteicas bien definidas y, aunque es

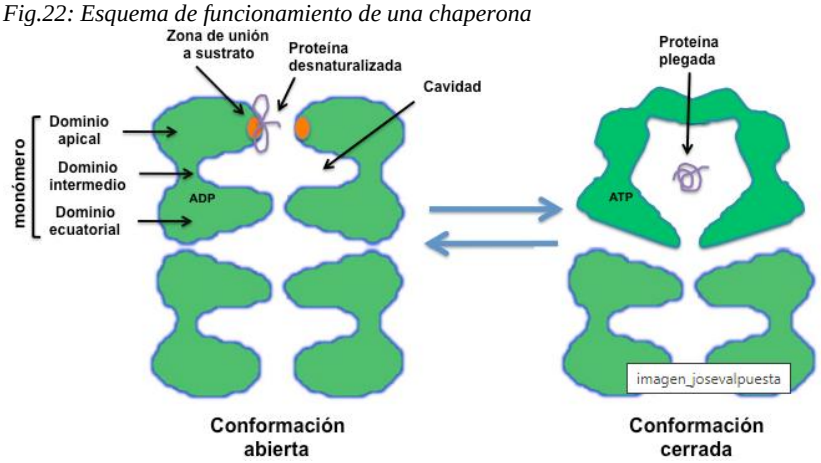
un grupo bien representado en la base de datos, tiene un número de elementos manejable.

Las chaperonas son un grupo abundante de proteínas, tanto en procariotas como en eucariotas, que presentan como función principal el favorecimiento del plegamiento de:

- Proteínas nuevas.
- Proteínas que han sufrido un proceso de desnaturalización.
- Proteínas translocadas a algún nuevo compartimento celular.

Las chaperonas no forman parte de la estructura primaria de la proteína funcional a la que modifican, sino que actúan uniéndose a la superficie hidrofóbica de la proteína evitando un plegamiento aberrante a la vez que les confieren la suficiente estabilidad para evitar su agregación con otras proteínas y favorecen que puedan adquirir su conformación funcional correcta. **[55,56]**

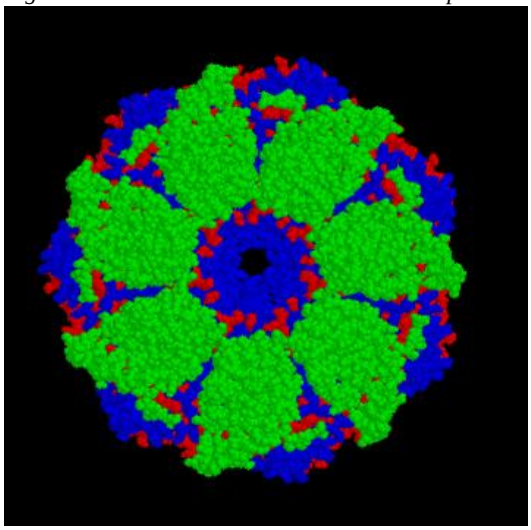
Fig.22: Esquema de funcionamiento de una chaperona



Varias chaperonas pueden colaborar simultáneamente en los cambios de conformación tridimensional de las proteínas, en función de la complejidad de la estructura a plegar y de la disponibilidad de las chaperonas.

Las chaperonas son oligómeros con una estructura común cilíndrica compuestas por uno o dos anillos, dispuestos paralelamente, en los que cada anillo presenta una cavidad dónde se realiza el plegado de las proteínas. **[57]**

Fig. 23: Estructura tridimensional de una chaperona



La proteína *Q81VE1* es un ejemplo de *chaperona de 60 kDa* presente en la base de datos, pero no es la única, por lo que se ha procedido a detectar el resto de proteínas de este tipo y crear una nueva base de datos, *List60kDa*, únicamente con las observaciones de este grupo de proteínas.

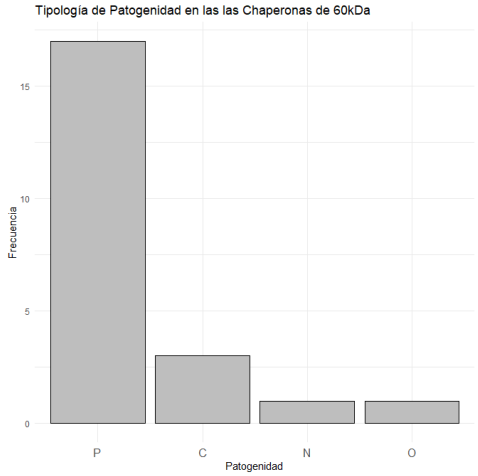
Tab 4: Sumario de *List60kDa*

```
Organisme
 Aggregatibacter actinomycetemcomitans serotype c: 1
 Bacillus anthracis                               : 1
 Bartonella baciliformis                          : 1
 Borrelia burgdorferi                             : 1
 Brucella abortus                                 : 1
 Chlamydia pneumoniae                             : 1
 (Other)                                          :16

UniProt         Moon    Review   Estructura      PDB             Patogen*
 A4IWC5 : 1     N: 2    N: 3     N:22            1iok_A :5       C: 3
 E7C160 : 1     S:20    S:19     S: 0            4v4o_A :3       N: 1
 P0A1D3 : 1                                      1aon_A :2       O: 1
 P0C923 : 1                                      2eu1_A :2       P:17
 P0CB35 : 1                                      4pkn_A :2
 P31294 : 1                                      (Other):4
 (Other):16                                      NA's   :4
```

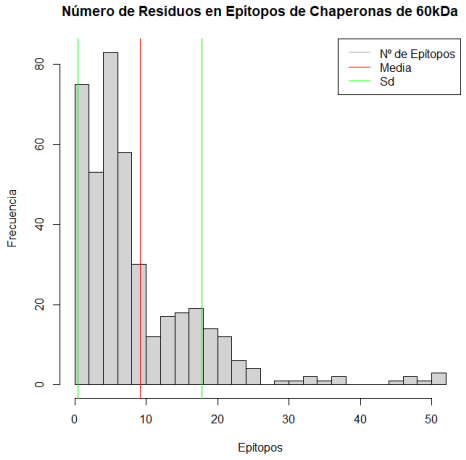
A partir de *List60kDa*, se ha realizado un análisis de patogenia de las proteínas del grupo, en el que se observa que un **77,27%** de las mismas son de organismos patógenos.

Fig. 24: Tipología de patogenia en las chaperonas de 60 kDa



La mayor parte de los epítopos de las *chaperonas de 60kDa* presentan entre **1 y 10 epítopos**, aunque existe un pequeño número de proteínas que tiene entre **45 y 52** epítopos.

Fig. 25: Distribución del número de residuos en los epítopos de las chaperonas de 60 kDa



El análisis gráfico de la localización y tamaño de los epítopos del grupo de proteínas *60 kDa chaperonin* permite visualizar una importante diferencia entre cuatro proteínas y el resto de componentes del grupo.

Mientras que la mayoría que presentan una distribución y número de epítopos relativamente similar, con un número de residuos medio de **entre 5 y 10 epítopos**, con la excepción de *P46398*, y una distribución de los epítopos entre **el primer residuo y el 535**, aproximadamente, las cuatro proteínas presentan una distribución de los epítopos bastante diferente al resto.

Las proteínas que presentan diferencias significativas respecto al resto de proteínas del grupo son:

- A4IWC5: Tiene un pequeño número de epítopos que se localizan entre los residuos 0 y 70 y presentan un número máximo de residuos inferior a 10.
- Q8KP52: Tiene todos los epítopos localizados en los primeros 200 residuos y presentan un número máximo de residuos inferior a 10.
- E7C160: Tiene todos los epítopos localizados entre los residuos 50 y 150 y una media de residuos ligeramente superior a 10.
- Q48900: Tiene los epítopos situados en los primeros 200 residuos y un número máximo de residuos inferior a 10.

Cabe destacar que con el análisis específico se ha descubierto la gran utilidad de los diagramas de puntos y de cajas a la hora de comparar la distribución de los epítopos en los mismos grupos, ya que no sólo permiten visualizar claramente las similitudes y diferencias entre las diferentes proteínas, sino que incluso permite detectar errores y incoherencias en los datos.

Gracias a la observación de los diagramas y la posterior revisión de la base de datos de origen se han detectado la duplicación de los epítopos de dos proteínas diferentes, P31294 y P31681, detectada al observarse una distribución idéntica de ambas en el diagrama de cajas del nº de residuos de los epítopos.

Fig. 26: Diagrama de puntos de la distribución del número de residuos en los epítopos de las chaperonas de 60 kDa

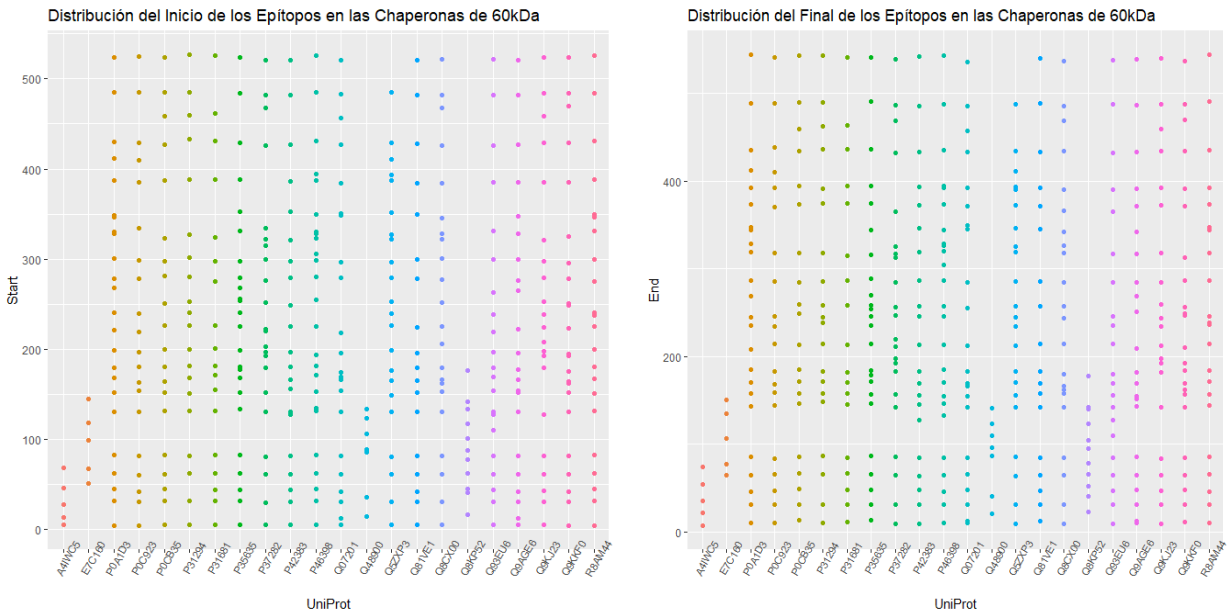
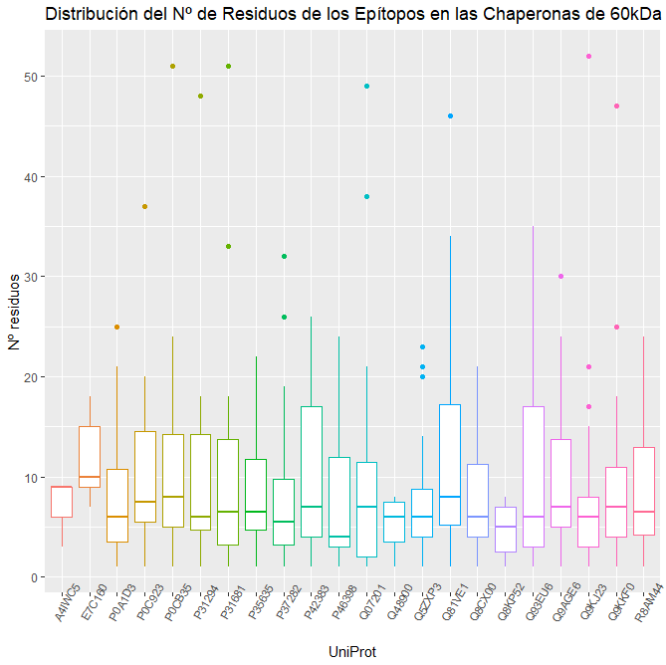
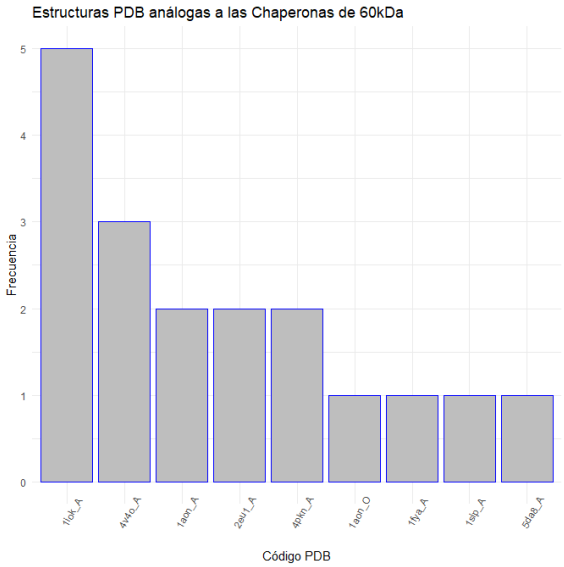


Fig. 27: Boxplot de distribución del número de residuos en los epítopos de las chaperonas de 60 kDa



También se han identificado los códigos *PDB* de las proteínas con estructuras análogas al grupo, obteniéndose **9** estructuras análogas, siendo *1iok_A* la estructura mayoritaria al compartir analogía con **5** proteínas de la base de datos.

Fig. 28: Distribución de las estructuras PDB análogas a las proteínas del grupo de las chaperonas de 60 kDa



Los análisis realizados han dejado patente la importante similitud en la distribución de los epítopos y el número de residuos por epítopo de un grupo de 17 proteínas del tipo *60 kDa chaperonin* y la casi idéntica distribución y tamaño de los epítopos de las proteínas con estructura análoga a *1iok_A*.

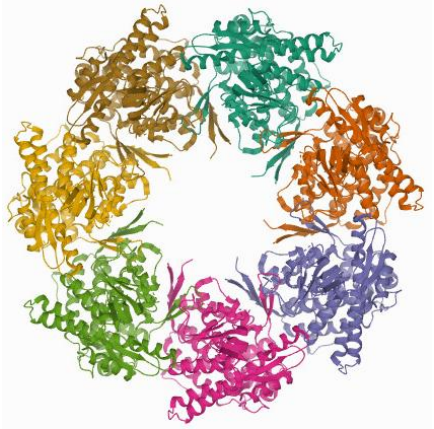
Estas similitudes hacen creer que puede ser factible identificar mimótopos en este grupo de proteínas para poder compararlos con proteínas humanas, por lo que ambos tipos serán los primeros candidatos para realizar la identificación de mimotopos.

3. Análisis de las Chaperonas de 60 kDa con Estructura Análoga a 1iok_A

3.1 Análisis General de las Chaperonas de 60 kDa con Estructura Análoga a 1iok_A

De todas las proteínas de la base de datos con estructura análoga a *1iok*, **5** de ellas forman parte del grupo de *60 kDa chaperonin*, **6** si se incluye *Q1D3Y5*, que es una chaperona de tipo 1.

Fig. 29: Estructura de la 60 kDa chaperonin de Paracoccus denitrificans.



Para poder analizar en detalle las *Chaperonas de 60 kDa con estructura análoga a 1iok* se ha procedido a crear una nueva base de datos específica.

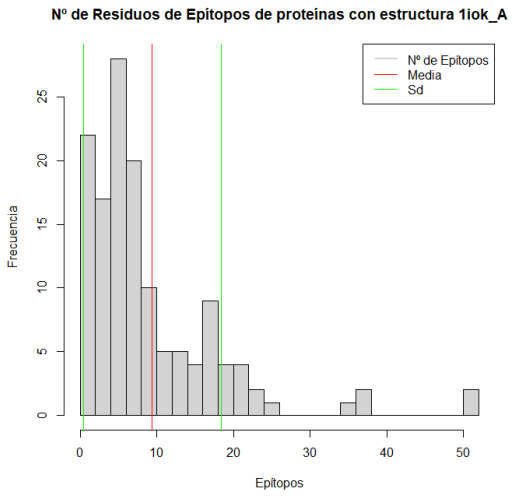
La base de datos está compuesta de las siguientes proteínas:

Tab 5: Tabla de chaperonas de 60 kDa con estructura análoga a 1iok_A

	Organisme	Proteïna	UniProt
1587	Borrelia burgdorferi	60 kDa chaperonin	P0C923
1607	Brucella abortus	60 kDa chaperonin	P0CB35
1822	Bartonella baciliformis	60 kDa chaperonin	P35635
2482	Myxococcus xanthus	60 kDa chaperonin 1	Q1D3Y5
3330	Enterococcus faecalis	60 kDa chaperonin	Q93EU6
3504	Lactobacillus johnsonii	60 kDa chaperonin	Q9KJ23

El grupo de *Chaperonas de 60 kDa con estructura análoga a 1iok_A* tiene una media de **9,41 residuos** por epítopo, una desviación estándar de **8,955** y está presente únicamente en **5 epítopos con más de 30 residuos**.

Fig. 30: Distribución del número de residuos en los epítopos de las Chaperonas de 60 kDa con estructura 1iok_A



Observando la localización de los epítopos de las proteínas analizadas se puede ver que todas presentan una distribución y rango muy similares, existiendo pequeñas diferencias en la presencia o localización de sólo alguno de los epítopos.

En la distribución del número de residuos por epítopo se observan diferencias algo más significativas pero un detalle muy importante es que, pese a tener en algunos casos rangos del número de residuos importante, todas las proteínas presentan una media muy similar.

De igual manera que en los diagramas de puntos representados anteriormente, se ha detectado un nuevo error en la base de datos que se ha podido corregir gracias a la observación de los gráficos.

En este caso se trataba de una mala anotación del código PDB en la proteína POCY35, que estaba incluida erróneamente en el grupo de proteínas *1iok_A*, y que fue detectada al observarse diferencias significativas en la distribución de los epítopos entre ésta y el resto de proteínas del grupo.

Fig. 31: Distribución del número de residuos en los epítopos de las Chaperonas de 60 kDa con estructura a 1iok_A

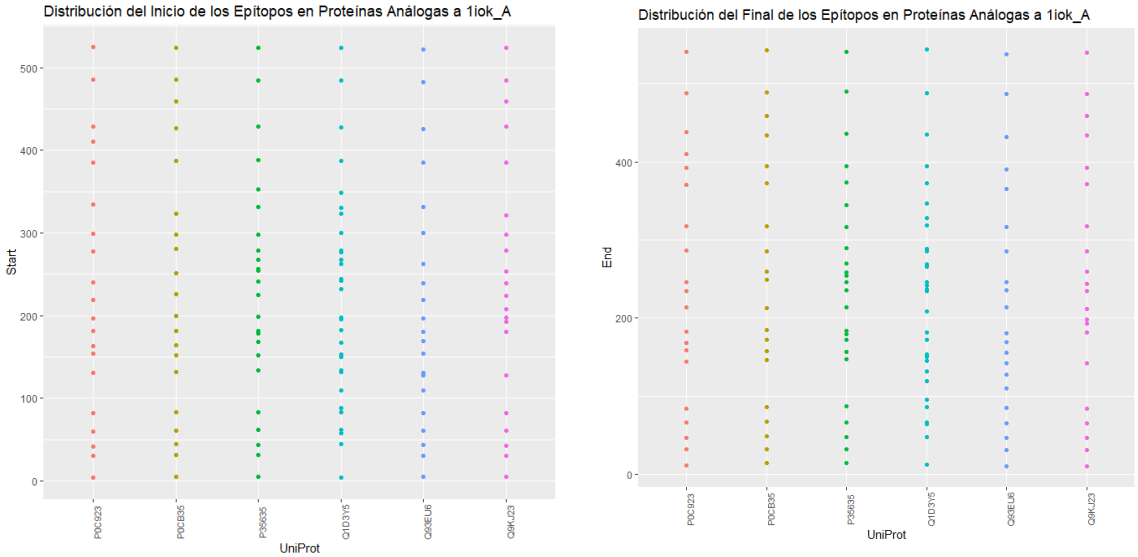
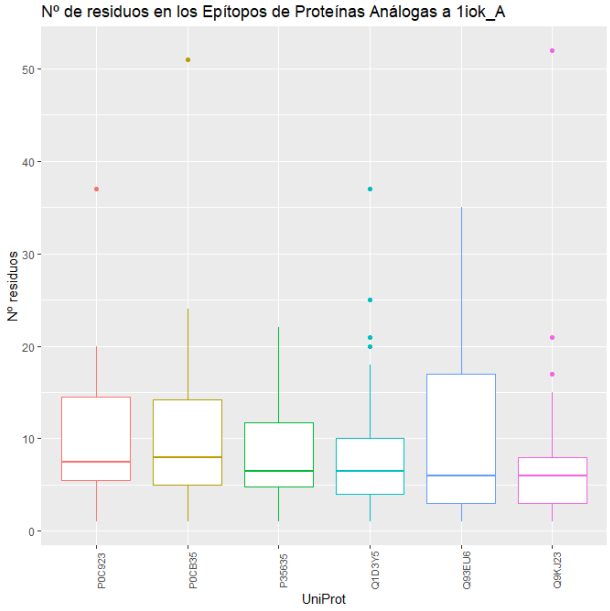


Fig. 32: Distribución del nº de residuos en los epítopos de Chaperonas de 60 kDa con estructura análoga a 1iok_A



3.2 Obtención de los Mimotopos de las Chaperonas de 60 kDa con Estructura Análoga a 1iok_A

3.2.1 Alineación de secuencias de las Chaperonas de 60 kDa con Estructura Análoga a 1iok_A

El primer paso para la obtención de los mimotopos de las chaperonas de 60 kDa con estructura análoga a 1iok_A ha consistido en la alineación de las secuencias proteicas con Uniprot Align por medio de Clustal Omega (https://www.uniprot.org/align/) y la selección de los epítopos coincidentes (en localización).

Para poder realizar la alineación de secuencias ha sido necesario introducir todas las secuencias en la ventana de selección y activar la consulta. Las seis secuencias conjuntas han sido guardadas en un archivo denominado *1iok_A FASTA.txt*.

Fig. 33: Pantalla de selección de la alineación de secuencias realizada con Uniprot Align

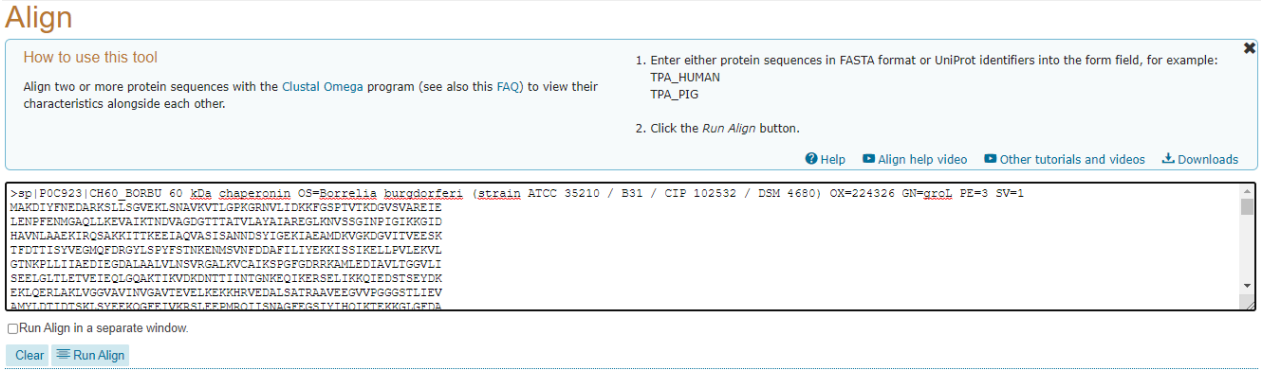
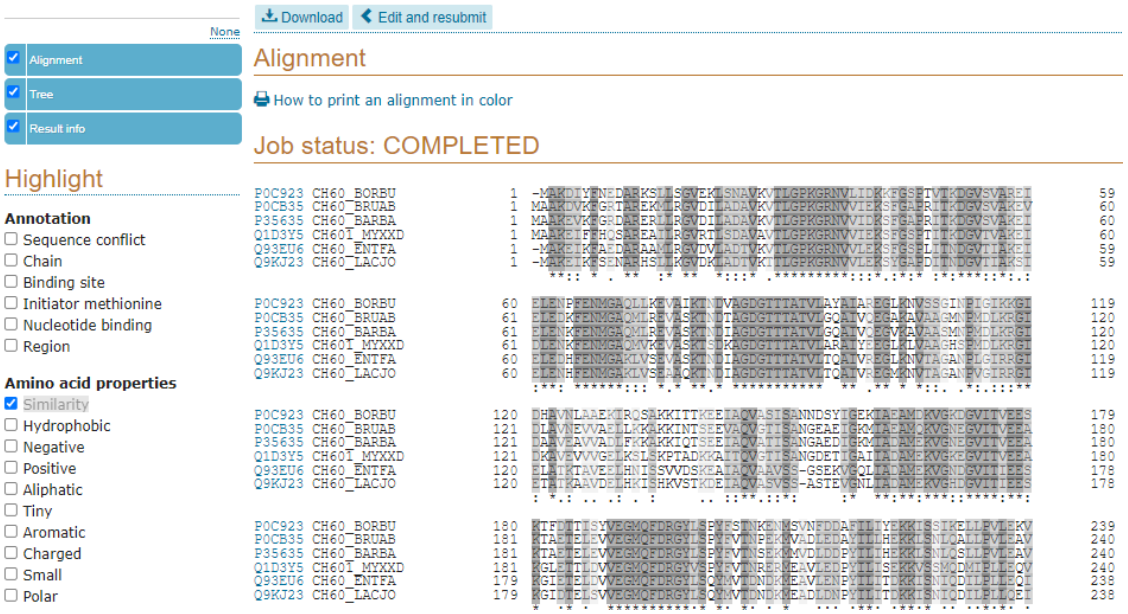


Fig. 34: Captura de pantalla de la alineación de secuencias generada con Uniprot Align



La alineación de secuencias con *Uniprot Align* ha dado como resultado un **39.312% de** identidad, con **176 posiciones similares (un 32.23% más)**, lo que muestra lo parecidas que son en estructura las seis proteínas. Los resultados de la alineación se ha guardado en un archivo denominado *align 1iok-A.aln*.

Fig. 35: Información de los resultados de la alineación de secuencias generada con Uniprot Align

Date of job execution	2021-11-08
Job identifier	A20211108A084FC58F6BBA219896F365D15F2EB4400B6DB7 (jobs are stored for 7 days)
Running time	14.1 seconds
Identical positions	217
Identity	39.312%
Similar positions	176
Program	CLUSTALO

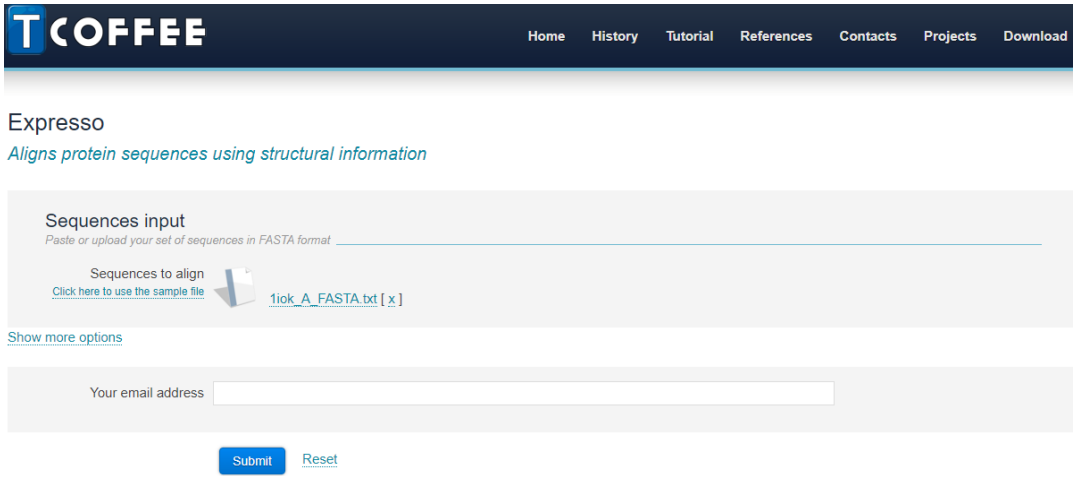
BLAST Align Download

Entry	Entry name	Protein names	Organism	Gene name
P0C923	CH60_BORBU	**60 kDa chaperonin**	Borrelia burgdorferi (strain ATCC 35210 / B31 / CIP 102532 / DSM 4680)	**groL** groEL, mopA, BB_0649
P0CB35	CH60_BRUAB	**60 kDa chaperonin**	Brucella abortus biovar 1 (strain 9-941)	**groL** groEL, mopA, BruAb2_0190
P35635	CH60_BARBA	**60 kDa chaperonin**	Bartonella bacilliformis	**groL** 7B2, Bb63, Bb65, groEL, mopA
Q1D3Y5	CH601_MYXXD	**60 kDa chaperonin 1**	Myxococcus xanthus (strain DK1622)	**groL1** groEL1, MXAN_4467
Q93EU6	CH60_ENTFA	**60 kDa chaperonin**	Enterococcus faecalis (strain ATCC 700802 / V583)	**groL** groEL, EF_2633
Q9KJ23	CH60_LACJO	**60 kDa chaperonin**	Lactobacillus johnsonii (strain CNCM I-12250 / La1 / NCC 533)	**groL** groEL, LJ_0461

Una acción complementaria al planteamiento inicial ha sido, para comprobar de forma más exacta las regiones coincidentes de los epítopos, una segunda alineación de secuencias con *T-Coffee Expresso (http://tcoffee.crg.cat/apps/tcoffee/do:expresso)*.

En este caso, para poder realizar la alineación de secuencias ha sido necesario incluir todas las secuencias en un único archivo, que se ha denominado *1iok_A FASTA.txt*.

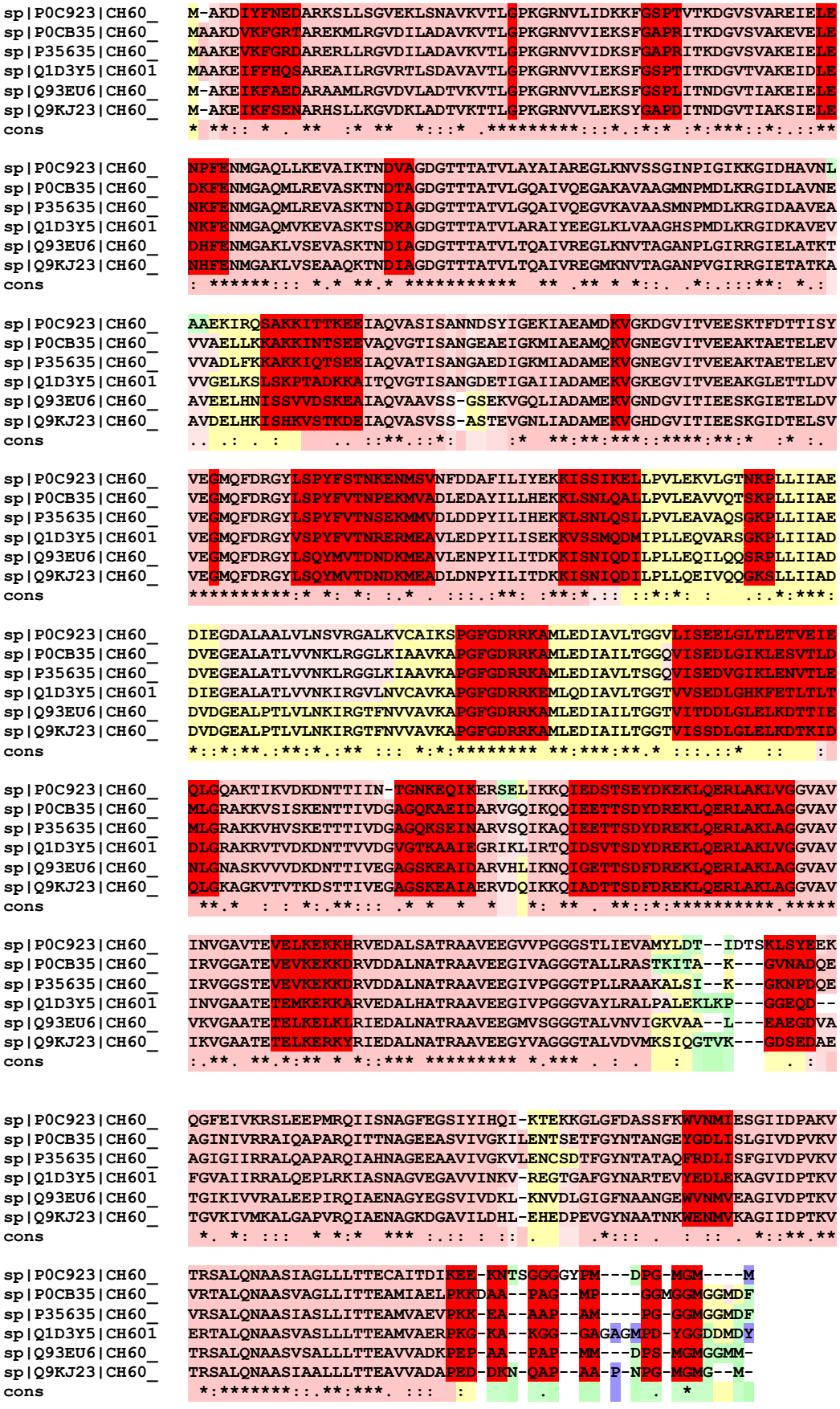
Fig. 36: Pantalla de selección de la alineación de secuencias realizada con T-Coffee



La alineación con *T-Coffee* ha dado resultados muy similares a *Align*, con un **Score de 89**, lo que ha permitido detectar las regiones peptídicas comunes sobre las que se pretende obtener los mimotopos. La secuencia obtenida con *T-Coffee*, con las **regiones comunes coloreadas en rojo**, ha sido la siguiente:

```
T-COFFEE, Version_11.00 (Version_11.00)
Cedric Notredame
CPU TIME:0 sec.
SCORE=89
 BAD AVG GOOD

sp|P0C923|CH60_    :  88
sp|P0CB35|CH60_    :  90
sp|P35635|CH60_    :  90
sp|Q1D3Y5|CH601    :  88
sp|Q93EU6|CH60_    :  88
sp|Q9KJ23|CH60_    :  88
cons               :   8
```


sp|P0C923|CH60_
sp|P0CB35|CH60_
sp|P35635|CH60_
sp|Q1D3Y5|CH601
sp|Q93EU6|CH60_
sp|Q9KJ23|CH60_
cons

3.2.2 Obtención de los Epítopos Consenso de las Chaperonas de 60 kDa con Estructura Análoga a 1iok_A

Una vez detectada las regiones peptídicas comunes de los epítopos de cada proteína, se ha procedido a realizar la comparación de secuencias con Tomtom (https://meme-suite.org/meme/tools/tomtom) para, de esta manera, obtener las secuencias de epítopos consenso (mimotopos).

Tomtom no realiza comparaciones de motivos de diferente tamaño, por lo que previamente se ha debido comparar todos los epítopos de las seis proteínas y seleccionar sólo aquellas regiones en que coincidieran los epítopos de las seis (las regiones marcadas en rojo en el apartado anterior).

Con los motivos de igual tamaño y localización, éstos se han entrado en la pestaña de consulta seleccionando proteína como tipo de motivo de consulta y como motivo objetivo y *Eukaryotic Linear Motif (ELM 2018)* como base de datos de comparación de motivos.

Fig. 37: Comparación de motivos del primer mimotopo realizada con Tomtom

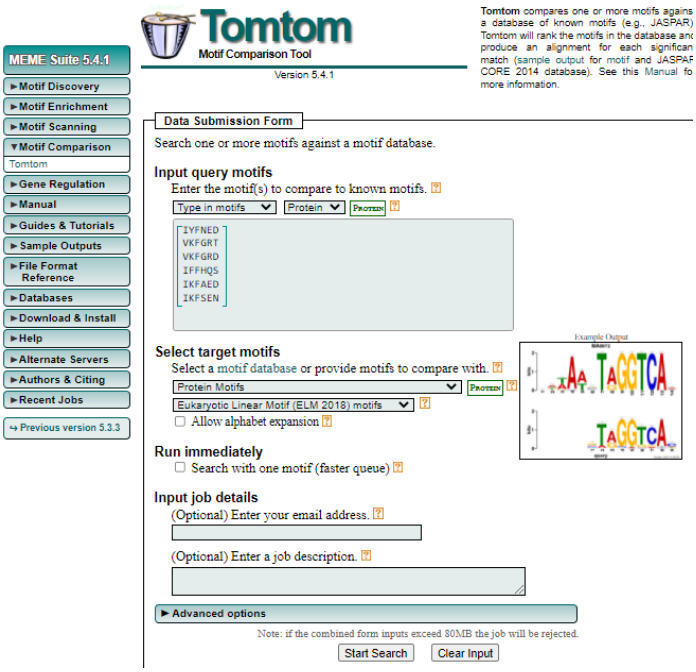
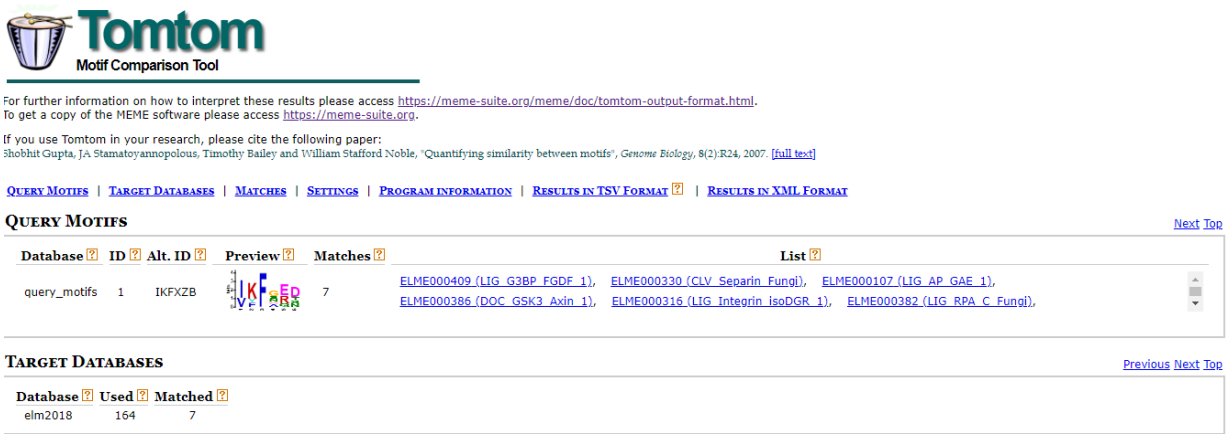


Fig. 38: Secuencia consenso de la primera región de epítopos seleccionada obtenida de Tomtom



A partir de las regiones comunes de epítopos preseleccionadas, *Tomtom* ha generado 19 epítopos consenso (mimotopos), guardados en un archivo denominado *Mimotop_1iok_A.txt*, cada uno con una lista de motivos proteicos análogos que, a su vez, incluyen una serie de proteínas asociadas sobre las que se ha investigado en fases posteriores del TFM si presentan patologías asociadas, de enfermedades del sistema inmune o de neoplasmas.

Tab 6: Tabla de Mimotopos de las Chaperonas de 60 kDA análogas en estructura a 1iok_A

No.	Start	End	Peptide	Number of residues
1	6	11	IKFXZB	6
2	32	32	G	1
3	45	48	GXPX	4
4	62	67	LEBXFE	6
5	82	84	DXA	3
6	133	142	XXKKXXTKEE	10
7	168	170	KV	2
8	182	182	G	1
9	199	212	LSPYFVTBXEKMXA	14
10	226	234	KJSNJQXJL	9
11	243	246	XKP	3
12	278	286	PGFGDRRKA	9
13	299	317	VISEDLGJXLEXXTJXXLG	19
14	334	342	AGXKXXIXX	9
15	351	372	IXXTTSDYDREKLQERLAKLAG	22
16	387	394	XEXKEKKH	8
17	430	434	GXXXD	5
18	485	489	XXBXX	5
19	526	542	PXXXAXAXXXPGXGM	15

4. Análisis de las Chaperonas de 60 kDa con Estructura Análoga a 1iok_A

4.1 Creación de la Base de datos de Proteínas análogas al Conjunto de Mimotopos

4.1.1 Análisis Comparativo del Conjunto de Mimotopos

Una vez obtenidos los mimotopos consenso de las *Chaperonas de 60 kDa con estructura análoga a 1iok_A*, se ha procedido a realizar un análisis comparativo con *FASTM* para obtener una lista de proteínas humanas con epítopos con un alto grado de similitud.

FASTM no acepta motivos con valores no específicos (X), por lo que para poder realizar la comparación con secuencias similares se ha seleccionado una secuencia lo más representativa de las secuencias consenso obtenidas con *Tomtom*.

En algunos epítopos, especialmente en los epítopos 17 y 19, que presentan gaps y discontinuidad en algunos de los aminoácidos, ha sido muy difícil seleccionar una secuencia representativa por lo que se ha realizado la comparación con diferentes combinaciones.

Aunque los resultados obtenidos no han sido exactamente iguales en función de la combinación seleccionada, la lista de proteínas obtenida en cada caso ha sido casi idéntica, cambiando simplemente el orden o el e-value de las proteínas elegidas excepto en los últimos valores de la lista, que presentan e-values bastante altas, en que puntualmente ha aparecido o desaparecido alguna de las proteínas listadas.

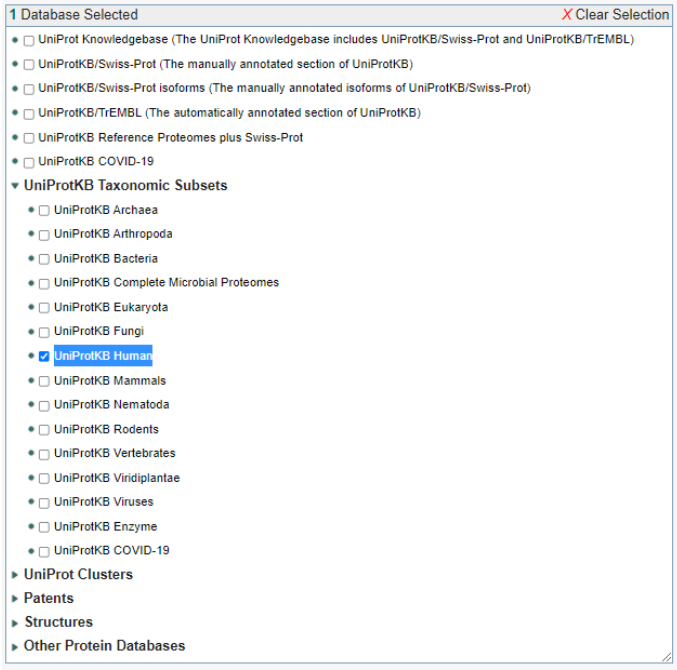
Por tanto, aunque puede haber pequeñas variaciones en los resultados obtenidos, se considera bastante representativa la secuencia analizada con *FASTM* (https://www.ebi.ac.uk/Tools/sss/fastm/):

> 1iok_A
IKFGZB,
G,
GSPT,
LEBKF,
DIA,
KSKKITTKEE,
KV,
G,
LSPYFVTBNEKMEA,
KJSNJQDJL,
GKP,
PGFGDRRKA,
VISEDLGJKLETVTJEQLG,
AGSKEAIDA,
IEETTSDYDREKLQERLAKLAG,
VELKEKKD,
GDSGD,
WVBMI,
PKKAAAPGMPPGMGM

Una cuestión importante cuando se pretende realizar la comparación de secuencias, para acotar los resultados, es seleccionar únicamente la base de datos con la que se pretende realizar la comparación de secuencias, *UniProtKB Human*, puesto que en caso contrario

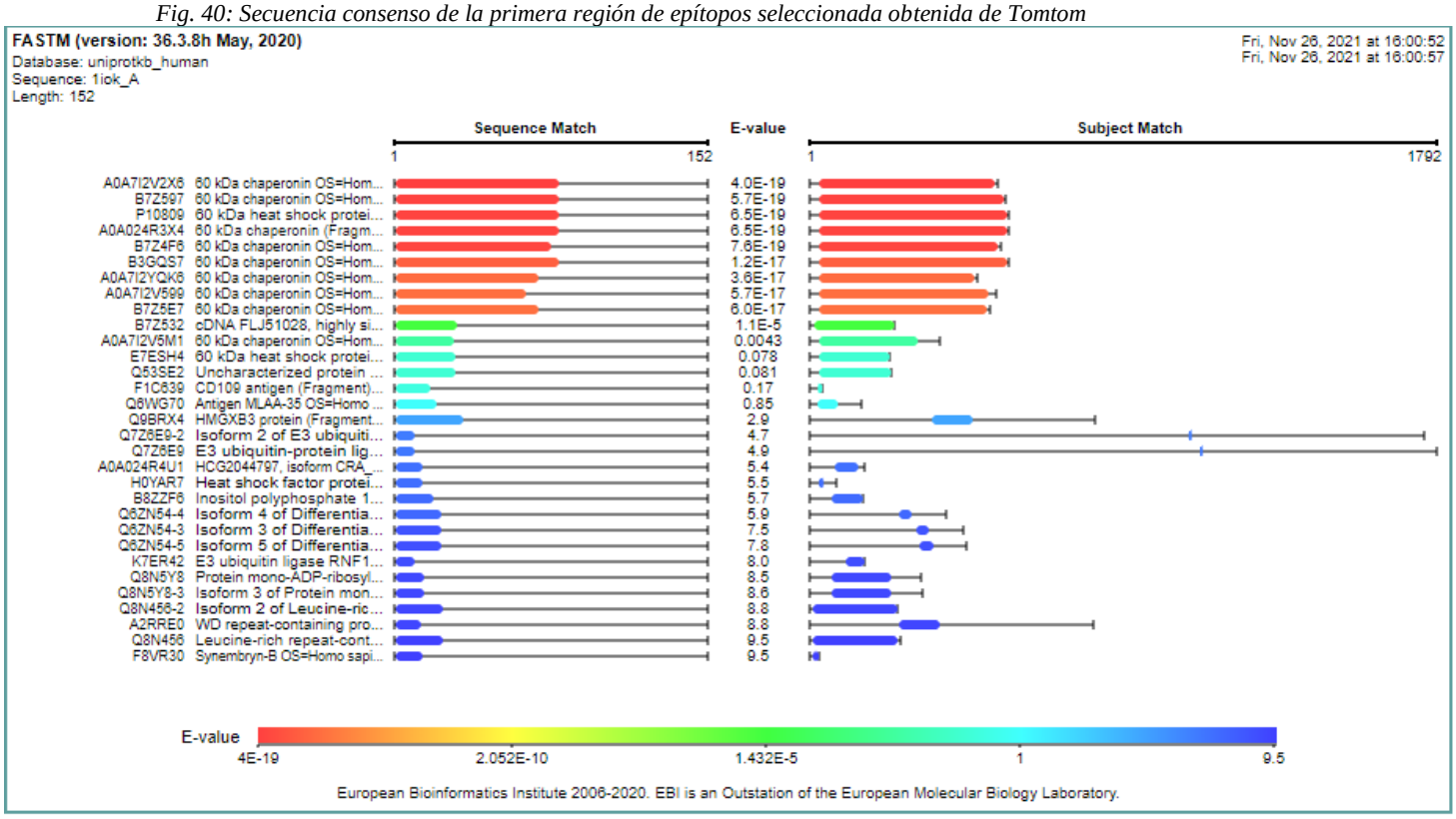
los resultados presentados por *FASTM* podrían no proporcionar secuencias proteicas humanas.

Fig. 39: Pantalla de selección de las bases de datos con las que hacer la comparación de secuencias en FASTM



El análisis ha proporcionado una lista de 40 proteínas con un alto grado de similitud respecto a la totalidad de mimotopos de 1iok_A, denominado *1iok_A FASTM.txt*, mediante la herramienta de comparación de secuencias *FASTM*.

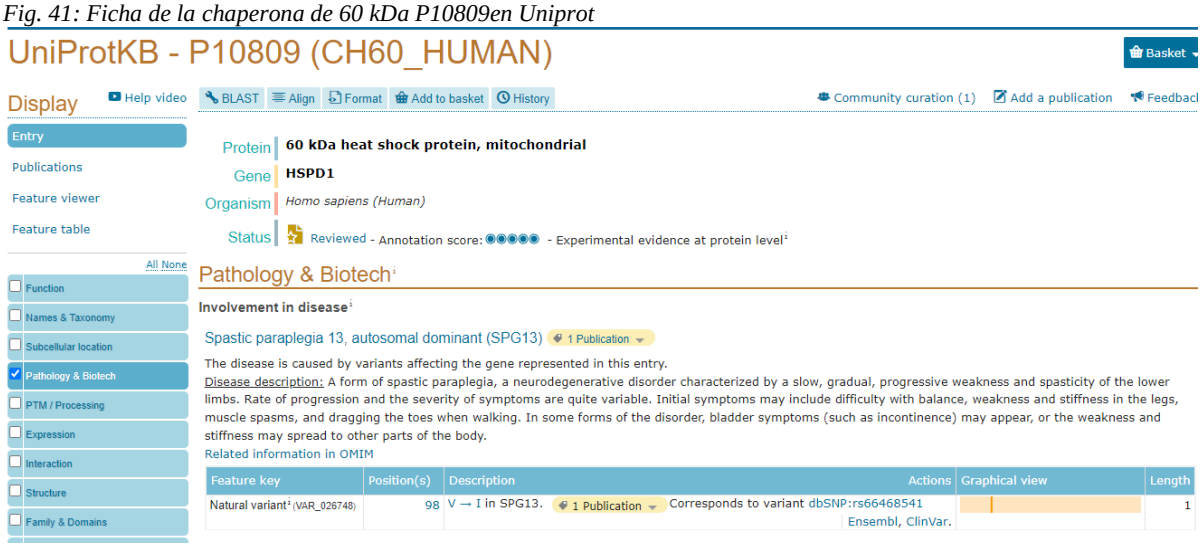
Fig. 40: Secuencia consenso de la primera región de epítopos seleccionada obtenida de Tomtom



4.1.2 Recopilación de Datos de Proteínas con Estructura Análoga al Conjunto de Mimotopos

Antes de proceder a realizar la base de datos se ha buscado información individual de las proteínas obtenidas en *Uniprot* (https://www.uniprot.org/uniprot/). Dentro de cada ficha de información de la proteína hay un apartado dedicado a las patologías asociadas de cada proteína, *Pathology & Biotech* en el que hay enlazadas varias páginas de bases de datos dedicadas específicamente a estas patologías asociadas.

Fig. 41: Ficha de la chaperona de 60 kDa P10809en Uniprot



Las dos bases de datos de detección de asociaciones entre proteínas y patologías que se han utilizado para referenciar el número de patologías asociadas a cada proteína han sido *DisGeNet* (https://www.disgenet.org/) y *Open Targets* (https://platform.opentargets.org/).

Fig. 41: Ficha de patologías asociadas a la chaperona de 60 kDa P10809 en DisGeNet

Fig. 41: Ficha de patologías asociadas a la chaperona de 60 kDa P10809 en Open Targets

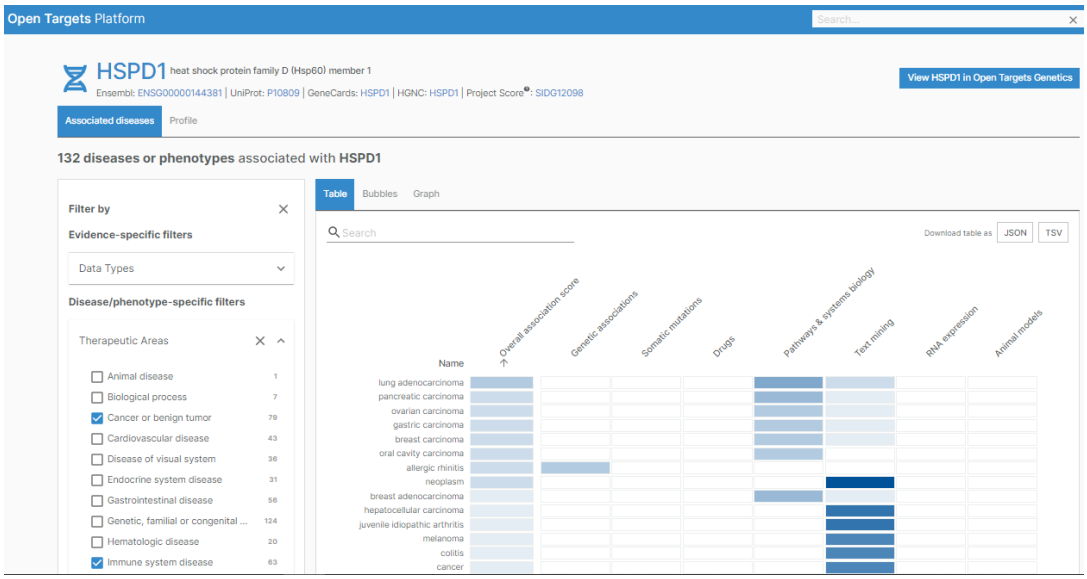


Fig. 42: Cabecera de la base de datos 1iok_A.xlsx

A	B	C	D	E	F	G	H	I	J	K	L	M
E-Value	UniProt	Nombre	Proteína	Gen	Organismo	Patología	Fuente	Nº Fuente	Nº Referencias	Start Epitopo	End Epitopo	AACC Secuencia
2.3e-19	A0A7I2V2X6	A0A7I2V2X6_HUMAN	60 kDa chaperonin	HSPD1	Homo sapiens	ND	Uniprot	1	0	30	533	533
3.7e-19	P10809	CH60_HUMAN	60 kDa heat shock protein, mitochondrial	HSPD1	Homo sapiens	Neoplasm	DisGeNET	1	90	30	533	533
3.7e-19	P10809	CH60_HUMAN	60 kDa heat shock protein, mitochondrial	HSPD1	Homo sapiens	Immune System Disease	DisGeNET	2	28	30	533	533
3.7e-19	P10809	CH60_HUMAN	60 kDa heat shock protein, mitochondrial	HSPD1	Homo sapiens	Cancer or benign tumor	OpenTargets	3	82	30	533	533
3.7e-19	P10809	CH60_HUMAN	60 kDa heat shock protein, mitochondrial	HSPD1	Homo sapiens	Immune System Disease	OpenTargets	4	68	30	533	533
2.1e-17	A0A7I2YQK6	A0A7I2YQK6_HUMAN	60 kDa chaperonin	HSPD1	Homo sapiens	ND	Uniprot	1	0	30	474	482
3.2e-17	A0A7I2V599	A0A7I2V599_HUMAN	60 kDa chaperonin	HSPD1	Homo sapiens	ND	Uniprot	1	0	30	514	537
0.0024	A0A7I2V5M1	A0A7I2V5M1_HUMAN	60 kDa chaperonin	HSPD1	Homo sapiens	ND	Uniprot	1	0	30	311	375
9.6	A0A7I2V5K3	A0A7I2V5K3	60 kDa chaperonin	HSPD1	Homo sapiens	ND	Uniprot	1	0	30	251	296
2.7	Q7Z6E9	RBBP6_HUMAN	E3 ubiquitin-protein ligase RBBP6	RBBP6	Homo sapiens	Neoplasm	DisGeNET	1	34	1085	1094	1758
2.7	Q7Z6E9	RBBP6_HUMAN	E3 ubiquitin-protein ligase RBBP6	RBBP6	Homo sapiens	Immune System Disease	DisGeNET	2	2	1085	1094	1758
2.7	Q7Z6E9	RBBP6_HUMAN	E3 ubiquitin-protein ligase RBBP6	RBBP6	Homo sapiens	Cancer or benign tumor	OpenTargets	3	32	1085	1094	1758
2.7	Q7Z6E9	RBBP6_HUMAN	E3 ubiquitin-protein ligase RBBP6	RBBP6	Homo sapiens	Immune System Disease	OpenTargets	4	6	1085	1094	1758
2.8	Q7Z6E9	RBBP6_HUMAN	E3 ubiquitin-protein ligase RBBP6	RBBP6	Homo sapiens	ND	FASTM	5	0	1119	1128	1792
3.3	B8ZZF6	B8ZZF6_HUMAN	Inositol polyphosphate 1-phosphatase	INPP1	Homo sapiens	Cancer or benign tumor	OpenTargets	1	14	65	155	155

A partir de la información obtenida de *FASTM*, *DisGeNet* y *Open Targets* se ha creado una base de datos, denominada *1iok_A*, con **41 anotaciones** de proteínas compuestas de **12 variables** y otra base de datos, denominada *1iok_A_Disease*, en la que se incluye una lista de patologías asociadas a las proteínas humanas análogas a *1iok_A* compuesta de **437 anotaciones** con **4 variables**.

4.2 Análisis de la Base de Datos 1iok_A

4.2.1 Estructura y Resumen de la Base de Datos 1iok_A

A partir de la lista de proteínas obtenidas, obtenidas por *DisGeNET* y/o *Open Targets*, se ha realizado una base de datos, *1iok_A.xlsx*, en la cual se han incluido el número de referencias de patologías asociadas a cada una de las proteínas.

El primer paso del análisis estadístico de la base de datos *1iok_A* ha sido el pre-procesamiento de la misma para darle un formato más adecuado para poder ser procesada en *R*.

El pre-procesamiento ha consistido en eliminar la variable *Organismo*, al ser todas proteínas humanas, en trasformar la variable *E-Value* en variable numérica y en cambiar todas las variables de tipo *carácter* a variables de tipo *factor*.

La base de datos se estructura en **41 observaciones**, una por cada fuente asociada, compuesta cada una de ellas de **12 variables**. Las variables incluidas en cada una de las observaciones son:

- *E-Value*: Valor estadístico que indica el número de alineamientos que se espera para una puntuación (score) X (o superior) en el análisis comparativo.
- *Uniprot*: Variable factorial que indica el código *Uniprot* de la proteína análoga a la secuencia consenso de epítopos.
- *Nombre*: Variable factorial que indica el nombre que se le da a la proteína humana análoga.
- *Proteína*: Variable factorial que indica el tipo de proteína a la cual pertenece la proteína humana análoga.
- *Gen*: Variable factorial que indica el gen a la cual pertenece la proteína humana análoga.
- *Patología*: Variable factorial que indica el tipo de patología asociado a la proteína humana análoga.
- *Fuente*: Variable factorial que indica la fuente donde se han obtenidos las patologías asociadas a la proteína humana análoga o, en el caso de estar repetida la fuente, el origen de la observación (*FASTM* o *Uniprot*).
- *Nº Fuente*: Variable factorial que indica el número de fuente asociado al gen al que pertenece la proteína humana análoga.
- *Nº Referencia*: Variable factorial que indica el número de enfermedades asociadas a la proteína presentes en la Fuente de origen.
- *Start Epítopo*: Variable numérica que indica el punto inicial de similitud de los epítopos con la proteína humana análoga.
- *End Epítopo*: Variable numérica que indica el punto final de similitud de los epítopos con la proteína humana análoga.
- *AACC Secuencia*: Variable numérica que indica el número de aminoácidos de la proteína.

A partir del análisis comparativo se han obtenido **477 referencias de enfermedades asociadas** pertenecientes a **19 proteínas** de **12 tipos de proteína diferentes**, que se distribuyen entre **10 genes**.

Analizando los datos se puede observar que:

- **6** de las 19 proteínas obtenidas forman parte del grupo de las *chaperonas de 60 kDa*, algo predecible teniendo en cuenta que las proteínas a partir de las cuales se han obtenido los mimótopos consenso también son *chaperonas de 60 kDa* sintetizadas por el gen *HSPD1*.
- Dentro de los **10 genes** representados en la base de datos, aparte del gen *HSPD1*, hay dos más con varias proteínas representadas en la lista: el gen *INPP1*, con **4 proteínas**, todas del tipo *Inositol polyphosphate 1-phosphatase* y el gen *DEF8* con **2 proteínas**, una del tipo *Differentially expressed in FDCP 8 homolog* y la otra del tipo *Serine palmitoyltransferase 3 (Fragment)*.
- La proteína con mayor número de anotaciones es *Q6ZN54*, con **6 anotaciones**, siendo tres de las mismas variantes de la proteína.

Tab 7: Proteínas incluidas en la base de datos 1iok_A

1	2.3e-19	A0A7I2V2X6	60 kDa chaperonin	HSPD1
2	3.7e-19	P10809	60 kDa heat shock protein, mitochondrial	HSPD1
3	2.1e-17	A0A7I2YQK6	60 kDa chaperonin	HSPD1
4	3.2e-17	A0A7I2V599	60 kDa chaperonin	HSPD1
5	2.4e- 3	A0A7I2V5M1	60 kDa chaperonin	HSPD1
6	9.6e+ 0	A0A7I2V5K3	60 kDa chaperonin	HSPD1

```
 7   2.7e+ 0 Q7Z6E9     E3 ubiquitin-protein ligase RBBP6           RBBP6
 8   3.3e+ 0 B8ZZF6     Inositol polyphosphate 1-phosphatase        INPP1
 9   6.1e+ 0 E7ET59     Inositol polyphosphate 1-phosphatase        INPP1
10   8.5e+ 0 E7EUX4     Inositol polyphosphate 1-phosphatase        INPP1
11   8.6e+ 0 E7ENF2     Inositol polyphosphate 1-phosphatase        INPP1
12   3.4e+ 0 Q6ZN54     Differentially expressed in FDCP 8 homolog  DEF8
13   4.6e+ 0 K7ER42     E3 ubiquitin ligase RNF157                  RNF157
14   4.9e+ 0 Q8N5Y8     Protein mono-ADP-ribosyltransferase PARP16  PARP16
15   5.1e+ 0 Q8N456     Leucine-rich repeat-containing protein 18   LRRC18
16   5.7e+ 0 F8VR30     Synembryn-B                                 RIC8B
17   6.1e+ 0 Q9Y3A4     Ribosomal RNA-processing protein 7 homolog  RRP7A
18   5.7e+ 0 B1AKS2     Serine palmitoyltransferase 3 (Fragment)    DEF8
19   9.8e+ 0 Q8IZ41     Ras and EF-hand domain-containing protein   RASEF
```

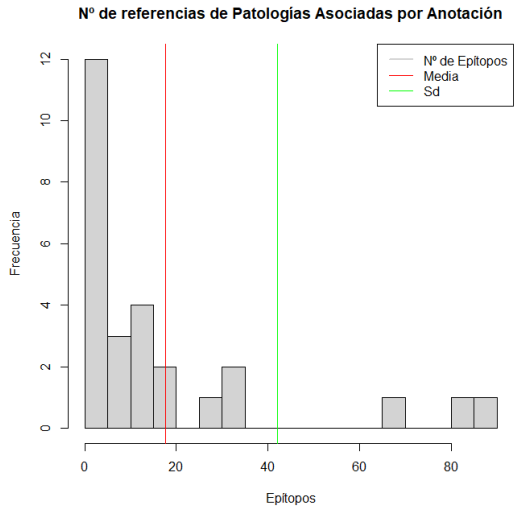
4.2.2 Análisis de las Referencias de Patologías Asociadas a las Proteínas Humanas Análogas a 1iok_A

El objetivo de esta base de datos es obtener una lista de proteínas con patologías asociadas, por lo que el siguiente proceso realizado ha sido observar la distribución del número de referencias por proteína.

4.2.2.1 Análisis de Referencias de Patologías Asociadas por Anotación

Analizando la distribución de referencias patológicas por proteína, se puede apreciar que hay **27 anotaciones** con patologías asociadas, con una distribución media de **17,67** referencias por anotación, aunque la mayor parte de las anotaciones tienen menos de 5 patologías asociadas y el 75% están por debajo de la media.

Fig. 43: Histograma de distribución del número de epítopos por organismo



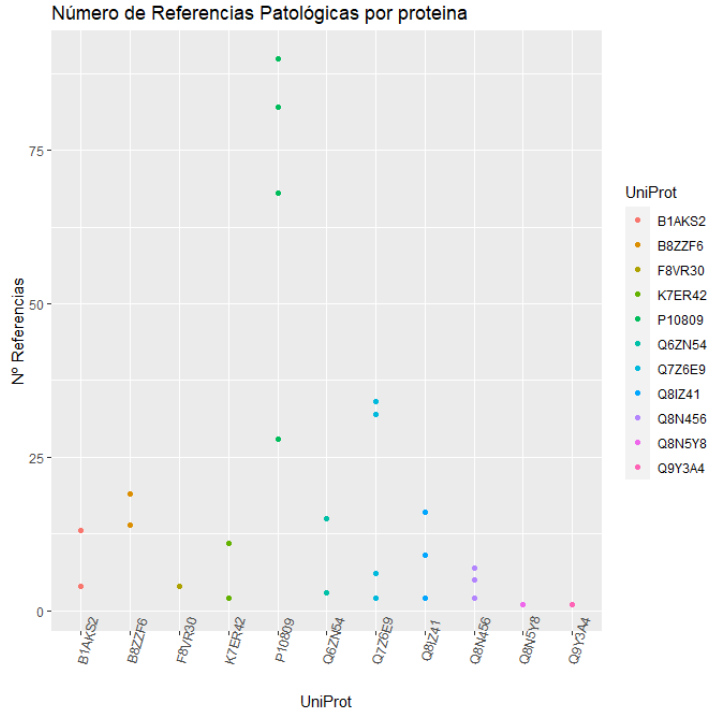
4.2.2.2 Análisis de Referencias de Patologías Asociadas por Proteína

El 25% de anotaciones superior se concentra en tres proteínas:

- *P10809* con **268 referencias**.
- *Q7Z6E9* con **74 referencias**.
- *B8ZZF6* con **33 referencias**.

Este hecho se ve claramente analizando el diagrama de puntos, en que las tres anotaciones destacan sobre el resto de proteínas, especialmente *P10809*.

Fig. 43: Histograma de distribución del número de epítopos por organismo



Tab 8: Anotaciones con Más de Quince Referencias de Patologías Asociadas

	UniProt	Gen	Patología	Fuente	`Nº Referencias`
1	P10809	HSPD1	Neoplasm	DisGeNET	90
2	P10809	HSPD1	Cancer or benign tumor	OpenTargets	82
3	P10809	HSPD1	Inmune System Disease	OpenTargets	68
4	Q7Z6E9	RBBP6	Neoplasm	DisGeNET	34
5	Q7Z6E9	RBBP6	Cancer or benign tumor	OpenTargets	32
6	P10809	HSPD1	Inmune System Disease	DisGeNET	28
7	B8ZZF6	INPP1	Inmune System Disease	OpenTargets	19
8	Q8IZ41	RASEF	Cancer or benign tumor	OpenTargets	16
9	Q6ZN54	DEF8	Cancer or benign tumor	OpenTargets	15

4.2.2.3 Análisis de Referencias de Patologías Asociadas por Fuente

Referencias extraídas de DisGeNet

En *DisGeNet* la mayor parte de las anotaciones son de *neoplasmas*, tratándose principalmente de patologías asociadas a *P10809* y, en menor medida, a *Q7Z6E9*.

Tab 9: Lista de Referencias extraídas de DisGeNet

	UniProt	Gen	Patología	Nº Referencias
1	P10809	HSPD1	Neoplasm	90
2	Q7Z6E9	RBBP6	Neoplasm	34
3	P10809	HSPD1	Inmune System Disease	28
4	Q8IZ41	RASEF	Neoplasm	9
5	Q6ZN54	DEF8	Neoplasm	3
6	Q7Z6E9	RBBP6	Inmune System Disease	2
7	Q8N456	LRRC18	Inmune System Disease	2

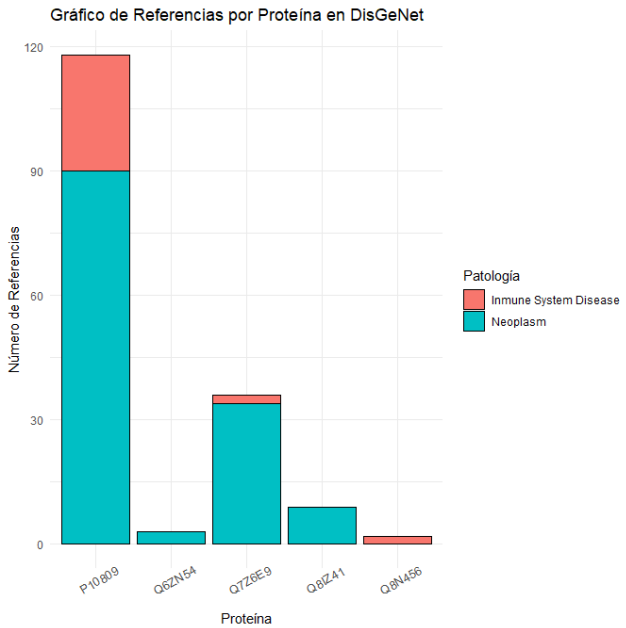


Fig. 44: Distribución del nº de referencias por proteína en DisGeNet

Referencias Extraídas de Open Targets

Por otro lado, las anotaciones de *neoplasmas* y de *ESI* están más equilibradas en *Open Targets*, siendo principalmente anotaciones de *P10809* y, en menor medida de *Q7Z6E9* y de *B8ZZF6*.

Cabe destacar que en esta última proteína hay más anotaciones de patologías asociadas a *ESI (enfermedades del sistema inmune)* que a *neoplasmas*.

Tab 10: Tabla referencias extraídas de Open Targets

	UniProt	Gen	Patología	Nº Referencias
1	P10809	HSPD1	Cancer or benign tumor	8
2	P10809	HSPD1	Inmune System Disease	68
3	Q7Z6E9	RBBP6	Cancer or benign tumor	32
4	B8ZZF6	INPP1	Inmune System Disease	19
5	Q8IZ41	RASEF	Cancer or benign tumor	16
6	Q6ZN54	DEF8	Cancer or benign tumor	15
7	B8ZZF6	INPP1	Cancer or benign tumor	14
8	B1AKS2	DEF8	Cancer or benign tumor	13
9	K7ER42	RNF157	Cancer or benign tumor	11
10	Q8N456	LRRC18	Cancer or benign tumor	7
11	Q7Z6E9	RBBP6	Inmune System Disease	6
12	Q8N456	LRRC18	Inmune System Disease	5
13	F8VR30	RIC8B	Cancer or benign tumor	4
14	F8VR30	RIC8B	Inmune System Disease	4
15	B1AKS2	DEF8	Inmune System Disease	4
16	Q6ZN54	DEF8	Inmune System Disease	3
17	K7ER42	RNF157	Inmune System Disease	2
18	Q8IZ41	RASEF	Inmune System Disease	2
19	Q8N5Y8	PARP16	Cancer or benign tumor	1
20	Q9Y3A4	RRP7A	Cancer or benign tumor	1

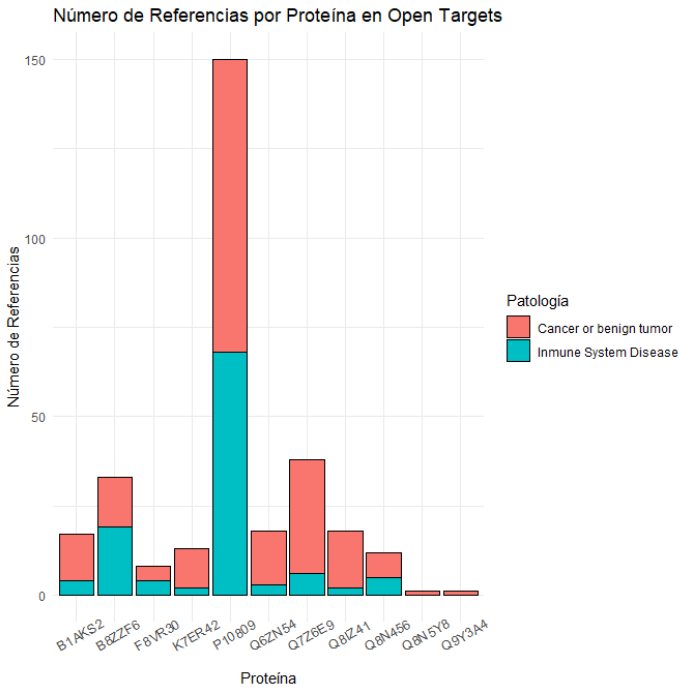


Fig. 45: Distribución del nº de referencias por proteína en Open Targets

4.3 Análisis de las Patologías Asociadas a Proteínas Humanas con Epítopos Análogos a 1iok_A

Para poder analizar más en profundidad el tipo de patologías asociadas, se ha creado otra base de datos, *1iok_A_Disease.xlsx*, en la que se han incluido todas las referencias de patologías asociadas a cada proteína analizada en *1iok_A.xlsx*.

4.3.1 Estructura y Resumen de la Base de Datos *1iok_A_Disease.xlsx*

En esta base de datos el pre-procesamiento ha consistido en convertir en *factor* todas las variables no numéricas de la base de datos.

La base de datos *1iok_A_Disease.xlsx* se estructura en **437 anotaciones** compuestas de **4 variables**. Las variables incluidas en cada una de las observaciones son:

- *Proteína*: Variable factorial que indica el código *Uniprot* de la proteína objeto de la anotación.
- *Patología*: Variable factorial que indica la patología asociada a la proteína.
- *Tipología*: Variable factorial que indica el tipo de patología asociado a la proteína.
- *Fuente*: Variable factorial que indica la fuente de dónde se han obtenidos las anotaciones.

A partir del análisis comparativo se han obtenido **305 patologías** asociadas a **11 proteínas**. Las anotaciones se han extraído de 2 fuentes:

- *DisGeNet*, con **121 observaciones (27,69%)**.
- *Open Targets*, con **316 observaciones**.

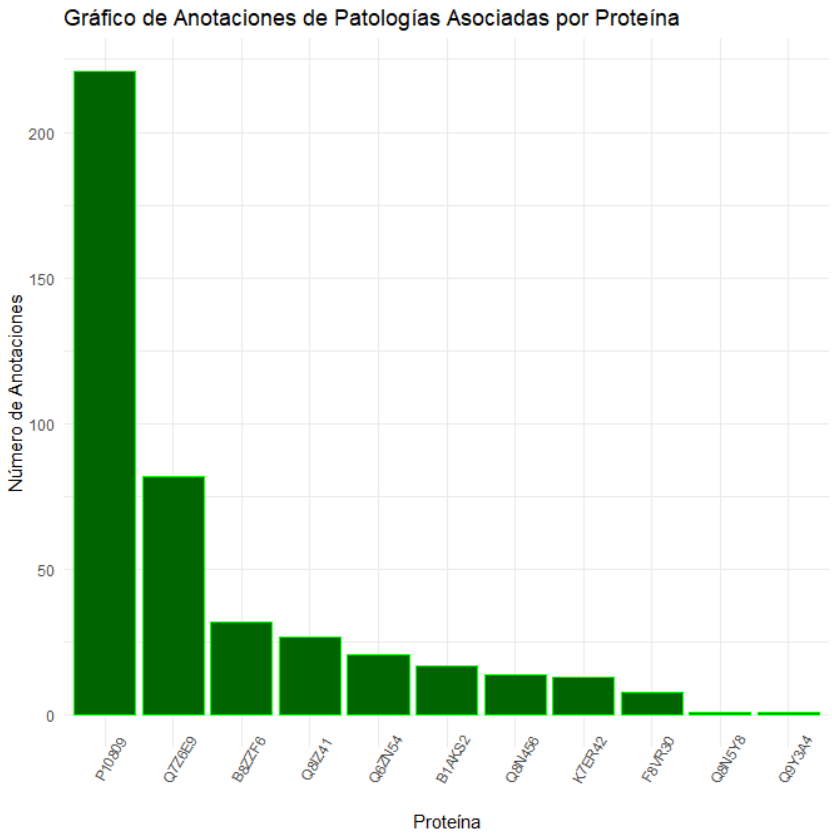
Del total de anotaciones, el **60,7% (274)** se corresponden a *Neoplasmas* mientras que el resto son *Enfermedades del Sistema Inmune*, perteneciendo mayoritariamente a la proteína *P10809*, que supone el **50,57% (221)** de las anotaciones de *neoplasmas*.

Tab 11: Sumario de 1iok_A_Disease

Proteína	Patología		
P10809 :221	cancer: 8		
Q7Z6E9 : 82	melanoma	:	8
B8ZZF6 : 32	neoplasm	:	6
Q8IZ41 : 27	acute myeloid leukemia	:	4
Q6ZN54 : 21	chronic myelogenous leukemia	:	4
B1AKS2 : 17	glioblastoma multiforme	:	4
(Other): 37	(Other)	:	403

Tipología		Fuente	
Inmune System Disease	: 163	DisGeNET	: 121
Neoplasma	: 274	OpenTargets	: 316

Fig. 46: Diagrama de barras de anotaciones con patologías asociadas por proteína



Si no se tienen en cuenta las denominaciones más genéricas como *cancer*, *melanoma* o *neoplasma*, las patologías más representadas en la base de datos son dos tipos de leucemia:

- *acute myeloid leukemia*
- *chronic myelogenous leukemia*

Tab 12: Tabla de frecuencias de patologías asociadas a la secuencia de mimotopos

	Patología	Frecuencia
1	cancer	8
2	melanoma	8
3	neoplasm	6
4	acute myeloid leukemia	4
5	chronic myelogenous leukemia	4
6	glioblastoma multiforme	4
7	hepatocellular carcinoma	4
8	multiple myeloma	4
9	ovarian neoplasm	4
10	breast carcinoma	3
11	breast ductal carcinoma in situ	3
12	carcinoma	3
13	Carcinoma of lung	3
14	Carcinoma, Lewis Lung	3
15	cervical cancer	3
16	colonic neoplasm	3
17	Crohn's disease	3
18	lung adenocarcinoma	3
19	lymph node metastatic carcinoma	3
20	Malignant neoplasm of lung	3
21	Neoplasms	3
22	neuroblastoma	3
23	non-small cell lung carcinoma	3
24	Primary malignant neoplasm of lung	3
25	psoriasis	3
26	ring dermoid of cornea	3
27	squamous cell carcinoma	3
28	Squamous cell carcinoma	3
29	systemic lupus erythematosus	3
30	type I diabetes mellitus	3
31	ulcerative colitis	3
32	urothelial carcinoma	3

4.3.2 Análisis Específico de la Base de Datos

4.3.2.1 Distribución de Patologías Asociadas por Fuente de Origen

En *DisGeNet* la mayor parte de las anotaciones, un **91,74%,** se corresponden a *neoplasmas*, presentando un porcentaje muy bajo de anotaciones asociadas a *enfermedades del sistema inmune*.

Esto se debe a que, aunque con el buscador de la página aparecen más anotaciones asociadas a *ESI*, cuando se descargan los archivos, éstos solo nombran un tipo de enfermedad, teniendo preferencia por anotar *neoplasmas* por encima de *ESI* cuando hay casos de enfermedades asociadas a ambos factores a la vez.

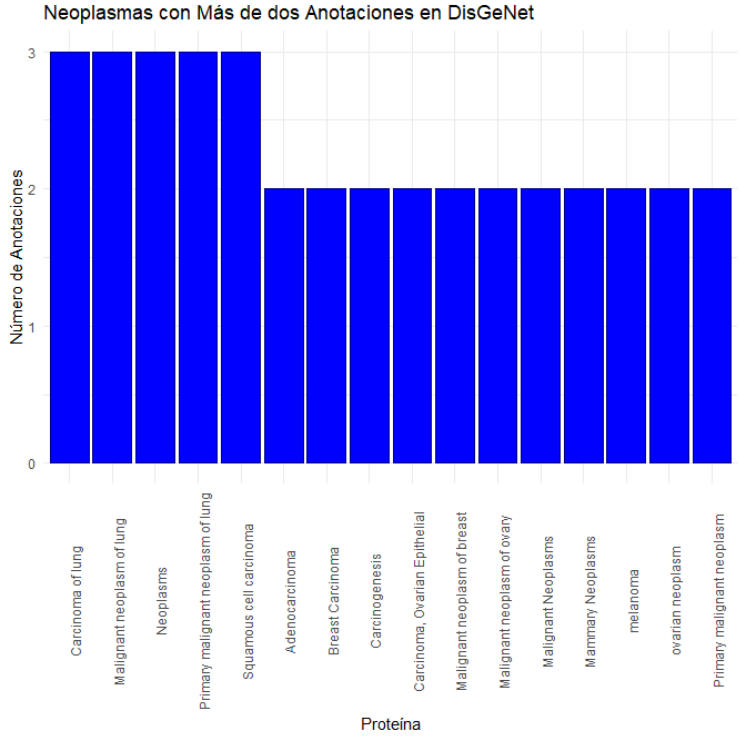
En *Open Targets*, en cambio, la distribución está más repartida, habiendo sólo una ligera diferencia en la proporción de patologías asociadas a *neoplasmas* **51,58%** respecto a las patologías asociadas a *enfermedades del sistema inmune*.

4.3.2.2 Distribución de Neoplasmas Asociados por Fuente de Origen

En *DisGeNet* la mayor parte de patologías con mayor número de anotaciones son *neoplasmas* originados en los pulmones:

- *Carcinoma of lung*.
- *Malignant neoplasm of lung*.
- *Primary malignant neoplasm of lung*.

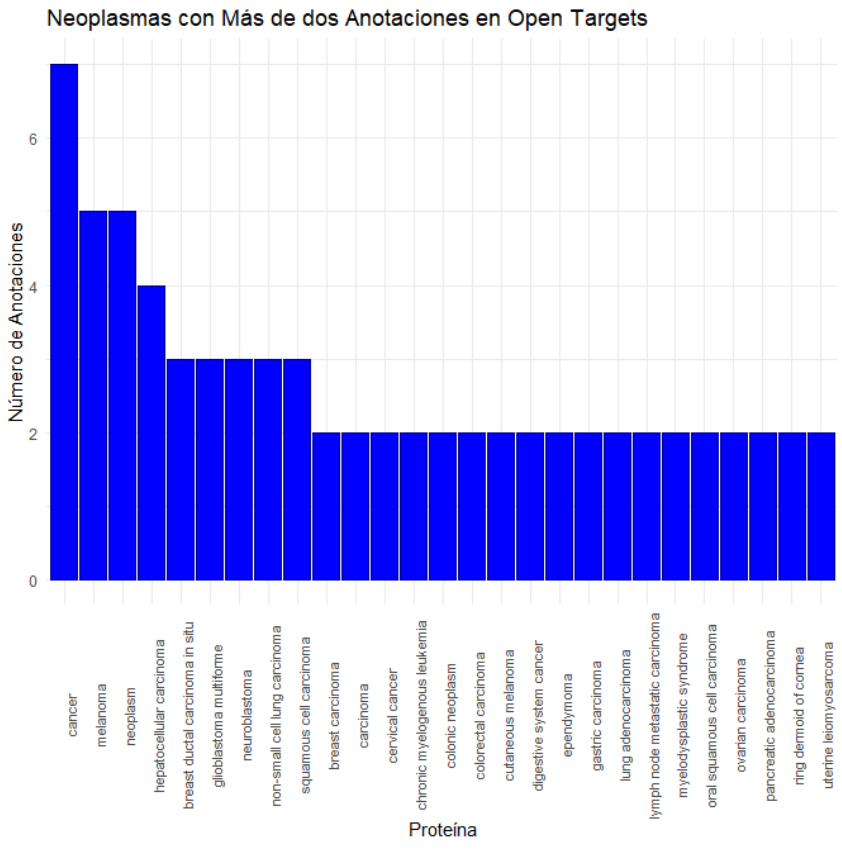
Fig. 47: Diagrama de barras de patologías con más de dos anotaciones en DisGeNet



En cambio, en *Open Targets* la representación de neoplasmas, sin tener en cuenta las denominaciones genéricas, es más variada, teniendo en las tres primeras posiciones *neoplasmas* de diferente origen:

- *hepatocellular carcinoma*.
- *breast ductal carcinoma in situ*.
- *glioblastoma multiforme*.

Fig. 46: Diagrama de barras de patologías con más de dos anotaciones en Open Targets



4.3.2.3 Distribución de Patologías Asociadas del Sistema inmune por Fuente de Origen

A nivel de patologías asociadas a *ESI* en *DisGeNet* hay muy pocas anotaciones, repitiéndose únicamente el *Guillain-Barre Syndrome*, a diferencia que en *Open Targets*, que sí que hay bastante variedad de patologías asociadas.

De esta forma, en *Open Targets* se observan **7 patologías asociadas** a *ESI* de diversa índole con **3 anotaciones**.

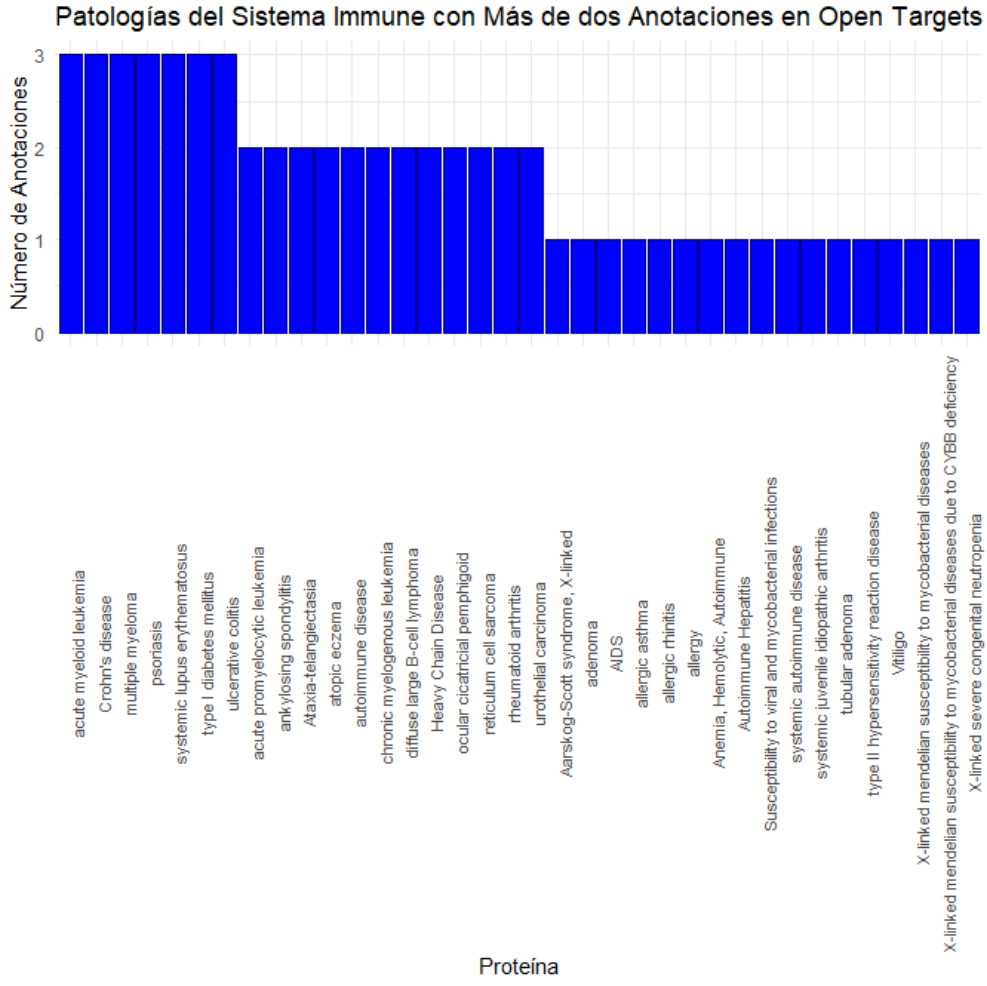
Patologías del sistema Inmune en DisGeNet

Tab 13: Frecuencia de Patologías asociadas en DisGeNet

	Patología	Frecuencia
1	*Autoimmune Diseases*	2
2	*AIDS-Associated Nephropathy*	1
3	*Allergic Reaction*	1
4	*Guillain-Barre Syndrome*	1
5	*Guillain-Barre Syndrome, Familial*	1
6	*Inmune System Diseases*	1
7	*Lupus Erythematosus*	1
8	*Multiple Sclerosis*	1
9	*Rheumatoid Arthritis*	1

Patologías del sistema Inmune en Open targets

Fig. 46: Diagrama de barras de patologías del sistema inmune con más de dos anotaciones en Open Targets



5 Análisis de las Proteínas Humanas con Motivos Análogos a cada Mimotopo

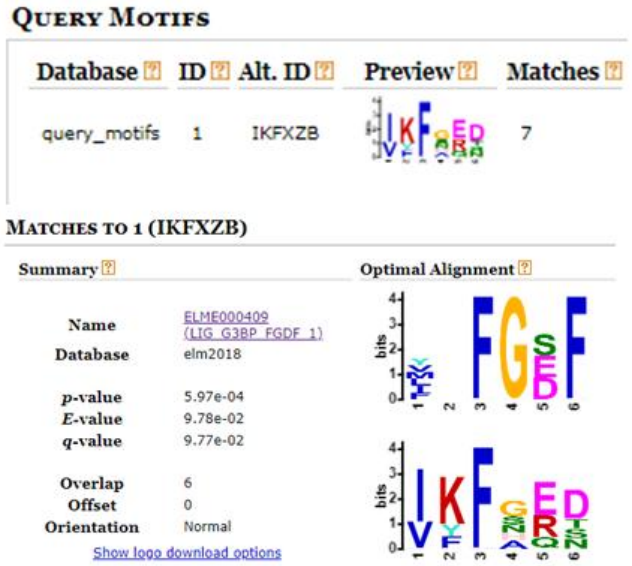
5.1 Creación de la Base de datos de Proteínas con motivos Análogos a cada Mimotopo

5.1.1 Selección de Proteínas con Motivos Análogos a cada Mimotopo

Una vez analizada la secuencia total de mimótopos obtenidos, se ha decidido realizar un análisis específico de los mimotopos de forma individualizada, ya que por medio de *Tomtom* se ha obtenido una lista de **130 motivos proteicos** análogos a alguno de los mimotopos de *1iok_A* que han sido guardados en un archivo denominado *Motivos.txt*.

De esta forma, la consulta de la primera región de epítopos comunes de las seis proteínas analizadas ha dado como resultado el mimotopo *IKFXZB*, que ha presentado analogía estructural con **7 motivos proteicos**.

Fig. 47: Pantalla de consulta de Tomtom con la primera coincidencia de motivos análogos al Mimotopo IKFXZB



La lista de motivos proteicos análogos al mimotopo *IKFXZB* es la siguiente:

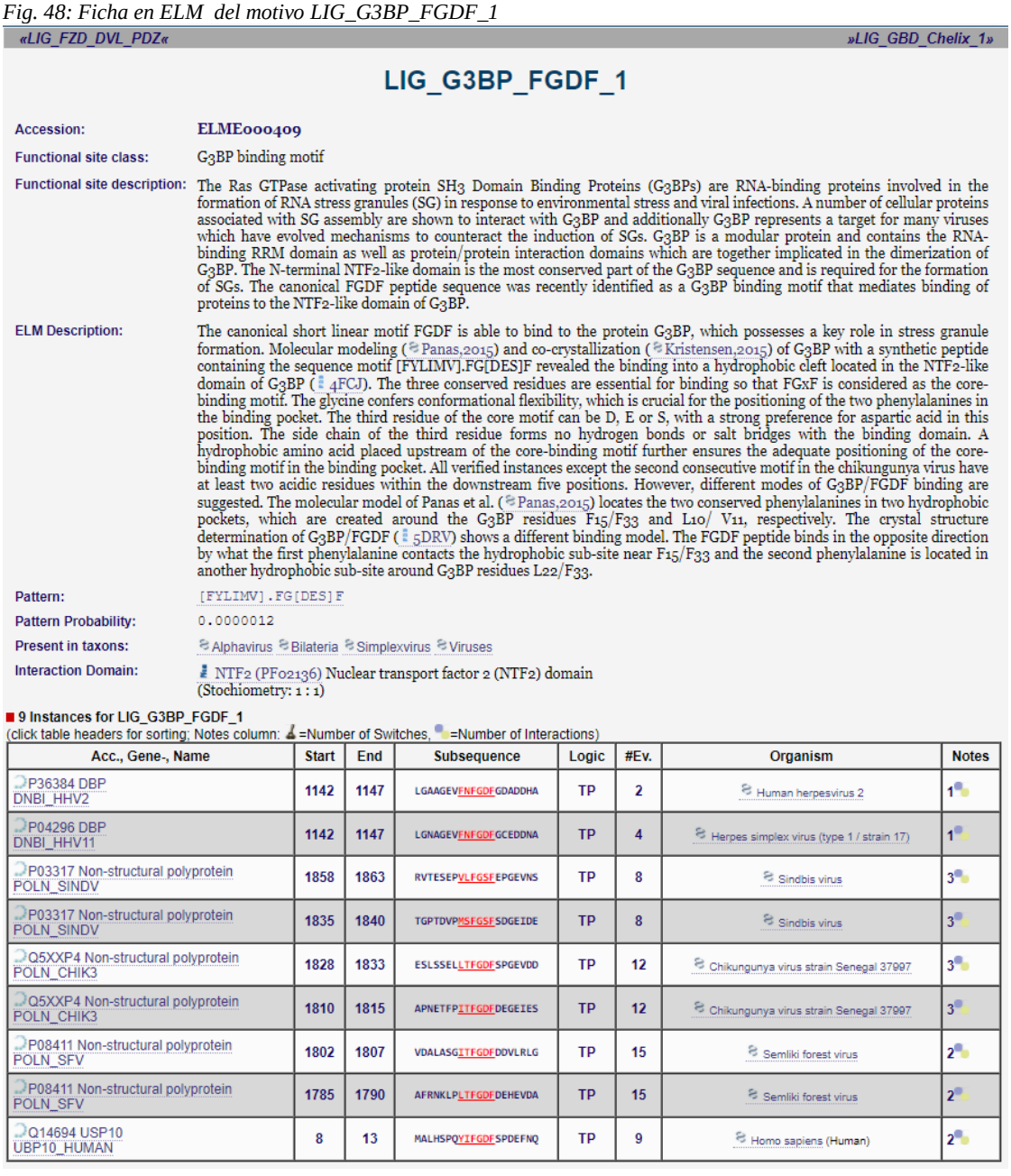
- ELME000409 (LIG_G3BP_FGDF_1)
- ELME000330 (CLV_Separin_Fungi)
- ELME000107 (LIG_AP_GAE_1)
- ELME000386 (DOC_GSK3_Axin_1)
- ELME000316 (LIG_Integrin_isoDGR_1)
- ELME000382 (LIG_RPA_C_Fungi)
- ELME000002 (MOD_SUMO_for_1)

Estos motivos proteicos se encuentran descritos en *ELM* (http://elm.eu.org/searchdb.html) y cada uno de ellos lleva asociado una lista de proteínas tanto humanas como de otros organismos con funciones similares.

Si se entra en el enlace de cada motivo de la base de datos *ELM*, se despliega una descripción de las características del motivo y una lista de las proteínas pertenecientes al mismo.

Por ejemplo, si se consulta el motivo *LIG_G3BP_FGDF_1*, análogo al mimotopo *IKFXZB* se observa que éste engloba un grupo de poliproteínas con funciones cruciales en los procesos de infección viral y la proteína humana *Ubiquitin carboxyl-terminal hydrolase 10*, implicada en procesos de autofagia y reparación de de ADN.

Fig. 48: Ficha en ELM del motivo LIG_G3BP_FGDF_1



«LIG_FZD_DVL_PDZ« »LIG_GBD_Chelix_1»

LIG_G3BP_FGDF_1

Accession: ELME000409

Functional site class: G3BP binding motif

Functional site description: The Ras GTPase activating protein SH3 Domain Binding Proteins (G3BPs) are RNA-binding proteins involved in the formation of RNA stress granules (SG) in response to environmental stress and viral infections. A number of cellular proteins associated with SG assembly are shown to interact with G3BP and additionally G3BP represents a target for many viruses which have evolved mechanisms to counteract the induction of SGs. G3BP is a modular protein and contains the RNA-binding RRM domain as well as protein/protein interaction domains which are together implicated in the dimerization of G3BP. The N-terminal NTF2-like domain is the most conserved part of the G3BP sequence and is required for the formation of SGs. The canonical FGDF peptide sequence was recently identified as a G3BP binding motif that mediates binding of proteins to the NTF2-like domain of G3BP.

ELM Description: The canonical short linear motif FGDF is able to bind to the protein G3BP, which possesses a key role in stress granule formation. Molecular modeling (Panas,2015) and co-crystallization (Kristensen,2015) of G3BP with a synthetic peptide containing the sequence motif [FYLIMV].FG[DES]F revealed the binding into a hydrophobic cleft located in the NTF2-like domain of G3BP (4FCJ). The three conserved residues are essential for binding so that FGxF is considered as the core-binding motif. The glycine confers conformational flexibility, which is crucial for the positioning of the two phenylalanines in the binding pocket. The third residue of the core motif can be D, E or S, with a strong preference for aspartic acid in this position. The side chain of the third residue forms no hydrogen bonds or salt bridges with the binding domain. A hydrophobic amino acid placed upstream of the core-binding motif further ensures the adequate positioning of the core-binding motif in the binding pocket. All verified instances except the second consecutive motif in the chikungunya virus have at least two acidic residues within the downstream five positions. However, different modes of G3BP/FGDF binding are suggested. The molecular model of Panas et al. (Panas,2015) locates the two conserved phenylalanines in two hydrophobic pockets, which are created around the G3BP residues F15/F33 and L10/ V11, respectively. The crystal structure determination of G3BP/FGDF (5DRV) shows a different binding model. The FGDF peptide binds in the opposite direction by what the first phenylalanine contacts the hydrophobic sub-site near F15/F33 and the second phenylalanine is located in another hydrophobic sub-site around G3BP residues L22/F33.

Pattern: [FYLIMV].FG[DES]F

Pattern Probability: 0.0000012

Present in taxons: Alphavirus Bilateria Simplexvirus Viruses

Interaction Domain: NTF2 (PF02136) Nuclear transport factor 2 (NTF2) domain (Stochiometry: 1 : 1)

9 Instances for LIG_G3BP_FGDF_1

(click table headers for sorting; Notes column: =Number of Switches, =Number of Interactions)

Acc., Gene-, Name	Start	End	Subsequence	Logic	#Ev.	Organism	Notes
P36384 DBP DNBI_HHV2	1142	1147	LGAAGEVFNFGDFGDADDHA	TP	2	Human herpesvirus 2	1
P04296 DBP DNBI_HHV11	1142	1147	LGNAGEVFNFGDFGCEDDNA	TP	4	Herpes simplex virus (type 1 / strain 17)	1
P03317 Non-structural polyprotein POLN_SINDV	1858	1863	RVTESEPVLFGSFEPGEVNS	TP	8	Sindbis virus	3
P03317 Non-structural polyprotein POLN_SINDV	1835	1840	TGPTDVPMSFGSFSDGEIDE	TP	8	Sindbis virus	3
Q5XXP4 Non-structural polyprotein POLN_CHIK3	1828	1833	ESLSSELLTFGDFSPGEVDD	TP	12	Chikungunya virus strain Senegal 37997	3
Q5XXP4 Non-structural polyprotein POLN_CHIK3	1810	1815	APNETFPITFGDFDEGEIES	TP	12	Chikungunya virus strain Senegal 37997	3
P08411 Non-structural polyprotein POLN_SFV	1802	1807	VDALASGITFGDFDDVLRLG	TP	15	Semliki forest virus	2
P08411 Non-structural polyprotein POLN_SFV	1785	1790	AFRNKLPLTFGDFDEHEVDA	TP	15	Semliki forest virus	2
Q14694 USP10 UBP10_HUMAN	8	13	MALHSPQYIFGDFSPDEFNQ	TP	9	Homo sapiens (Human)	2

El análisis de los mimotopos individuales ha proporcionado una gran cantidad de motivos análogos extraídos de *ELM*, cada uno de los cuales está compuesto de una lista de proteínas que, tras ser revisadas con *Uniprot*, en muchos casos presentan patologías asociadas, tanto de *neoplasmas* como de *enfermedades del sistema inmune*.

Este hecho ha obligado a replantearse objetivos, priorizando la creación y testeo de un protocolo de análisis de epítopos y la obtención de una lista de patologías asociadas a los mismos en lugar de crear una base de datos con toda la lista de proteínas con motivos análogos a los mimotopos de la *chaperonas de 60kDA con estructura análoga a 1iok_A.*

Por tanto, finalmente solo se han analizado en profundidad tres mimotopos, de forma que existiera una variedad de datos y motivos que hiciera más representativo el análisis pero que, a la vez, fuera posible completar el resto de objetivos específicos.

Los mimotopos analizados han sido:

- ***IKFXZB***
- ***G***
- ***GXPX***

Todas las anotaciones generadas se han guardado en una base de datos, *Mimotopos.xlsx*, en la que se han incluido el número de referencias de patologías asociadas a todas las proteínas de los motivos proteicos extraídos de *ELM*.

Fig. 49: Cabecera de la base de datos de Mimotopos.xlsx

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
Mimotopo	PDB	E-Value	Motivo	UniProt	Nombre	Organismo	Proteína	Patología	Fuente	Nº Fuente	Nº Referencias	Start	End	Peptide	Residuos
IKFXZB	1iok_A	1.07e-02	LIG_G3BP_FGDF_1	P36384	DNBL_HHV2	Human herpesvirus 2	Major DNA-binding protein	viral infection	ELM	1	2	1142	1147	FNFGDF	6
IKFXZB	1iok_A	1.07e-02	LIG_G3BP_FGDF_1	P04296	DNBL_HHV11	Human herpesvirus 1 (strain 17)	Major DNA-binding protein	viral infection	ELM	1	4	1142	1147	FNFGDF	6
IKFXZB	1iok_A	1.07e-02	LIG_G3BP_FGDF_1	P03317	POLN_SINDV	Sindbis virus	Polyprotein P1234	viral infection	ELM	1	8	1858	1863	VLFGSF	6
IKFXZB	1iok_A	1.07e-02	LIG_G3BP_FGDF_1	P03317	POLN_SINDV	Sindbis virus	Polyprotein P1234	viral infection	ELM	2	8	1835	1840	MSFGSF	6
IKFXZB	1iok_A	1.07e-02	LIG_G3BP_FGDF_1	Q5XXP4	POLN_CHIK3	Chikungunya virus (strain 37997)	Polyprotein P1234	viral infection	ELM	1	12	1810	1815	LTFGDF	6
IKFXZB	1iok_A	1.07e-02	LIG_G3BP_FGDF_1	Q5XXP4	POLN_CHIK3	Chikungunya virus (strain 37997)	Polyprotein P1234	viral infection	ELM	2	2	1828	1833	ITFGDF	6
IKFXZB	1iok_A	1.07e-02	LIG_G3BP_FGDF_1	P08411	POLN_SFV	Semliki forest virus	Polyprotein P1234	viral infection	ELM	1	15	1802	1807	ITFGDF	6
IKFXZB	1iok_A	1.07e-02	LIG_G3BP_FGDF_1	P08411	POLN_SFV	Semliki forest virus	Polyprotein P1234	viral infection	ELM	2	15	1785	1790	LTFGDF	6
IKFXZB	1iok_A	1.07e-02	LIG_G3BP_FGDF_1	Q14694	UBP10_HUMAN	Homo sapiens	Ubiquitin carboxyl-terminal hydrolase 10	Neoplasm	DisGeNET	1	6	8	13	YIFGDF	6
IKFXZB	1iok_A	1.07e-02	LIG_G3BP_FGDF_1	Q14694	UBP10_HUMAN	Homo sapiens	Ubiquitin carboxyl-terminal hydrolase 10	Cancer or benign tumor	OpenTargets	2	15	8	13	YIFGDF	6
IKFXZB	1iok_A	1.07e-02	LIG_G3BP_FGDF_1	Q14694	UBP10_HUMAN	Homo sapiens	Ubiquitin carboxyl-terminal hydrolase 10	Immune System Disease	OpenTargets	3	2	8	13	YIFGDF	6
IKFXZB	1iok_A	2.64e-02	MOD_SUMO_for_1	Q2GIB5	AMPA_ANAPZ	Anaplasma phagocytophilum (strain HZ)	SUMOylated effector protein AmpA	bacterial infection	ELM	1	3	134	135	FKIE	4
IKFXZB	1iok_A	2.64e-02	MOD_SUMO_for_1	Q16665	HIF1A_HUMAN	Homo sapiens	Hypoxia-inducible factor 1-alpha	Neoplasm	DisGeNET	1	334	531	534	FKIE	4
IKFXZB	1iok_A	2.64e-02	MOD_SUMO_for_1	Q16665	HIF1A_HUMAN	Homo sapiens	Hypoxia-inducible factor 1-alpha	Immune System Disease	DisGeNET	2	50	531	534	FKIE	4
IKFXZB	1iok_A	2.64e-02	MOD_SUMO_for_1	Q16665	HIF1A_HUMAN	Homo sapiens	Hypoxia-inducible factor 1-alpha	Neoplasm	OpenTargets	3	338	531	534	FKIE	4
IKFXZB	1iok_A	2.64e-02	MOD_SUMO_for_1	Q16665	HIF1A_HUMAN	Homo sapiens	Hypoxia-inducible factor 1-alpha	Immune System Disease	OpenTargets	4	88	531	534	FKIE	4
IKFXZB	1iok_A	2.64e-02	MOD_SUMO_for_1	Q16665	HIF1A_HUMAN	Homo sapiens	Hypoxia-inducible factor 1-alpha	ND	ELM	5	7	476	479	LKLE	4
IKFXZB	1iok_A	2.64e-02	MOD_SUMO_for_1	Q16665	HIF1A_HUMAN	Homo sapiens	Hypoxia-inducible factor 1-alpha	ND	ELM	6	7	390	393	LKKE	4
IKFXZB	1iok_A	4.50e-02	MOD_ProDKin_1	Q16665	HIF1A_HUMAN	Homo sapiens	Hypoxia-inducible factor 1-alpha	ND	ELM	7	7	638	644	LIASPSP	7
GXPX	1iok_A	4.50e-02	MOD_CDK_SPxK_1	Q16665	HIF1A_HUMAN	Homo sapiens	Hypoxia-inducible factor 1-alpha	ND	ELM	8	6	665	671	RTASPNR	7
IKFXZB	1iok_A	2.64e-02	MOD_SUMO_for_1	P29590	PML_HUMAN	Homo sapiens	Protein PML	Neoplasm	DisGeNET	1	118	489	492	FKLE	4
IKFXZB	1iok_A	2.64e-02	MOD_SUMO_for_1	P29590	PML_HUMAN	Homo sapiens	Protein PML	Immune System Disease	DisGeNET	2	37	489	492	FKLE	4

5.2 Análisis de la Base de Datos de Mimotopos

5.2.1 Estructura y Resumen de la Base de Datos de Mimotopos

El pre-procesamiento de esta base de datos ha sido similar al realizado en *1iok_A.xlsx*, consistiendo en eliminar la variable *PDB*, al ser todas las proteínas análogas a la estructura *1iok_A*, en trasformar la variable *E-Value* en variable numérica y en cambiar todas las variables de tipo *carácter* en variables de tipo *factor*.

La base de datos se estructura en **595 anotaciones**, una por cada fuente asociada, compuesta cada una de ellas de **15 variables**. Las variables incluidas en cada una de las observaciones son:

- *Mimotopo*: Variable factorial que indica el mimotopo a partir del cual se ha realizado el análisis comparativo de motivos.
- *E-Value*: Valor estadístico que indica el número de alineamientos que se espera para una puntuación (score) X (o superior) en el análisis comparativo.

- *Uniprot*: Variable factorial que indica el código Uniprot de la proteína análoga a la secuencia consenso de epítopos de las proteínas con estructura *1iok_A*.
- *Nombre*: Variable factorial que indica el nombre de la proteína humana análoga a la secuencia consenso de epítopos de las proteínas con estructura *1iok_A*.
- *Proteína*: Variable factorial que indica el tipo de proteína a la cual pertenece la proteína humana análoga.
- *Patología*: Variable factorial que indica el tipo de patología asociado a la proteína humana análoga.
- *Fuente*: Variable factorial que indica la fuente donde se han obtenidos las patologías asociadas a la proteína humana análoga o, en el caso de estar repetida la fuente, el origen de la observación (*FASTM* o *Uniprot*).
- *Nº Fuente*: Variable factorial que indica el número de fuente asociado al gen al que pertenece la proteína humana análoga.
- *Nº Referencia*: Variable factorial que indica el número de patologías asociadas a la proteína presente en la Fuente de origen.
- *Start*: Variable numérica que indica el punto inicial de similitud de los epítopos en la proteína humana análoga.
- *End*: Variable numérica que indica el punto final de similitud de los epítopos en la proteína humana análoga.

A partir del análisis comparativo se han obtenido **41416 referencias de enfermedades asociadas** que pertenecen a **158 proteínas** de **30 organismos** diferentes formando parte de **145 tipos de proteína diferentes**, agrupadas en **13 motivos proteicos**.

Tab 14: Estructura de la base de datos Mimotopos.xlsx

```
'data.frame':   595 obs. of  15 variables:
Mimotopo      : Factor w/ 3 levels
Motivo        : Factor w/ 13 levels
UniProt       : Factor w/ 158 levels
Nombre       : Factor w/ 158 levels
Organismo     : Factor w/ 30 levels
Proteína      : Factor w/ 145 levels
Patología     : Factor w/ 7 levels
Fuente        : Factor w/ 4 levels
Peptide       : Factor w/ 191 levels
```

Distribución de Anotaciones por Mimotopo

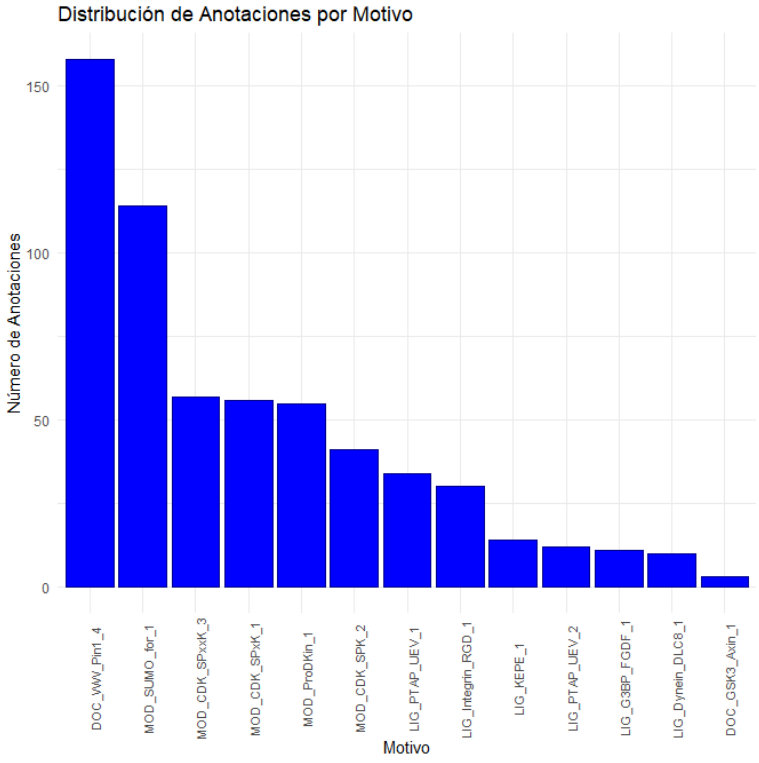
Analizando los datos por mimotopo se puede observar que:

- El primer mimotopo analizado, *IKFXZB* cuenta con **154 anotaciones**. Este mimotopo presenta **6** motivos análogos, siendo el motivo *MOD_SUMO_for_1* el que tiene mayor número de anotaciones **114 (72,61% del total de anotaciones del mimotopo)**.

- El segundo mimotopo analizado, *G*, consta de **1** único motivo, *LIG_Integrin_RGD_1*, con **30 anotaciones**.

- Finalmente, el tercer mimotopo analizado, *GXPX*, supone el **68,95% (411)** de las anotaciones totales de la base de datos y presenta **7** motivos análogos, de los cuales *DOC_WW_Pin1_4* es el mayoritario, con **158 anotaciones** (**38,44% del total** de anotaciones del mimotopo).

Fig. 50: Distribución de anotaciones de organismos patológicos por motivo proteico



Tab 15: Número de Proteínas por Motivo Proteico en Mimotopos.xlsx

IKFXZB

LIG_G3BP_FGDF_1	*DOC_GSK3_Axin_1*	*LIG_Dynein_DLC8_1*	*LIG_KEPE_1*
11	*3*	*10*	*14*
MOD_ProDKin_1	*MOD_SUMO_for_1*		
2	*114*		

G

LIG_Integrin_RGD_1			
30			

GXPX

DOC_WW_Pin1_4	*LIG_PTAP_UEV_1*	*LIG_PTAP_UEV_2*	*MOD_CDK_SPK_2*
158	*34*	*12*	*41*
MOD_CDK_SPxK_1	*MOD_CDK_SPxxK_3*	*MOD_ProDKin_1*	
56	*57*	*53*	

5.2.2 Análisis de Organismos Patógenos con Motivos Proteicos Análogos a los Mimotopos de 1iok_A

5.2.2.1 Tipología de las Proteínas de Organismos Patógenos Presentes en la Base de Datos de Mimotopos

Uno de los objetivos del TFM es intentar encontrar relaciones entre los procesos de infección de organismos patógenos y el desarrollo de enfermedades, principalmente asociados a *patologías del sistema inmune* y *neoplasmas*.

Con la obtención de la lista de motivos proteicos análogos a los mimotopos de *1iok_A* se ha observado que varios organismos patógenos presentan motivos análogos implicados en procesos de infección que coinciden con diversas proteínas asociadas a enfermedades de diversa índole.

En esta sección del TFM se analizan qué organismos patógenos presentan estos motivos.

Un primer detalle interesante del análisis es que todos los mimotopos analizados presentan algún organismo patógeno con motivos análogos en proteínas implicadas en procesos de infección.

Así tenemos que en el mimotopo *IKFXZB* como motivos análogos hay *LIG_G3BP_FGDF_1*, del que 9 de las 10 proteínas que lo componen son poliproteínas de virus con funciones cruciales en los procesos de infección, *MOD_SUMO_for_1* que presenta dos proteínas de organismos patógenos, la bacteria *Anaplasma phagocytophilum* y el virus *Human herpesvirus 5* y *LIG_Dynein_DLC8*_1 que incluye dentro del grupo una proteína de *Rabies virus*.

Cabe destacar que ciertos estudios sugieren que la *fosfoproteína P* del virus de la rabia (*P22363*), junto con otros cuatro productos truncados en el extremo amino (P2, P3, P4, P5), interactúa directamente con la *leucemia promielocítica* inducida por interferón (*LMP*) al reorganizar los cuerpos nucleares (NB) de la *LMP*. **[58]**

Estos cuerpos nucleares**,** podrían desempeñar un papel en la respuesta al *interferón* (*IFN*) y, aunque la función de los *NB PML* aún no está clara, algunos resultados indican que pueden representar objetivos preferenciales para las infecciones virales ya que la *LMP* podría desempeñar un papel en el mecanismo de la acción antiviral de los *IFN* y los virus que requieren de la maquinaria celular para su replicación intentan contrarrestar la acción del *IFN* inhibiendo la señalización del *IFN*. **[59]**

Uno de los motivos análogos al mimotopo *G* es el motivo *LIG_Integrin_isoDGR_2*. Éste motivo incluye las integrinas, que son receptores mediadores de la adhesión celular presentes en todos los metazoos. Todas las células humanas expresan uno o más de los 24 tipos de integrinas que componen la membrana plasmática **[60]** y que actúan como mediadoras de señales entre el medio intracelular y el extracelular **[61, 62, 63]**.

Debido a su papel fundamental en la comunicación celular, su la mala regulación endógena puede implicar varias enfermedades como la enfermedad de Alzheimer **[64]**, la fibrosis quística **[65]**, el trastorno del espectro autista, la esquizofrenia **[66]** y el cáncer **[67]**, ya que las integrinas juegan un papel fundamental en la angiogénesis y la metástasis.

Esta importancia las convierte en dianas de varios virus, como *el virus de la fiebre aftosa*, el *VIH*, el *virus del Nilo Occidental* o el *VPH-16* **[68, 69]** y de otros organismos patógenos, incluidas bacterias y eucariotas, que tienen motivos similares a *RGD* incrustados en sus proteínas para unirse a las integrinas en la superficie de la célula huésped y ayudar a la entrada de la célula.

Finalmente, el motivo *GXPX* presenta dos motivos, *LIG_PTAP_UEV_1* y *MOD_ProDKin_1*, con proteínas en procesos de infección de virus como *Human T-cell leukemia virus 1*, *Human immunodeficiency virus* y *Hepatitis B virus.*

5.2.2.2 Estructura y Resumen del Análisis de las Anotaciones Pertenecientes a Organismos Patógenos

A partir de la base de datos de mimotopos se puede observar que **13 de las 34 anotaciones** indican que la proteína analizada pertenece a algún tipo de poliproteína y

que dentro de la lista están representados tres grandes grupos de organismos diferentes: virus, bacterias y protistas.

Finalmente, se observa que casi la mitad de los organismos patógenos estudiados, un **47,05% (16)**, tienen el motivo *GXPX* en alguna de sus proteínas, principalmente el motivo *LIG_PTAP_UEV_1*, que presenta **14 anotaciones**.

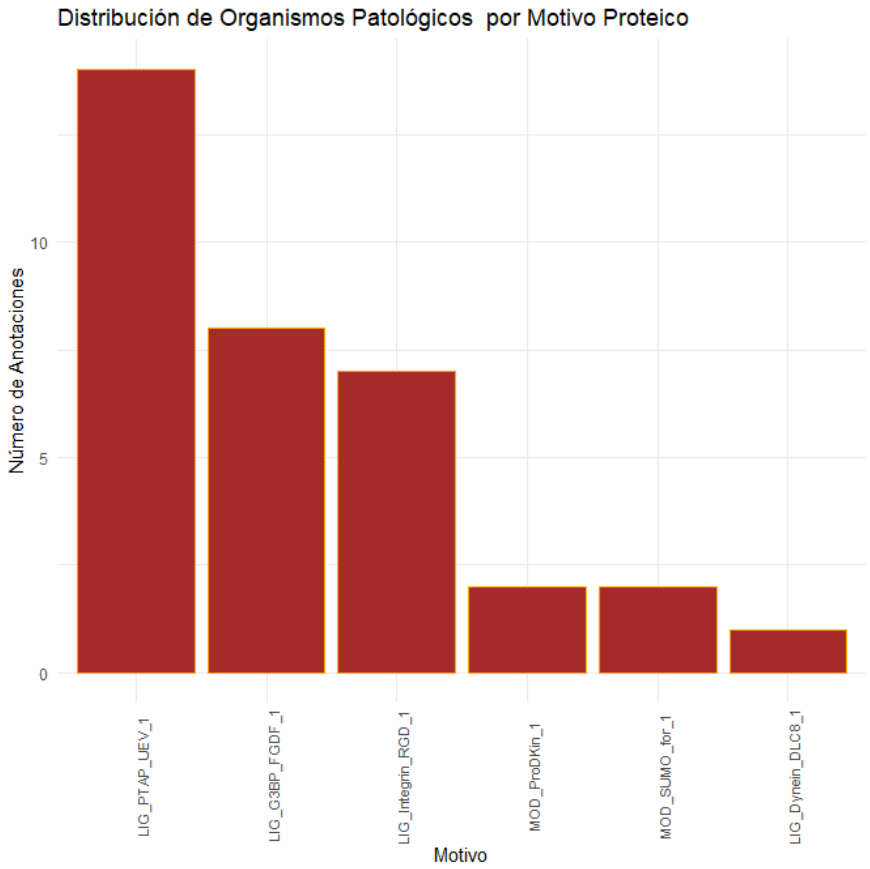
Tab 16: Sumario de Proteínas de organismos patógenos presentes en Mimotopos.xlsx

Mimotopo		Motivo		UniProt		Nombre	
G	: 7	LIG_PTAP_UEV_1	:14	D0V559	: 2	D0V559_TOXGO	: 2
GXPX	:16	LIG_G3BP_FGDF_1	: 8	P03317	: 2	POLN_CHIK3	: 2
IKFXZB	:11	LIG_Integrin_RGD_1	: 7	P08411	: 2	POLN_SFV	: 2
		MOD_ProDKin_1	: 2	Q5XXP4	: 2	POLN_SINDV	: 2
		MOD_SUMO_for_1	: 2	H2VFV1	: 1	AMPA_ANAPZ	: 1
		LIG_Dynein_DLC8_1	: 1	O25272	: 1	CAPSP_ADE02	: 1
		(Other)	: 0	(Other)	:24	(Other)	:24

Organismo		Proteína	
Chikungunya virus	: 2	Polyprotein P1234	: 6
Human T-cell leukemia virus	: 12	Gag polyprotein	: 4
Semliki forest virus	: 2	Genome polyprotein	: 3
Sindbis virus	: 2	Filamentous hemagglutinin	: 2
Toxoplasma gondii	: 2	Major DNA-binding protein	: 2
Human herpesvirus 5	: 1	Rhoptry neck protein 5	: 2
(Other)	:23	(Other)	:15

5.2.2.3 Distribución de Anotaciones Pertenecientes a Organismos Patógenos por Motivo Proteico

Fig. 51: Distribución de anotaciones de organismos patológicos por motivo proteico



Tab 17: Organismos con Motivos Análogos al Mimotopo IKFXZB

	Motivo	UniProt	Nombre	Organismo
1	LIG_G3BP_FGDF_1	P36384	DNBI_HHV2	Human herpesvirus 2
2	LIG_G3BP_FGDF_1	P04296	DNBI_HHV11	Human herpesvirus 1 (strain 17)
3	LIG_G3BP_FGDF_1	P03317	POLN_SINDV	Sindbis virus
4	LIG_G3BP_FGDF_1	P03317	POLN_SINDV	Sindbis virus
5	LIG_G3BP_FGDF_1	Q5XXP4	POLN_CHIK3	Chikungunya virus (strain 37997)
6	LIG_G3BP_FGDF_1	Q5XXP4	POLN_CHIK3	Chikungunya virus (strain 37997)
7	LIG_G3BP_FGDF_1	P08411	POLN_SFV	Semliki forest virus
8	LIG_G3BP_FGDF_1	P08411	POLN_SFV	Semliki forest virus
12	MOD_SUMO_for_1	Q2GIB5	AMPA_ANAPZ	Anaplasma phagocytophilum (strain HZ)
82	MOD_SUMO_for_1	P13202	VIE1_HCMVA	Human herpesvirus 5 strain AD169
145	LIG_Dynein_DLC8_1	P22363	PHOSP_RABVC	Rabies virus (strain CVS-11) (RABV)

Tab 18: Organismos con Motivos análogos al Mimotopo G

	Motivo	UniProt	Nombre	Organismo
160	LIG_Integrin_RGD_1	S0HPF7	PILY1_PSEAW	Pseudomonas aeruginosa (strain PAK)
161	LIG_Integrin_RGD_1	O25272	O25272_HELPY	Helicobacter pylori
162	LIG_Integrin_RGD_1	P12255	FHAB_BORPE	Bordetella pertussis
167	LIG_Integrin_RGD_1	P03276	CAPSP_ADE02	Human adenovirus C serotype 2
168	LIG_Integrin_RGD_1	P03305	POLG_FMDVO	Foot-and-mouth disease virus
169	LIG_Integrin_RGD_1	Q66578	POLG_HPE1H	Human parechovirus 1
170	LIG_Integrin_RGD_1	P21404	POLG_CXA9	Coxsackievirus A9

Tab 19: Organismos con Motivos análogos al Mimotopo GXPX

	Motivo	UniProt	Nombre	Organismo
235	LIG_PTAP_UEV_1	D0V559	D0V559_TOXGO	Toxoplasma gondii
236	LIG_PTAP_UEV_1	D0V559	D0V559_TOXGO	Toxoplasma gondii
237	LIG_PTAP_UEV_1	P69616	ORF3_HEVBU	Hepatitis E virus genotype 1
238	LIG_PTAP_UEV_1	P27588	NCAP_MABVM	Lake Victoria marburgvirus
239	LIG_PTAP_UEV_1	P08363	VP8_BTV10	Bluetongue virus 10
240	LIG_PTAP_UEV_1	P14350	GAG_FOAMV	Human spumaretrovirus
241	LIG_PTAP_UEV_1	P03345	GAG_HTL1A	Human T-cell leukemia virus 1 (HTLV-1)
254	LIG_PTAP_UEV_1	P03519	MATRX_VSIVA	Vesicular stomatitis Indiana virus
255	LIG_PTAP_UEV_1	O73557	Z_LASSJ	Lassa virus (LASV)
256	LIG_PTAP_UEV_1	Q05128	VP40_EBOZM	Zaire ebolavirus
257	LIG_PTAP_UEV_1	P18095	GAG_HV2BE	Human immunodeficiency virus type 2
272	LIG_PTAP_UEV_1	P04591	GAG_HV1H2	Human immunodeficiency virus type 1
405	LIG_PTAP_UEV_1	P69713	X_HBVA3	Hepatitis B virus genotype A2
406	LIG_PTAP_UEV_1	P03409	TAX_HTL1A	Human T-cell leukemia virus 1 (HTLV-1)
443	MOD_ProDKin_1	P13128	TACY_LISMO	Listeria monocytogenes serovar
444	MOD_ProDKin_1	H2VFV1	INCA_CHLCV	Chlamydophila caviae

5.2.3 Análisis de Patologías Asociadas a Motivos Proteicos Análogos a los Mimotopos de 1iok_A

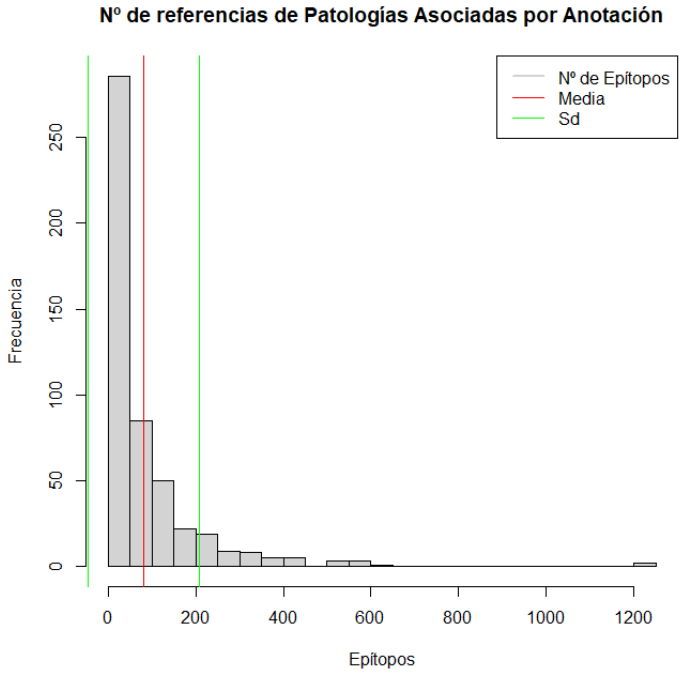
5.2.3.1 Estructura y Resumen del Análisis de Patologías Asociadas a Motivos Proteicos Análogos a los Mimotopos de 1iok_A

El objetivo de esta base de datos es obtener una lista de proteínas con patologías asociadas, por lo que el siguiente análisis realizado ha sido observar la distribución del número de referencias por proteína.

La base de datos presenta **253** anotaciones con patologías asociadas a *neoplasmas* y **245** anotaciones con patología asociadas a *enfermedades del sistema inmune*.

Analizando la distribución de referencias patológicas por proteína se puede apreciar que más de la mitad de anotaciones presentan menos de **37,5** referencias de patologías asociadas pese a tener una media de **87,1** referencias por anotación.

Fig. 52: Histograma de frecuencias del número de referencias de patologías por anotación



Distribución de Referencias de Patologías Asociadas por Mimotopo

De un total de **40946** referencias de patologías asociadas, **26825 (65,51%)** pertenecen al mimotopo *GXPX*, **11817 (28,86%)** pertenecen a *IKFXZB* y el resto pertenece a *G*

Comparando fuentes y patologías se puede observar que de media hay más referencias de *neoplasmas*, encontrándose mayoritariamente en ellas los outliers, y presentando una distribución mucho más amplia, que referencias a *enfermedades del sistema inmune*.

Si se compara el número de referencias por tipo de fuente, hay un media de referencias ligeramente mayor y una distribución de valores algo más amplia, especialmente en el tercer cuartil, en *Open Targets*.

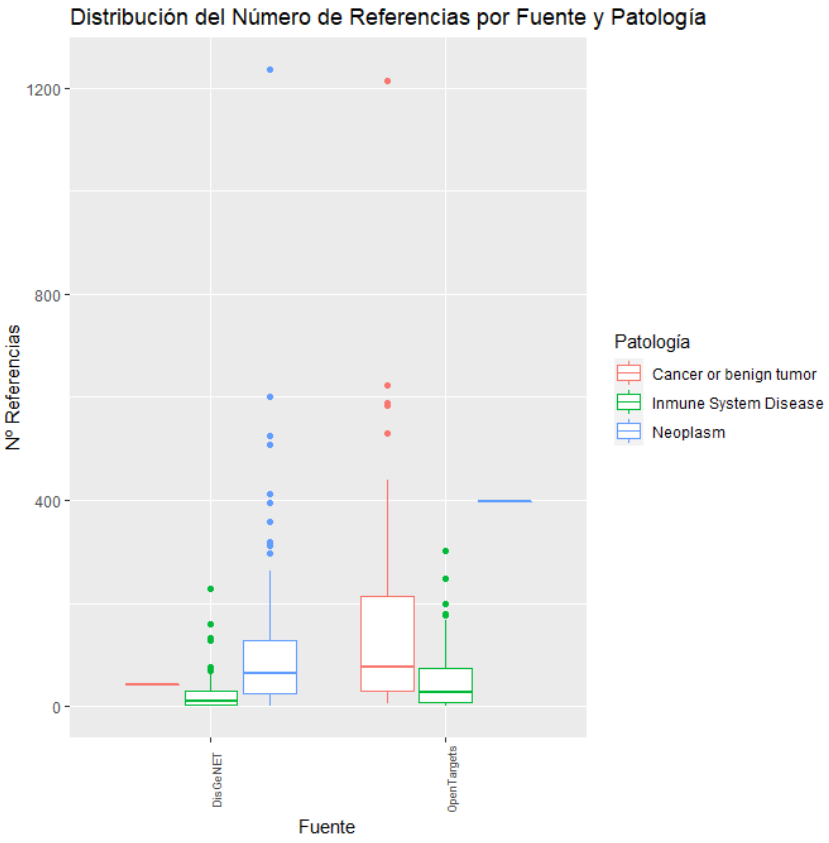


Fig. 53: Boxplot de distribución del número de referencias de patologías por fuente y por tipo de patología

Distribución de Referencias de Patologías Asociadas por Fuente y Mimotopo

Se observa una distribución de referencias de patologías entre mimotopos homogénea, si se dividen las anotaciones por fuente, excepto en la media de referencias del mimotopo *G* en *Open Targets* que, aunque presenta una amplitud de distribución similar, es más alta que el resto, lo que quizás se deba al pequeño número de anotaciones de este tipo que presenta el mimotopo *G*.

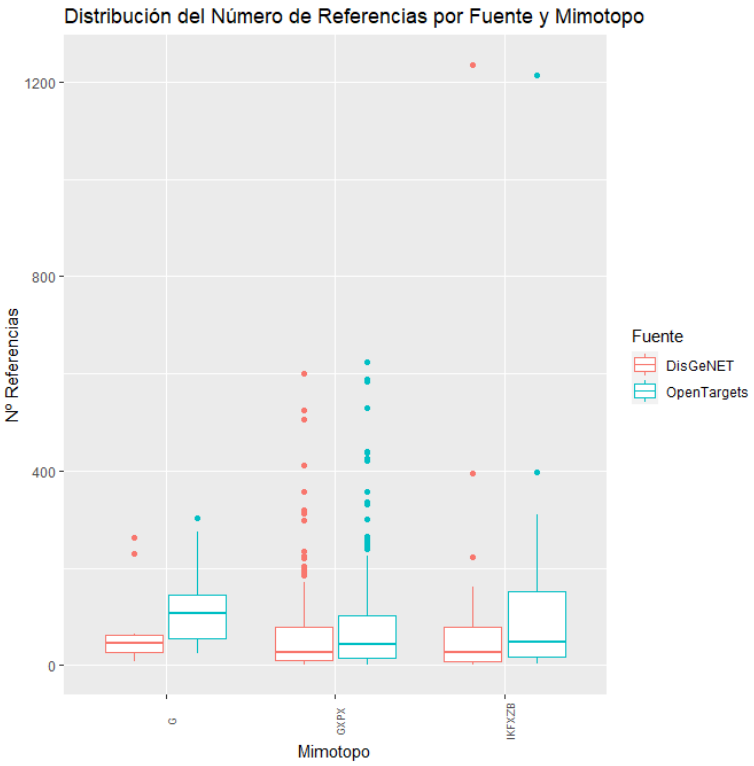


Fig. 54: Boxplot de distribución del número de referencias de patologías por fuente y por mimotopo

Distribución de Referencias de Patologías Asociadas por Patología y Mimotopo

Se observa una situación similar si se compara la distribución de referencias por patologías, con una media mayor de referencias en las anotaciones del mimotopo *G* pero con una amplitud de valores de referencias asociadas a *neoplasmas* menor.

Comparando *GXPX* e *IKFXZB* se observa una distribución de valores similar, aunque con una amplitud de valores y media de patologías asociadas a *cánceres y tumores benignos* ligeramente mayores en *GXPX*

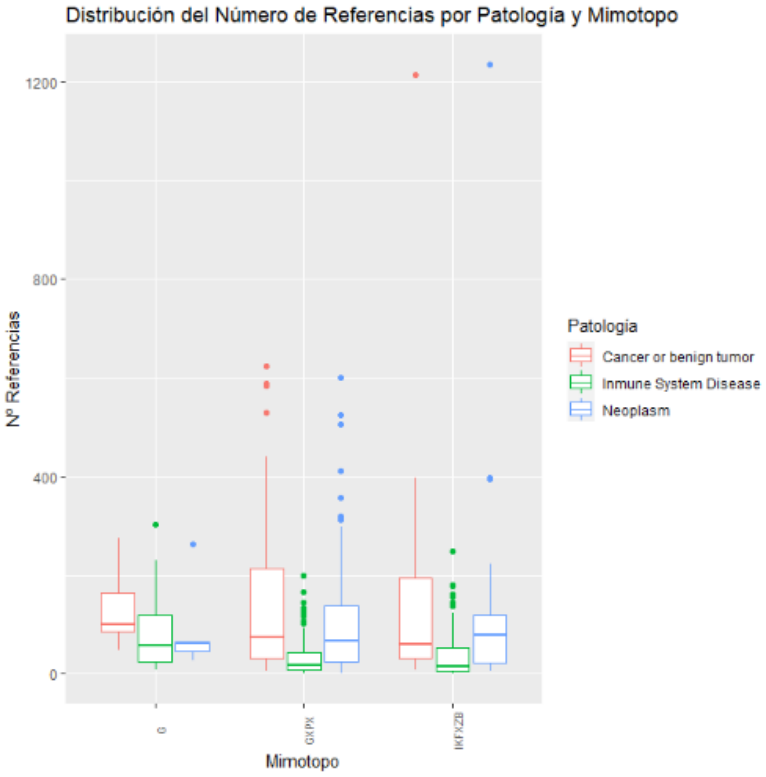


Fig. 55: Boxplot de distribución del número de referencias de patologías por fuente y por mimotopo

5.2.3.2 Distribución de Anotaciones por Tipo de Fuente

Distribución de Referencias de Patologías Asociadas Extraídas de DisGeNet

Esta parte del análisis consiste en comparar las referencias de patologías asociadas de ambas fuentes con el fin de inferir si ambas presentan datos similares o, en cambio, muestran diferencias significativas.

En las anotaciones obtenidas con *DisGeNet* se observa que el mimotopo *GXPX* es el que presenta un mayor número de motivos y una distribución más amplia del número de referencias.

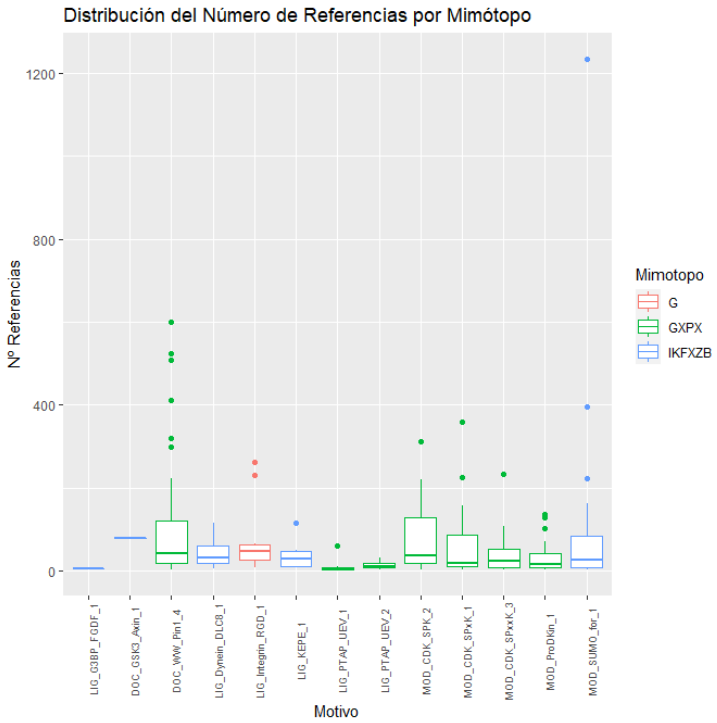


Fig. 56: Boxplot de distribución del número de referencias por mimotopo extraídas de DisGeNet

A nivel individual, la proteína con más referencias es *P04637*, del motivo *MOD_SUMO_for_1*, aunque las cuatro proteínas siguientes con más referencias pertenecen al motivo *DOC_WW_Pin1_4*.

Tab 20: Proteínas con mayor número de referencias en DisGeNet

	Mimotopo	Motivo	UniProt	Nº Referencias
1	IKFXZB	MOD_SUMO_for_1	P04637	1235
2	GXPX	DOC_WW_Pin1_4	P35222	600
3	GXPX	DOC_WW_Pin1_4	P31749	525
4	GXPX	DOC_WW_Pin1_4	P01106	507
5	GXPX	DOC_WW_Pin1_4	P40763	412
6	IKFXZB	MOD_SUMO_for_1	Q16665	394
7	GXPX	MOD_CDK_SPxK_1	P25054	358
8	GXPX	DOC_WW_Pin1_4	P46531	320
9	GXPX	MOD_CDK_SPK_2	P10275	312
10	GXPX	DOC_WW_Pin1_4	P46527	298

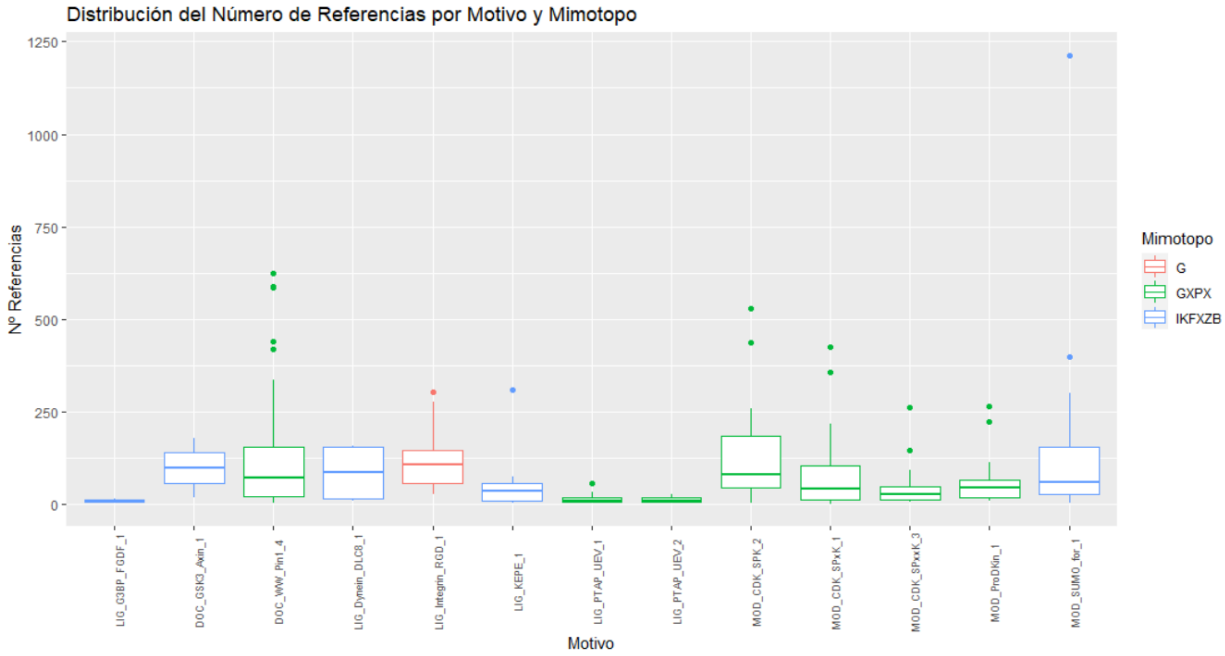
Distribución de Referencias de Patologías Asociadas Extraídas de Open Targets

En las referencias de patologías asociadas extraídas de *Open Targets*, en cambio, los tres mimotopos presentan patrones de distribución del número de referencias de patologías asociadas similares.

En los tres grupos se combinan motivos con medias altas del *nº de referencias* y distribuciones amplias y motivos con una distribución del *nº de referencias* más limitada y medias bajas.

Se puede considerar que la variabilidad del número de referencias por motivo es bastante alta en los tres mimotopos.

Fig. 57: Boxplot de distribución del número de referencias por mimotopo extraídas de Open Targets



Pese a las diferencias en la distribución del *nº de referencias* por motivo, cuando se analiza individualmente esa distribución se puede observar que el ranking de proteínas y el número de referencias individual es muy similar.

Estos resultados dan a entender que, aunque en las proteínas más estudiadas, o con mayor número de anotaciones, los resultados son similares, en el caso de las proteínas con menor número de referencias, quizás menos estudiadas, *Open Targets* presenta un mayor nivel de sensibilidad y, como consecuencia, de anotaciones.

Tab 21: Proteínas con mayor número de referencias en Open Targets

	Mimotopo	Motivo	UniProt	Nº Referencias
1	IKFXZB	MOD_SUMO_for_1	P04637	1214
2	GXPX	DOC_WW_Pin1_4	P31749	624
3	GXPX	DOC_WW_Pin1_4	P35222	589
4	GXPX	DOC_WW_Pin1_4	P01106	585
5	GXPX	MOD_CDK_SPK_2	P06400	530
6	GXPX	DOC_WW_Pin1_4	P46531	440
7	GXPX	MOD_CDK_SPK_2	P10275	438
8	GXPX	MOD_CDK_SPxK_1	P25054	426
9	GXPX	DOC_WW_Pin1_4	P40763	420
10	IKFXZB	MOD_SUMO_for_1	Q16665	398

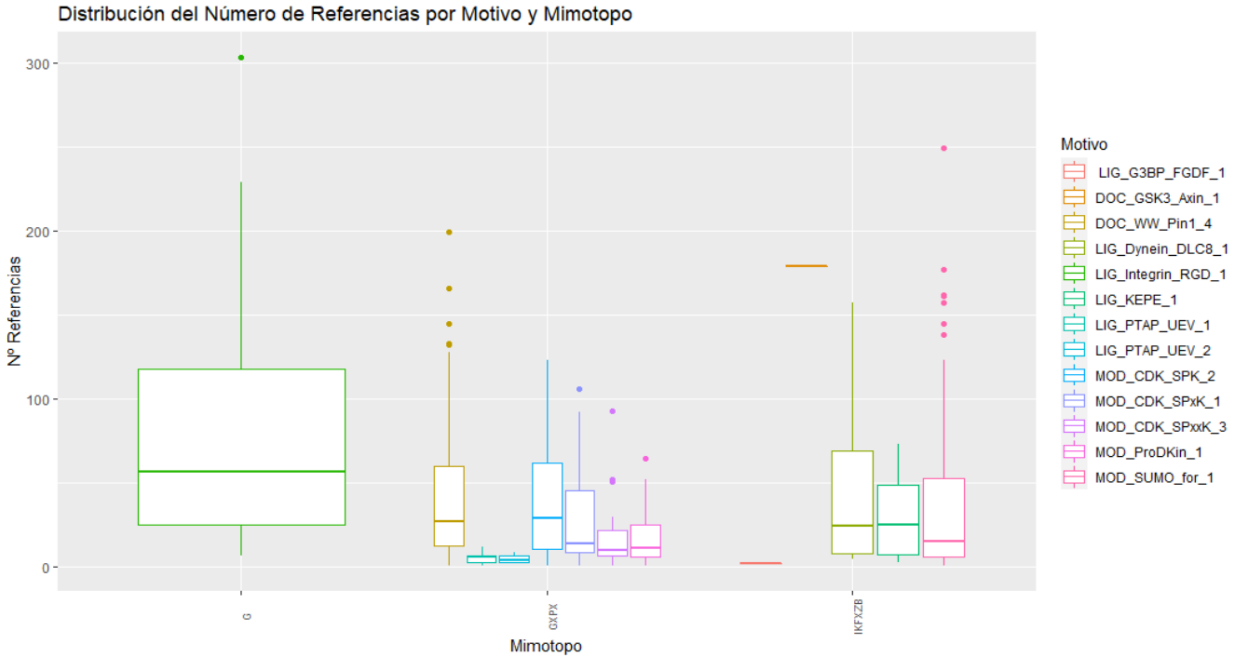
5.2.3.3 Distribución de Anotaciones por Tipo de Patología

Finalmente, se ha realizado el análisis de la distribución de anotaciones de patologías asociadas en función de tipo de patología analizada, que quizás sea el más significativo de los análisis realizados.

Distribución de Anotaciones con Patologías Asociadas al Sistema Inmune

A nivel de las *ESI*, se puede ver que las proteínas del motivo *LIG_Integrin_RGD_1* son las que presentan una mayor media y distribución de referencias de patologías asociadas a *ESI*, tanto globalmente como a nivel individual, en que la proteína *P02751* es la que presenta un mayor número de referencias.

Fig. 58: Distribución del número de referencias de patologías asociadas al sistema inmune por motivo y mimotopo

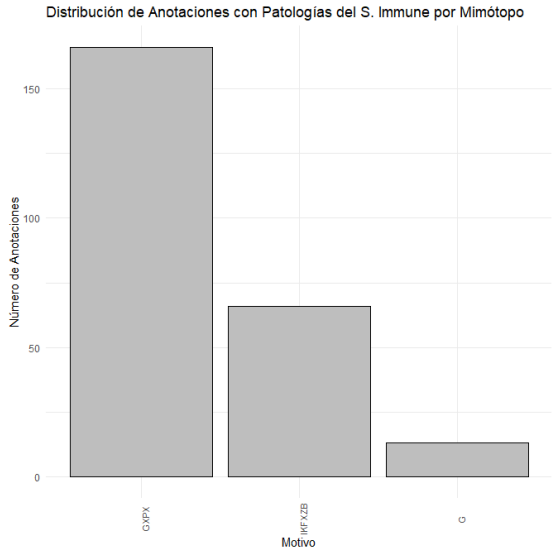


Tab 22: Proteínas con mayor número de referencias de patologías asociadas al sistema inmune

	Mimotopo	Motivo	UniProt	Nº Referencias
1	G	LIG_Integrin_RGD_1	P02751	303
2	IKFXZB	MOD_SUMO_for_1	P04637	249
3	G	LIG_Integrin_RGD_1	P02751	229
4	GXPX	DOC_WW_Pin1_4	P01106	199
5	IKFXZB	DOC_GSK3_Axin_1	O15169	179
6	IKFXZB	MOD_SUMO_for_1	P04150	177
7	GXPX	DOC_WW_Pin1_4	P40763	166
8	IKFXZB	MOD_SUMO_for_1	P10242	162
9	IKFXZB	MOD_SUMO_for_1	P04637	161
10	IKFXZB	MOD_SUMO_for_1	Q15596	157

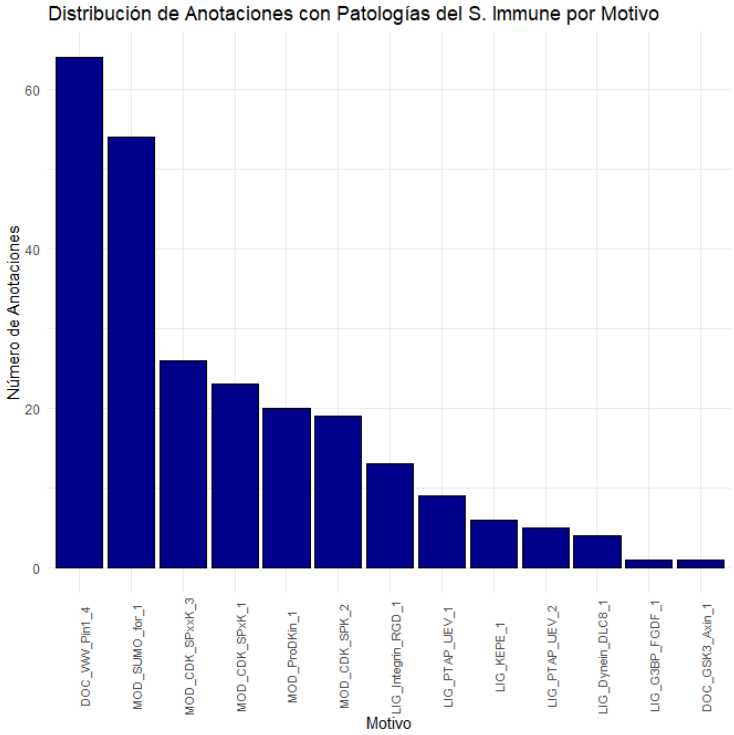
Cabe destacar, que pese a la importancia de las anotaciones relativas a *ESI* de las proteínas de *G* es alta, a nivel global este mimotopo es el que presenta un menor número de referencias, debido al bajo número de proteínas existentes en la base de datos de este mimotopo.

Fig. 59: Distribución del número de referencias de patologías asociadas al sistema inmune por mimotopo



Si se analiza la distribución de patologías asociadas a *ESI*, se observa que los motivos con más anotaciones de *ESI* son *DOC_WW_Pin1_4* y *MOD_SUMO_for_1*

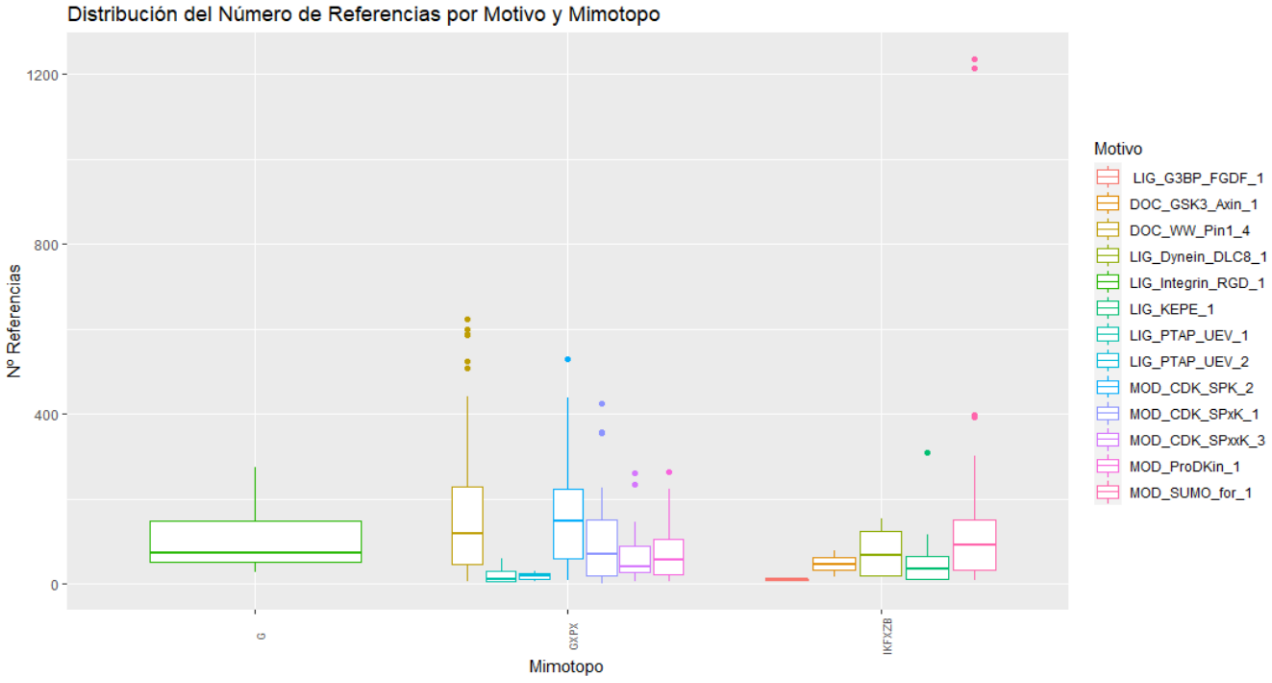
Fig. 60: Distribución de anotaciones con patologías asociadas al sistema inmune por motivo



Distribución de Anotaciones con Patologías Asociadas a Neoplasmas

Por otro lado, se puede ver que el mimotopo con mayor número de motivos con anotaciones asociadas a *neoplasmas* es *GXPX* que, además, presenta motivos con una media de referencias mayor a la del resto de motivos.

Fig. 61: Distribución del número de referencias de patologías asociadas al neoplasmas por motivo y mimotopo



Pese al mayor número de referencias globales de *GXPX*, la proteína con más referencias de patologías asociadas a neoplasmas es *P04637* del motivo *MOD_SUMO_for_1*, análogo al mimotopo *IKFXZB*.

Tab 23: Nº de Referencias de neoplasmas por proteína

	Mimotopo	Motivo	UniProt	Nº Referencias
1	IKFXZB	MOD_SUMO_for_1	P04637	1235
2	IKFXZB	MOD_SUMO_for_1	P04637	1214
3	GXPX	DOC_WW_Pin1_4	P31749	624
4	GXPX	DOC_WW_Pin1_4	P35222	600
5	GXPX	DOC_WW_Pin1_4	P35222	589
6	GXPX	DOC_WW_Pin1_4	P01106	585
7	GXPX	MOD_CDK_SPK_2	P06400	530
8	GXPX	DOC_WW_Pin1_4	P31749	525
9	GXPX	DOC_WW_Pin1_4	P01106	507
10	GXPX	DOC_WW_Pin1_4	P46531	440

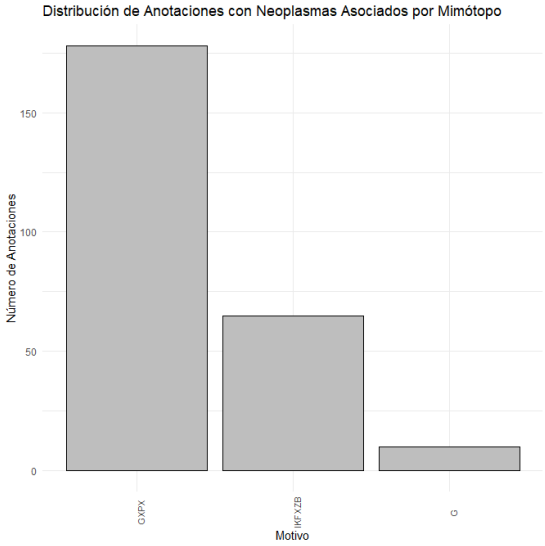
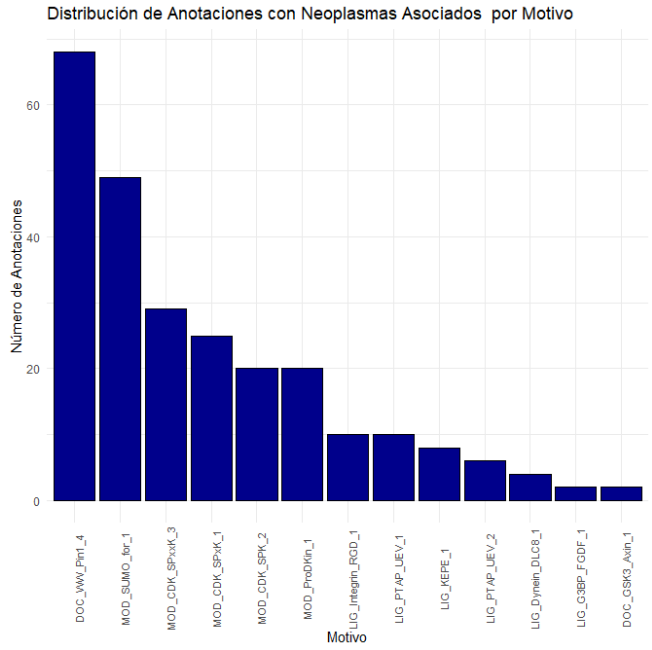


Fig. 62: Distribución del número de referencias de patologías asociadas al sistema inmune mimotopo

La distribución de anotaciones de motivos proteicos con *neoplasmas* asociados es muy similar, con pequeñas diferencias en el número de referencias y en el ranking de anotaciones bastante poco significativas.

Fig. 63: Distribución de anotaciones con patologías asociadas a neoplasmas por motivo



El último análisis realizado ha sido la extracción de las proteínas análogas a *IKFXZB* existentes en la base de datos, generando una lista de **33** proteínas.

Este procedimiento se ha realizado porqué se ha decidido realizar el análisis específico de patologías asociadas de este mimotopo únicamente a partir de los datos de *Open Targets*.

Se ha escogido esta fuente porqué la extracción de información es más fácil y presenta una mayor cantidad de referencias de patologías asociadas, especialmente de aquellas asociadas a *ESI*.

Tab 24: Proteínas con motivos análogos a IKFXZB

Anotación	Motivo	UniProt	Nombre
9	LIG_G3BP_FGDF_1	Q14694	UBP10_HUMAN
13	MOD_SUMO_for_1	Q16665	HIF1A_HUMAN
21	MOD_SUMO_for_1	P29590	PML_HUMAN
29	MOD_SUMO_for_1	Q02078	MEF2A_HUMAN
34	MOD_SUMO_for_1	Q9NSC2	SALL1_HUMAN
38	MOD_SUMO_for_1	Q14526	HIC1_HUMAN
42	MOD_SUMO_for_1	P61956	SUMO2_HUMAN
46	MOD_SUMO_for_1	P10242	MYB_HUMAN
51	MOD_SUMO_for_1	Q15596	NCOA2_HUMAN
57	MOD_SUMO_for_1	Q9UPW6	SATB2_HUMAN
62	MOD_SUMO_for_1	P56524	HDAC4_HUMAN
66	MOD_SUMO_for_1	P55854	SUMO3_HUMAN
70	MOD_SUMO_for_1	P23497	SP100_HUMAN
74	MOD_SUMO_for_1	Q02447	SP3_HUMAN
78	MOD_SUMO_for_1	Q13569	TDG_HUMAN
83	MOD_SUMO_for_1	P04637	P53_HUMAN
90	MOD_SUMO_for_1	P49716	CEBPD_HUMAN
94	MOD_SUMO_for_1	Q15744	CEBPE_HUMAN
98	MOD_SUMO_for_1	P06401	PRGR_HUMAN
102	MOD_SUMO_for_1	P49715	CEBPA_HUMAN
106	MOD_SUMO_for_1	Q92754	AP2C_HUMAN
110	MOD_SUMO_for_1	P27540	ARNT_HUMAN
114	MOD_SUMO_for_1	P17676	CEBPB_HUMAN
118	MOD_SUMO_for_1	P04150	GCR_HUMAN
123	MOD_SUMO_for_1	Q13547	HDAC1_HUMAN
128	MOD_SUMO_for_1	Q00613	HSF1_HUMAN
133	DOC_GSK3_Axin_1	O15169	AXIN1_HUMAN
136	LIG_Dynein_DLC8_1	Q15326	ZMY11_HUMAN
140	LIG_Dynein_DLC8_1	O43521	B2L11_HUMAN
146	LIG_KEPE_1	P14316	IRF2_HUMAN
150	LIG_KEPE_1	O14686	KMT2D_HUMAN
154	LIG_KEPE_1	P54845	NRL_HUMAN
157	LIG_KEPE_1	O15525	MAFG_HUMAN

5.3 Análisis Estadístico de la Base de Datos de Patologías de las Proteínas Humanas con Epítopos Análogos a IKFXZB

5.3.1 Estructura y Resumen de la Base de Datos *IKFXZB_Disease*

Para poder analizar más en profundidad el tipo de patologías asociadas, se ha creado otra base de datos, denominada *IKFXZB_Disease.xlsx*, en la que se han incluido todas las referencias de patologías asociadas a cada proteína con motivos análogos a *IKFXZB*.

De igual forma que en los análisis anteriores, se han convertido en *factor* todas las variables no numéricas de la base de datos.

La base de datos se estructura en **6.918 anotaciones** compuestas de **4 variables**. Las variables incluidas en cada una de las observaciones son:

- *Motivo*: Variable factorial que indica el motivo proteico al que pertenece la proteína humana análoga.
- *Proteína:* Variable factorial que indica el código *Uniprot* de la proteína humana análoga.
- *Patología*: Variable factorial que indica la patología asociada a la proteína humana análoga.
- *Tipología*: Variable factorial que indica el tipo de patología asociado a la proteína humana análoga.

Fig. 64: Cabecera de la Base de datos IKFXZB_Disease

Motivo	Proteína	Patología	Tipología
LIG_G3BP_FGDF_1	Q14694	breast adenocarcinoma	Neoplasma
LIG_G3BP_FGDF_1	Q14694	esophageal cancer	Neoplasma
LIG_G3BP_FGDF_1	Q14694	non-small cell lung carcinoma	Neoplasma
LIG_G3BP_FGDF_1	Q14694	neoplasm	Neoplasma
LIG_G3BP_FGDF_1	Q14694	cancer	Neoplasma
LIG_G3BP_FGDF_1	Q14694	glioblastoma multiforme	Neoplasma
LIG_G3BP_FGDF_1	Q14694	adrenal gland neoplasm	Neoplasma
LIG_G3BP_FGDF_1	Q14694	hepatocellular carcinoma	Neoplasma
LIG_G3BP_FGDF_1	Q14694	gastric carcinoma	Neoplasma
LIG_G3BP_FGDF_1	Q14694	adrenal gland pheochromocytoma	Neoplasma
LIG_G3BP_FGDF_1	Q14694	adrenal cortex carcinoma	Neoplasma
LIG_G3BP_FGDF_1	Q14694	acute myeloid leukemia	Neoplasma
LIG_G3BP_FGDF_1	Q14694	squamous cell lung carcinoma	Neoplasma
LIG_G3BP_FGDF_1	Q14694	lung adenocarcinoma	Neoplasma
LIG_G3BP_FGDF_1	Q14694	adrenocortical adenoma	Neoplasma
LIG_G3BP_FGDF_1	Q14694	acute myeloid leukemia	Inmune System Disease
LIG_G3BP_FGDF_1	Q14694	juvenile der	Inmune System Disease
MOD_SUMO_for_1	Q16665	kidney neoplasm	Neoplasma
MOD_SUMO_for_1	Q16665	breast ductal adenocarcinoma	Neoplasma
MOD_SUMO_for_1	Q16665	lymphoid neoplasm	Neoplasma
MOD_SUMO_for_1	Q16665	brain glioblastoma	Neoplasma
MOD_SUMO_for_1	Q16665	Ovarian Endometrioid Adenocarcinoma with Squamous Differentiation	Neoplasma
MOD_SUMO_for_1	Q16665	carcinoma of liver and intrahepatic biliary tract	Neoplasma

A partir del análisis comparativo se han obtenido **1730 patologías asociadas** a **32 proteínas** pertenecientes a **5 motivos proteicos**.

Del total de observaciones, el **74,75% (5102)** se corresponden a *Neoplasmas* mientras que el resto son *Enfermedades del Sistema Inmune*.

La proteína con más anotaciones es *P04637* con **1447 anotaciones**, que pertenece al motivo *MOD_SUMO_for_1*, que, a su vez, agrupa el **84,55% (5849)** de las anotaciones totales de la base de datos, siendo 4 de las 5 patologías asociadas con más anotaciones diferentes tipos de *leucemia*.

Tab 25: Sumario de la base de datos IKFXZB_Disease

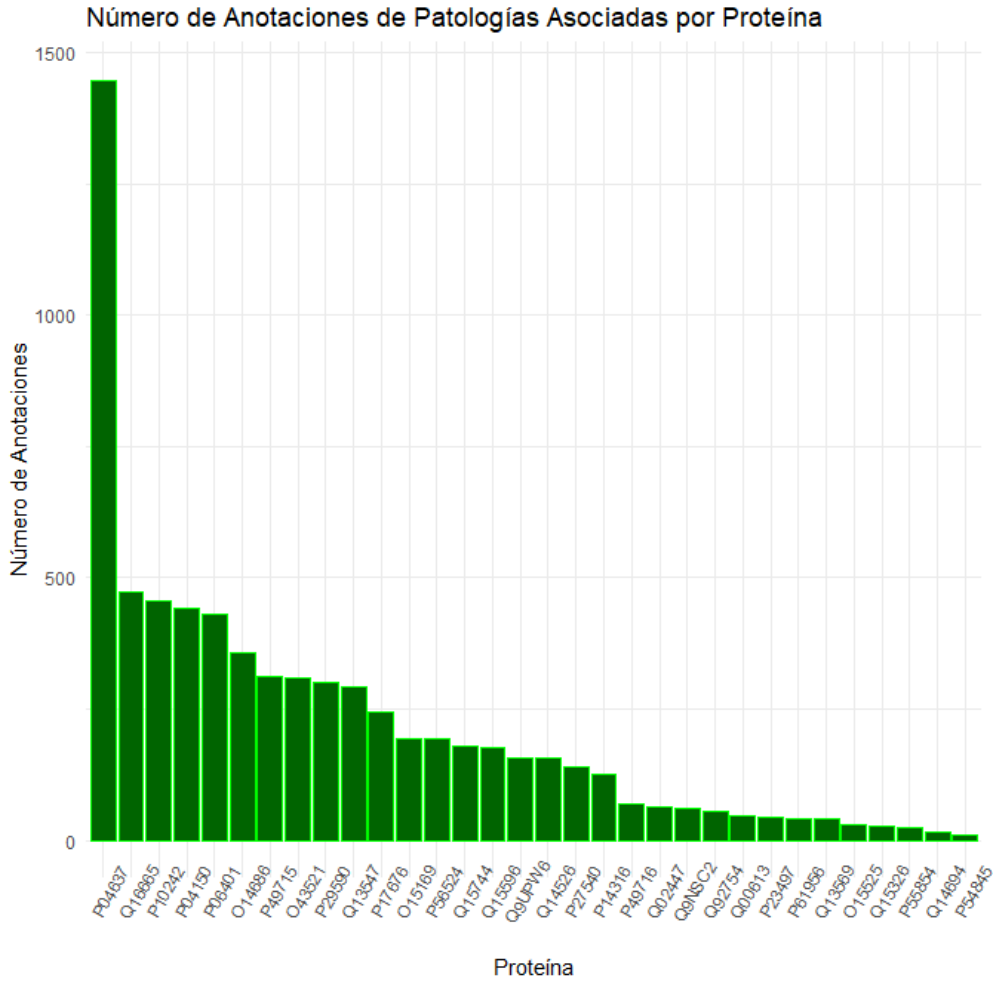
```
tibble [6,918 x 4]
Motivo   : Factor w/ 5 levels
Proteína : Factor w/ 32 levels
Patología: Factor w/ 1730 levels
Tipología: Factor w/ 2 levels

          Motivo                      Proteína              Patología
DOC_GSK3_Axin_1  : 194         P04637 :1447    acute myeloid leukemia      :  52
LIG_Dynein_DLC8_1: 336         Q16665 : 474    chronic lymphocytic leukemia:  50
LIG_G3BP_FGDF_1  :  17         P10242 : 455            acute lymphoblastic leukemia:  42
LIG_KEPE_1       : 522         P04150 : 443    multiple myeloma            :  42
MOD_SUMO_for_1   :5849         P06401 : 430    chronic myelogenous leukemia:  40
                               O14686 : 358    (Other)                     :6691
                               (Other):3311    NA's                        :   1

          Tipología
Inmune System Disease:1816
Neoplasma            :5102

     Proteína            Frecuencia
1    P04637              1447
2    Q16665              474
3    P10242              455
4    P04150              443
5    P06401              430
6    O14686              358
```

Fig. 65: Distribución de anotaciones de patologías asociadas por proteína



5.3.2 Distribución de proteínas por Patología Asociada

5.3.2.1 Distribución de Proteínas con Neoplasmas Asociados

Se ha obtenido una lista de **5105 referencias** de *neoplasmas* asociados a **32 proteínas**.

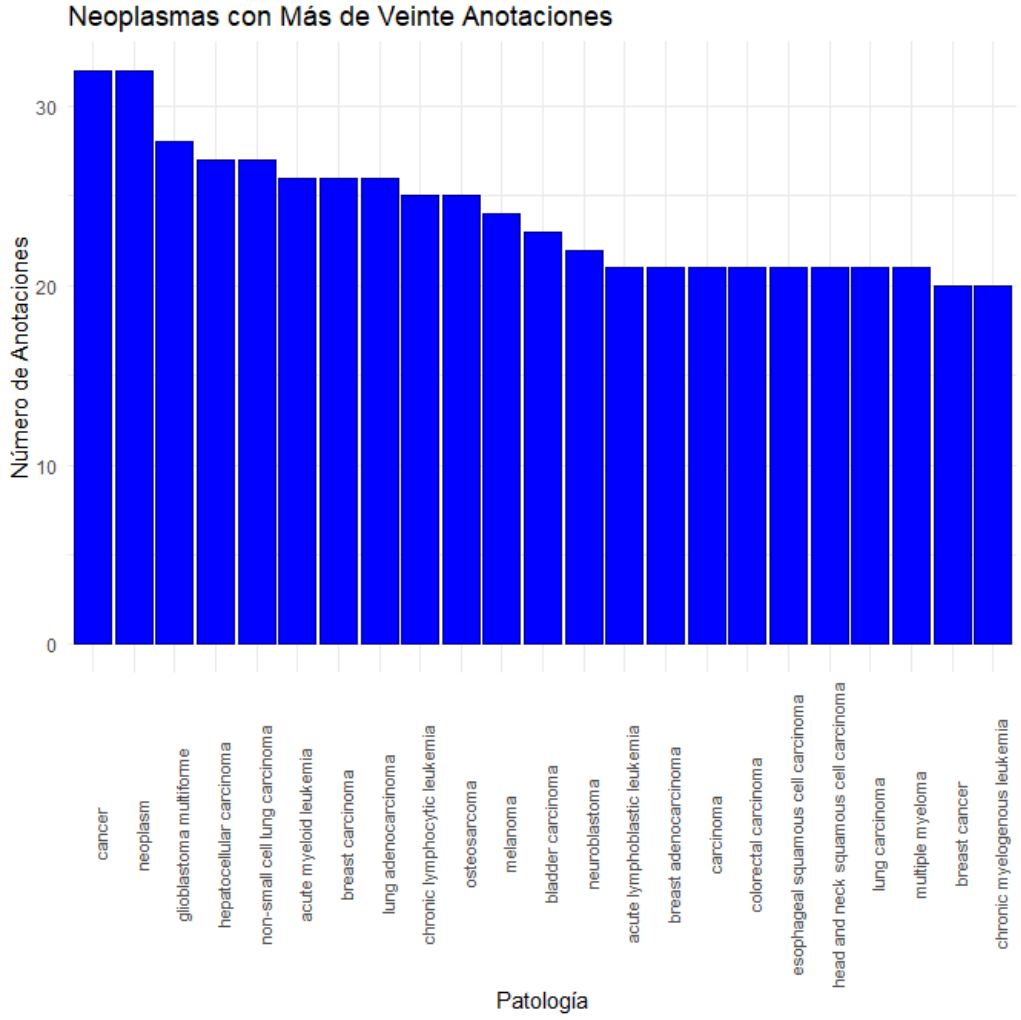
La proteína con más anotaciones de neoplasmas ha sido *P04637*, con **1204 anotaciones,** perteneciendo al motivo mayoritario de la base de datos, *MOD_SUMO_for_1*.

El tipo de patología mayoritario, sin tener en cuenta los agrupamientos genéricos como *cancer* y *neoplasma*, es *glioblastoma multiforme*, con **28 anotaciones**, seguida de *hepatocellular carcinoma* y *non-small cell lung carcinoma* con **27 anotaciones** y *acute myeloid leukemia* con **26 anotaciones**.

Tab 26: Sumario de anotaciones con patologías asociadas a neoplasmas

Motivo	Proteína	Patología
DOC_GSK3_Axin_1 : 177	P04637 :1204	cancer : 32
LIG_Dynein_DLC8_1: 162	Q16665 : 387	neoplasm : 32
LIG_G3BP_FGDF_1 : 15	P06401 : 386	glioblastoma multiforme : 28
LIG_KEPE_1 : 385	O14686 : 309	hepatocellular carcinoma : 27
MOD_SUMO_for_1 :4363	P10242 : 291	non-small cell lung carcinoma: 27
	P04150 : 265	acute myeloid leukemia : 26
	(Other):2260	(Other) :4930

Fig. 66: Distribución de patologías asociadas con más de veinte anotaciones en la base de datos IKFXZB_Disease



Tab 27: Tabla de frecuencias de patologías asociadas a neoplasmas

	Patología	Frecuencia
1	cancer	32
2	neoplasm	32
3	glioblastoma multiforme	28
4	hepatocellular carcinoma	27
5	non-small cell lung carcinoma	27
6	acute myeloid leukemia	26
7	breast carcinoma	26
8	lung adenocarcinoma	26
9	chronic lymphocytic leukemia	25
10	osteosarcoma	25

5.3.2.2 Distribución de Proteínas con Patologías Asociadas al Sistema Inmune

Tab 28: Tabla de frecuencias por proteína de anotaciones de patologías asociadas a ESI

Prot.	O14686	O15169	O15525	O43521	P04150	P04637	P06401	P10242	P14316	P17676	P23497
Freq.	49	17	7	165	178	243	44	164	78	112	15
Prot.	P27540	P29590	P49715	P49716	P54845	P55854	P56524	P61956	Q00613	Q02447	Q13547
Freq.	16	91	125	20	3	6	41	8	14	13	64
Prot.	Q13569	Q14526	Q14694	Q15326	Q15596	Q15744	Q16665	Q92754	Q9NSC2	Q9UPW6	
Freq.	4	28	2	9	25	145	87	5	17	21	

Se ha obtenido una lista de **1816 referencias** de *ESI* asociadas a **32 proteínas.**

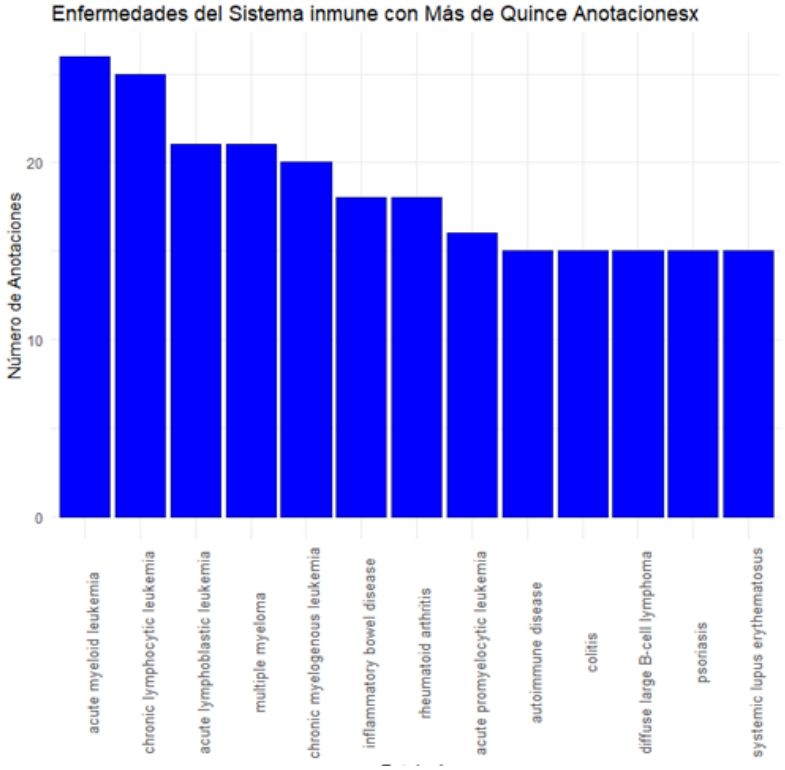
La proteína con más anotaciones de asociaciones a *ESI* es *P04637*, con **243 anotaciones,** que pertenece al motivo mayoritario de la base de datos, *MOD_SUMO_for_1*, que presenta 1486 anotaciones de patologías asociadas.

El tipo de patología mayoritario, es *acute myeloid leukemia*, con **26 anotaciones**, seguidos de *chronic lymphocytic leukemia*, con **25 anotaciones** y de *multiple myeloma* y *acute lymphoblastic leukemia* con **21 anotaciones**.

Tab 29: Sumario de anotaciones con patologías asociadas a ESI

Motivo	Proteína	Patología
DOC_GSK3_Axin_1 : 17	P04637 :243	acute myeloid leukemia : 26
LIG_Dynein_DLC8_1: 174	P04150 :178	chronic lymphocytic leukemia: 25
LIG_G3BP_FGDF_1 : 2	O43521 :165	acute lymphoblastic leukemia: 21
LIG_KEPE_1 : 137	P10242 :164	multiple myeloma : 21
MOD_SUMO_for_1 :1486	Q15744 :145	chronic myelogenous leukemia: 20
	P49715 :125	(Other) :1702
	(Other):796	NA's : 1

Fig. 66: Distribución de patologías asociadas a ESI con más de quince anotaciones



Tab 30: Tabla de frecuencias de patologías asociadas a ESI

	Patología	Frecuencia
1	acute myeloid leukemia	26
2	chronic lymphocytic leukemia	25
3	acute lymphoblastic leukemia	21
4	multiple myeloma	21
5	chronic myelogenous leukemia	20
6	inflammatory bowel disease	18
7	rheumatoid arthritis	18
8	acute promyelocytic leukemia	16
9	autoimmune disease	15
10	colitis	15

6. Conclusiones

6.1 Discusión

El Trabajo Final de Máster del cual es objeto esta memoria se ha centrado en la consecución de dos grandes objetivos:

- La obtención del primer objetivo del proyecto consiste en la realización de una base de datos de epítopos de proteínas multitarea para ampliar la base de datos MultitaskProtDB-II.
- El segundo objetivo delTFM consiste en la identificación de mimotopos de proteínas multitarea implicadas en enfermedades y en la determinación de posibles relaciones causales entre ambos factores.

Se puede considerar que los dos objetivos del TFM se han completado satisfactoriamente, puesto que se ha generado una base de datos con **3719 anotaciones de epítopos** de las **253 proteínas multitarea** implicadas en procesos de virulencia además de obtenerse dos bases de datos de patologías asociadas a proteínas con motivos proteicos análogos a los mimotopos del grupo *de chaperonas de 60 kDA con estructura análoga a 1iok_A* presentes en la base de datos de epítopos.

Al ser un TFM de tipo analítico, el trabajo se han concentrado en la obtención las diferentes bases de datos y en su análisis estadístico pero, y no menos importante, también ha servido para crear un protocolo de identificación de epítopos y de obtención de patologías asociadas a proteínas análogas a epítopos de organismos patógenos que se divide en las siguientes fases:

- Selección de los códigos *PDB* de las estructuras proteicas análogas a cada proteína que se quiera analizar y anotación de las mismos en una base de datos de tipo tabulada (.tab, .xlsx, …).
- Revisión de los códigos *PDB* de las proteínas a analizar mediante la base de datos *RCSB PDB* (https://www.rcsb.org/) para comprobar si son específicos de la proteína anotada o, por el contrario, son simplemente estructuras análogas.
- Para las proteínas que no tengan un código *PDB* específico, obtención de su estructura primaria en formato *FASTA* a partir de la base de datos de *Uniprot* (https://www.uniprot.org/) y anotación de la misma en un archivo de texto individualizado.
- Predicción de los epítopos a partir de las estructuras primarias obtenidas anteriormente mediante la *herramienta de predicción de epítopos de células B de IEDB* (http://tools.iedb.org/bcell/), mediante *bepipred-2.0*, y copia y anotación de la lista de epítopos generada en un archivo tabulado individualizado.
- Si la proteína analizada tiene un código *PDB* específico, predicción de los epítopos a partir del código *PDB* mediante la herramienta *Ellipro* de *IEDB* (http://tools.iedb.org/ellipro/) y copia y anotación de las dos listas de epítopos generadas en dos archivos tabulados, uno para los epítopos lineales y el otro para los epítopos discontinuos.
- Copia de los epítopos lineales guardados en los archivos tabulados y anotación de los mismos en una base de datos conjunta.

- Análisis estadístico de la base de datos generada y comparación de la localización y distribución de los epítopos mediante diagramas de puntos y de cajas para analizar patrones y/o distribuciones comunes entre los grupos de proteínas analizados.
- Alineación de secuencas con *T-Coffee Expresso* (*http://tcoffee.crg.cat/apps/tcoffee/do:expresso*) de las proteínas sobre las que se busca obtener las secuencias de epítopos consenso.
- Localización de las regiones dónde se localizan los epítopos en cada secuencia y selección de los epítopos coincidentes en localización.
- Realización de la comparación de secuencias con *Tomtom* (https://meme-suite.org/meme/tools/tomtom) y obtención de las secuencias de epítopos consenso (mimotopos).
- Si se pretende realizar un análisis comparativo de la totalidad (o una parte) de los mimotopos, selección de los epítopos más parecidos a los mimotopos a analizar y realización de un análisis comparativo en la base de datos *UniprotKB Human* mediante *FASTM* (https://www.ebi.ac.uk/Tools/sss/fastm/).
- Si, en cambio, se pretende realizar un análisis individual de los motivos análogos a cada uno de los mimotopos obtenidos con *Tomtom*, revisión de los mismos en *ELM* (http://elm.eu.org/) y obtención de la lista de proteínas con patologías asociadas para cada mimotopo.
- Anotación de las proteínas humanas con secuencias peptídicas similares al conjunto de mimotopos obtenidos en una base de datos de tabulada y análisis estadístico de la misma.
- Revisión individual de la lista de proteínas generada mediante *Uniprot* (https://www.uniprot.org/), búsqueda de patologías asociadas mediante *DisGeNet* (https://www.disgenet.org/) y/o *Open Targets* (https://platform.opentargets.org/)
- Anotación de los resultados obtenidos en una nueva base de datos tabulada y análisis estadístico de la misma.

Aunque el protocolo definitivo ha sido testado y se considera adecuado, éste es el resultado de varios errores y modificaciones durante el proceso de realización del TFM, ya que ha sido necesario realizar cambios en el planteamiento de algunos de los objetivos específicos como consecuencia de errores de concepto y de resultados inesperados en los análisis previos.

El primer factor que limitó el planteamiento de TFM fue el gran tamaño de las base de datos *MultitaskProtDB-II*, 694 proteínas, que provocó que se optara por realizar únicamente la **predicción de 253 proteínas multitarea**, aquellas implicadas en procesos de virulencia.

Otra modificación del TFM se debió a un error de concepto, al creer que el código *PDB* indicado en la base de datos *Moon.xlsx* era la propia estructura tridimensional de la proteína, siendo incorrecto en la mayor parte de los casos, lo que provocó que se tuvieran que revisar todos los archivos de epítopos ya generados con *Ellipro* y que se descartaran para la base de datos los que no se correspondían con la estructura real de las proteínas analizadas.

Este error, además, provocó que se tuviera que volver a realizar las predicciones de epítopos para todas las proteínas sin estructura tridimensional conocida mediante *Bepipred 2.0*, que paso a ser la herramienta de predicción de epítopos principal al utilizarse en la predicción de **221 proteínas**.

Otro de los causantes de la reorientación de algunos de los planteamientos parciales han sido los resultados del análisis de los mimótopos individuales con *Tomtom*, que ha proporcionado una gran cantidad de proteínas provistas de patologías asociadas con motivos análogos a estos mimotopos, **130 motivos** en total, obligando a priorizar la creación y testeo de un protocolo de análisis de epítopos y la obtención de una lista de patologías asociadas a los mismos, creando una base de datos de los mimotopos *IKFXZB, G y GXPX* con **595 anotaciones,** en lugar de crear una base de datos con las referencias de patologías asociadas de todas proteínas con motivos análogos a los mimótopos de 1iok_A.

Por la misma razón comentada anteriormente, la imposibilidad realizar más bases de datos dentro de la temporización prevista, sólo se ha generado la base de datos de patologías asociadas a las proteínas de los motivos análogos al mimotopo *IKFXZB* con las anotaciones extraídas únicamente de Open Targets por las siguientes razones:

- El número de anotaciones generadas ha sido muy grande, sólo con *Open Targets* se han generado **6918 anotaciones** en la base de datos.
- Aunque con la herramienta de búsqueda de *DisGeNet* se puede acceder al número de referencias tanto de *neoplasmas* como de *ESI*, los archivos descargados con la lista de patologías asociadas sólo indican un tipo de clase de patología, por lo que se perdían anotaciones de patologías que estuvieran relacionadas tanto con *neoplasmas* como con *ESI*.
- *DisGeNet* proporciona principalmente anotaciones de patologías asociadas a *neoplasmas*, generando resultados muy parecidos a *Open Targets*, por lo que se ha considerado que ésta sólo aportaría anotaciones repetidas que, además, obviarían estadísticamente parte de la importancia de las patologías asociadas a *ESI*.

Finalmente, cabe indicar que en el plan de trabajo del TFM también estaba previsto encontrar secuencias similares en organismos humanos con *PSI-BLAST*, al ser una buena herramienta para detectar proteínas humanas ortólogas pero, pese a haberse realizado el análisis comparativo, no se ha realizado ningún análisis posterior con los datos obtenidos, al considerarse más significativo centrarse en el análisis de las secuencias proteicas obtenidas con *FASTM* y de los motivos proteicos obtenidos con *Tomtom*.

6.2 Análisis Final

Aparte de los resultados obtenidos en el TFM, tanto en forma de bases de datos como de protocolo de predicción de epítopos, de comparación de mimotopos y de búsqueda de patologías asociadas con proteínas humanas análogas, el análisis de los resultados obtenidos aporta **bastantes indicios sobre la relación entre infecciones patógenas y la aparición de enfermedades asociadas al sistema inmune o a neoplasmas**.

En primer lugar, analizando la base de datos de epítopos se ha podido observar que las proteínas implicadas en procesos de virulencia de diferentes especies, en muchos casos, presentan estructuras análogas con una secuencia y localización de epítopos similar, tal como se ha podido comprobar tanto con el análisis de las *proteínas con estructura análoga a 1iok_A* como, en más detalle, con el análisis de las *chaperonas de 60kDA*, lo

que induce a suponer que los diferentes organismos analizados puedan presentan mecanismos similares de infección y de mimetismo.

En relación a este punto, cabe destacar la utilidad de la realización de diagramas de puntos y de cajas para comparar la localización y distribución de los epítopos de las proteínas analizados y agruparlas por patrones comunes.

En segundo lugar, pese a la dificultad de conseguir secuencias análogas a un conjunto de mimotopos tan grande como el obtenido con el grupo de las *chaperonas de 60kDa con estructura análoga a 1iok_A*, no solo se ha generado una lista de proteínas con un un alto grado de similitud, **13 de ellas con un E-value < 0.05**, sino que la proteína humana con mayor grado de similitud, la *chaperona de 60kDa P10809*, presenta una importante cantidad de patologías asociadas, tanto del sistema inmune como neoplasmas, lo que podría ser un indicio entre la relación del desarrollo de enfermedades relacionadas con la proteína *P10809* e infecciones de alguno de los organismos patógenos con proteínas de este tipo.

Un tercer indicio de esta conexión entre el mimetismo de patógenos y el desarrollo de enfermedades en el huésped se puede detectar con el análisis individual de los mimotopos obtenidos.

Pese a haberse analizado en profundidad sólo tres mimotopos, todos presentan en sus motivos análogos proteínas de organismos patógenos implicados en procesos de infección, habiéndose comprobado que éstos interactúan con proteínas humanas pudiendo inducir la aparición de patologías del sistema autoinmune o neoplasmas.

De esta forma, la *fosfoproteína P* del virus de la rabia, con motivo análogo al mimotopo *IKFXZB*, interactúa directamente con la *leucemia promielocítica inducida por interferón* al reorganizar los cuerpos nucleares de la *LMP*.

Otro ejemplo de interacción entre infecciones y patologías se da en el motivo *LIG_Integrin_isoDGR_2*, análogo al mimotopo *G*, que incluye a las integrinas, un tipo de proteínas que, debido a su papel fundamental en la comunicación celular, suelen ser dianas de varios virus, como el *virus de la fiebre aftosa*, el *VIH*, el *virus del Nilo Occidental* o el *VPH-16* y de otros organismos patógenos, incluidas bacterias y organismo patógenos eucariotas.

Este tipo de proteínas pueden estar implicadas en la aparición de varias patologías como la *enfermedad de Alzheimer*, la *fibrosis quística*, el *trastorno del espectro autista*, la *esquizofrenia* y el *cáncer*, cuando existe una la mala regulación de las mismas, por lo que se observa una nueva coincidencia entre procesos de mimetismo de organismos patógenos y proteínas con patologías asociadas, tanto del sistema inmune como de neoplasmas.

Finalmente, el último indicio importante de correlación entre ambos factores es la gran cantidad de proteínas con patologías asociadas con motivos análogos a los tres mimotopos analizados, obteniéndose **6918 anotaciones de patologías** asociadas a proteínas con motivos análogos únicamente al primer de los mimotopos, *IKFXZB*, dándose el caso de algunas proteínas con **más de 500 referencias propias**, como *P04637*, que cuenta ella sola con **1447 referencias de patologías asociadas**.

Todos estos datos encontrados durante el proceso de realización del TFM son indicios que **refuerzan la teoría que existe relación entre algunos procesos de infección por patógenos y la aparición de patologías asociadas a *ESI* y a *neoplasmas***, aunque, al haberse realizado todo el análisis virtualmente, sería necesario necesario comprobar experimentalmente la causalidad de ambos factores.

6.3 Líneas de Investigación Futuras

El gran volumen de datos generados y la amplitud de análisis y estudios que se podrían realizar con la información obtenida en cada una de las fases del trabajo ha provocado que varios planteamientos previos se hayan quedado en el tintero y que se haya debido renunciar a la realización de varios análisis interesantes, y que podrían ayudar a continuar con la investigación realizada en este TFM.

Un primer proyecto que sería muy interesante de realizar, aparte de completar la predicción de epítopos del resto de proteínas de *MultitaskProtDB-II*, sería conseguir el código *PDB* de todas las proteínas de la base de datos, no sólo para realizar una predicción de epítopos más completa con *Ellipro*, sino también para poder agrupar las proteínas por estructura y buscar secuencias de mimotopos comunes, ya que el análisis realizado ha dejado claro la importancia de la estructura proteica a la hora de buscar secuencias de epítopos consenso.

Otro proyecto interesante sería poder crear una página de consulta de datos en red con *Shiny*, de forma que se pudieran consultar directamente las bases de datos en red o, incluso, entrar los datos propios de otros usuarios, en un formato determinado previamente, y que la página realizara los mismos análisis en *R* que los realizados en el TFM.

Sería interesante, en continuación con el trabajo realizado en el TFM, completar la base de datos de mimotopos con el análisis individual del resto del mimotopos, de forma que se tuvier un panorama completo de las proteínas análogas a la totalidad de epítopos del grupo de *chaperonas de 60kDa con estructura análoga a 1iok_A* y de la lista completa de patologías que podrían estar asociadas a la misma.

Entrando en más detalle en este último análisis, también sería interesante investigar experimentalmente los indicios que relacionan infecciones patógenas y enfermedades asociadas a ESI o a neoplasmas, concentrado esfuerzos principalmente en aquellas proteínas con un alto número de referencias de patologías asociadas, como la proteína *P04637* o la *chaperona de 60kDa P10809*.

Finalmente, cabe indicar que, al ser un campo de estudio muy amplio y poco desarrollado, la lista de proyectos que se podrían realizar a partir de este TFM es muy grande, tanto a nivel de predicción de epítopos como de análisis comparativos en busca de proteínas análogas a mimotopos o de investigación de proteínas específicas con patologías asociadas pero como planteamiento para futuras investigaciones, las líneas de investigación aquí planteadas podrían ser un buen punto de partida.

7. Glosario

En esta página se muestra únicamente los conceptos clave del TFM. Para consultar el glosario de definiciones completo se puede acceder en el siguiente link.

Analogía *f* Similitud observada entre dos estructuras o secuencias que no tienen un origen común. El parecido se alcanza por convergencia.
BepiPred-2.0 *f* Herramienta de predicción de epítopos de células B.
Células B *f* Célula del sistema inmunitario que se forman a partir de las células madre en la médula ósea. También se llama linfocito B.
Clustal Omega *m* Programa de alineación de múltiples de secuencias divergentes biológicamente significativas.
Chaperona de 60 kDa *f* Proteína de 60 kDa de peso, cuya función es la de ayudar al plegamiento de otras proteínas recién formadas en la síntesis de proteínas.
DisGeNet *f* Plataforma web que contiene un repositorio de genes y variantes disponibles públicamente asociados a enfermedades humanas.
ElliPro *f* Herramienta web que implementa el método de Thornton y permite la predicción y visualización de epítopos de anticuerpos en una secuencia o estructura de proteína determinada.
The Eukaryotic Linear Motif Resource (ELM) *m* Herramienta que se centra en la anotación y detección de motivos lineales eucariotas al tener una base de datos de motivos anotados y una herramienta exploratoria para la predicción de motivos.
Enfermedades del sistema inmune (ESI) *f* Enfermedades que afectan al conjunto de elementos y procesos biológicos en el interior de un organismo que le permite mantener la homeostasis o hacer frente a agresiones externas.
Epítopo *m* Porción de una macromolécula que es reconocida por el sistema inmunitario, específicamente es la secuencia a la que se unen los anticuerpos, los receptores de las células B o de las células T.
1iok_A *f* Estructura cristalina de l bacteria *Paracoccus denitrificans.*
E-value *m* Parámetro que describe el número de aciertos que uno puede "esperar" ver por casualidad cuando se busca en una base de datos de un tamaño particular.
FASTA *m* formato de fichero informático basado en texto, utilizado para representar secuencias bien de ácidos nucleicos, bien de péptidos, que se representan usando códigos de una única letra.
FASTM *f* Herramienta de búsqueda de similitudes y comparación de péptidos sobre una base de datos de secuencias proteicas.
Función canónica *f* Primera función descrita para una proteína.
Matriz extracelular del (MEC) *f* Conjunto de materiales extracelulares que forman parte de un tejido, componen la sustancia del medio intersticial (intercelular).
Mimetismo molecular *m* e Posibilidad de que las similitudes de secuencia entre péptidos extraños y propios provoquen la activación cruzada de células T o B por péptidos derivados de patógenos.
Mimotopo *m* Macromolécula, a menudo un péptido, que imita la estructura de un epítopo. Debido a esta propiedad, provoca una respuesta de anticuerpos similar a la provocada por el epítopo imitado.
Motivo proteico *m* Combinaciones simples de elementos de la estructura secundaria que ocurren con frecuencia en la estructura de la proteína.
MultitaskProtDB-II *f* Base de datos de proteínas multitarea promovido por el Institut de Biotecnologia i Biomedicina de la Universitat Autònoma de Barcelona.
Neoplasma *m* Crecimiento descontrolado de células o tejidos anormales en el organismo.
Open Targets *f* Plataforma web que utiliza datos de genética humana y genómica para la identificación y priorización sistemáticas de objetivos de fármacos.
Código PDB *m* Código de identificación o acceso único a cada modelo molecular del Protein Data Bank (PDB)
Proteínas multitarea *f* Proteínas que presentan funciones alternativas a la canónica, realizadas por una sola cadena polipeptídica, como consecuencia de diferencias en el estado, condiciones y/o localización.
PSI-BLAST *f* Herramienta web que permite a los usuarios construir y realizar una búsqueda NCBI BLAST con una matriz de puntuación personalizada y específica de la posición que puede ayudar a encontrar relaciones evolutivas distantes.
R *m* Entorno y lenguaje de programación de software libre con un enfoque al análisis estadístico.
T_Coffe Expresso *m* Servdor de alineamiento múltiple de secuencias de tipo estructural.
Tomtom *f Herramienta de* comparación de motivos con una base de datos de motivos conocidos. Clasifica los motivos en la base de datos y produce una alineación para cada coincidencia significativa.
Uniprot Align *f* Herramienta de alineación de secuencias de The Universal Protein Resource (UniProt).

8. Bibliografía

[1] Jeffery, C. J. (1999). Moonlighting proteins. *Trends Biochem. Sci.* 24, 8–11. doi:10.1016/ S0968-0004(98)01335-8.

[2] Huberts DH, van der Klei IJ. Moonlighting proteins: an intriguing mode of multitasking. *Biochim Biophys Acta* 2010;**1803**:520–5.

[3] Copley SD. Moonlighting is mainstream: paradigm adjustment required. *Bioessays* 2012;**34**:578–88.

[4] Jeffery CJ. (2014). An introduction to protein moonlighting. *Biochem Soc Trans*.;**42**:1679–83.

[5] Franco-Serrano L, Hernández S, Calvo A, Severi MA, Ferragut G et al. (2018). Una base de datos de proteínas multitarea. Ácidos nucleicos, Res. https://doi.org/10.1093/nar/gkx1066

[6] Franco-Serrano, L., Huerta, M., Hernández, S. *et al.* (2018). Proteínas multifuncionales: participación en enfermedades humanas y objetivos de los fármacos actuales. *Protein J* 37, 444–453. https://doi.org/10.1007/s10930-018-9790x

[7] Amblee, V.; Jeffery, C.J. (2015). Physical Features of Intracellular Proteins that Moonlight on the Cell Surface. PLoS ONE, 10, e0130575.

[8] Luis Franco-Serrano, Juan Cedano, Josep Antoni Perez-Pons, Angel Mozo-Villarias, Jaume Piñol, Isaac Amela, Enrique Querol. (2018). Una hipótesis que explica por qué tantas proteínas de virulencia de patógenos son proteínas de pluriempleo, *Patógenos y enfermedades* , Volumen 76, Número 5 , Julio de 2018, fty046, https://doi.org/10.1093/femspd/fty046

[9] Wistow, G., and Piatigorsky, J. (1987). Recruitment of enzymes as lens structural proteins. *Science* 236, 1554–1556. doi:10.1126/science.3589669

[10] Chen C, Zabad S, Liu H, Wang W, Jeffery C (2018) MoonProt 2.0: una expansión y actualización de la base de datos de proteínas del pluriempleo. Res de ácidos nucleicos, https://doi.org/10.1093/nar/gkx1043

[11] Diogo M Ribeiro, Galadriel Briere, Benoit Bely, Lionel Spinelli, Christine Brun. (2019). MoonDB 2.0: an updated database of extreme multifunctional and moonlighting proteins, *Nucleic Acids Research*, Volume 47, Issue D1, 08 January 2019, Pages D398–D402, https://doi.org/10.1093/nar/gky1039

[12] Franco-Serrano L, Sánchez-Redondo D, Nájar-García A, Hernández S, Amela I, Perez-Pons JA, Piñol J, Mozo-Villarias A, Cedano J, Querol E. (2021). Pathogen Moonlighting Proteins: From Ancestral Key Metabolic Enzymes to Factores virulentos. *Microorganismos* . 9 (6): 1300. https://doi.org/10.3390/microorganisms9061300

[13] Negi, S.S. and Braun, W. (2009). Automated Detection of Conformational Epitopes using Phage Display Peptide Sequences. Bioinform. Biol. Insights, 3, 71-81. http://curie.utmb.edu/episearch.html

[14] Mayrose I, Penn O, Erez E, Rubinstein ND, Shlomi T, Tarnovitski Freund N, Bublil E, Rupin E, Sharan R, Gershoni JM, Martz E, Pupko T. (2007). Pepitope: epitope mapping from affinity-selected peptides. Bioinformatics 23(23):3244-3246. http://pepitope.tau.ac.il/

[15] Odorico M, Pellequer JL (2003) BEPITOPE: predicting the location of continuous epitopes and patterns in proteins. *J. Mol. Recogn.* **16**: 20-22. http://bepitope.ibs.fr/

[16] Reche PA, Glutting JP and Reinherz EL Prediction of MHC Class I Binding Peptides Using Profile Motifs. Human Immunology 63, 701-709 (2002).

[17] Reche PA, Glutting JP, Zhang H, Reinherz EL. (2004). Enhancement to the RANKPEP resource for the prediction of peptide binding to MHC molecules using profiles. 456:405-419. http://imed.med.ucm.es/Tools/rankpep.html.

[18] Manoj Bhasin, Ellis L. Reinhez and Pedro A. Reche. (2005). Modeling features of immunodominance into T-cell epitope identification. 2nd International Immunoinformatics Symposium, Boston University, March 7-9.

[19] Chen J, Liu H, Yang J, Chou K (2007) Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. Amino Acids 33: 423-428.

[20] EL-Manzalawy Y, Dobbs D, Honavar V. (2008). Predicting linear B-cell epitopes using string kernels. J Mol Recognit 21: 243-255.

[21] EL-Manzalawy Y, Dobbs D, Honavar V. (2008). Predicting flexible length linear xB-cell epitopes. 7th International Conference on Computational Systems Bioinformatics, Stanford, CA. pp. 121-131. http://ailab-projects1.ist.psu.edu:8080/bcpred/predict.html.

[22] Larsen JE, Lund O, Nielsen M. (2006). Improved method for predicting linear B-cell epitopes. *Immunome Res* **2**:2. PMID: 16635264.

[23] Jespersen MC, Peters B, Nielsen M, Marcatili P. (2017). BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. Nucleic Acids Res (Web Server issue). 2:2. PMID: 28472356.

[24] P. H. Andersen, M. Nielsen and O. Lund. (2006). Prediction of residues in discontinuous B cell epitopes using protein 3D structures. *Protein Science*. **15**:2558-2567. PMID: 17001032.

[25] J. V. Kringelum, C. Lundegaard, O. Lund, M. Nielsen. (2012). Reliable B cell epitope predictions: impacts of method development and improved benchmarking. *PLoS Comput Biol*. **8**:(12):e1002829. PMID: 23300419.

[26] Ponomarenko JV, Bui H, Li W, Fusseder N, Bourne PE, Sette A, Peters B. (2008). ElliPro: a new structure-based tool for the prediction of antibody epitopes. *BMC Bioinformatics* **9**:514. PMID: 19055730.

[27] The UniProt Consortium. (2021). UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. 49:D1 (2021)

[28] 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.

[29] R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

[30] Li W, Cowley A, Uludag M, et al. (2015). The EMBL-EBI bioinformatics web and programmatic tools framework. Nucleic Acids Research. 2015 Jul;43(W1):W580-4. DOI: 10.1093/nar/gkv279. PMID: 25845596; PMCID: PMC4489272.

[31] McWilliam H, Li W, Uludag M, et al. (2013). Analysis Tool Web Services from the EMBL-EBI. Nucleic Acids Research. 2013 Jul;41(Web Server issue):W597-600. DOI: 10.1093/nar/gkt376. PMID: 23671338; PMCID: PMC3692137.

[32] Camacho C, Coulouris G, Avagyan V, et al. (2009). BLAST+: architecture and applications. BMC Bioinformatics. 2009 Dec;10:421. DOI: 10.1186/1471-2105-10-421. PMID: 20003500; PMCID: PMC2803857.

[33] Altschul SF, Madden TL, Schäffer AA, et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research. 1997 Sep;25(17):3389-3402. DOI: 10.1093/nar/25.17.3389. PMID: 9254694; PMCID: PMC146917.

[34] Sievers F, Wilm A, Dineen D, et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular Systems Biology. 2011 Oct;7:539. DOI: 10.1038/msb.2011.75. PMID: 21988835; PMCID: PMC3261699.

[35] Pearson WR. (1999). Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. Genomics. 1991 Nov;11(3):635-650. DOI: 10.1016/0888-7543(91)90071-l. PMID: 1774068.

[36] Pearson WR. (1990). Rapid and sensitive sequence comparison with FASTP and FASTA. Methods in Enzymology. 1990 ;183:63-98. DOI: 10.1016/0076-6879(90)83007-v. PMID: 2156132.

[37] Mackey AJ, Haystead TA, Pearson WR. (2002). Getting more from less: algorithms for rapid protein identification with multiple short peptide sequences. Molecular & Cellular Proteomics: MCP. 2002 Feb;1(2):139-147. DOI: 10.1074/mcp.m100004-mcp200. PMID: 12096132.

[38] Notredame C, Higgins DG, Heringa J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. Journal of Molecular Biology. 2000 Sep;302(1):205-217. DOI: 10.1006/jmbi.2000.4042. PMID: 10964570.

[39] Timothy L. Bailey, James Johnson, Charles E. Grant, William S. Noble. (2015). The MEME Suite, *Nucleic Acids Research*, Volume 43, Issue W1, 1 July 2015, Pages W39–W49, https://doi.org/10.1093/nar/gkv416.

[40] Shobhit Gupta, JA Stamatoyannopolous, Timothy Bailey and William Stafford Noble, (2007). Quantifying similarity between motifs, *Genome Biology*, **8**(2):R24.

[41] Emi Tanaka, Timothy Bailey, Charles E. Grant, William Stafford Noble, Uri Keich. (2011). Improved similarity scores for comparing motifs, *Bioinformatics*, Volume 27, Issue 12, 15 June 2011, Pages 1603–1609, https://doi.org/10.1093/bioinformatics/btr257.

[42] Janet Piñero, Juan Manuel Ramírez-Anguita, Josep Saüch-Pitarch, Francesco Ronzano, Emilio Centeno, Ferran Sanz, Laura I Furlong. (2019). The DisGeNET knowledge platform for disease genomics: 2019 update. Nucl. Acids Res. doi:10.1093/nar/gkz1021

[43] Núria Queralt-Rosinach, Janet Piñero, Àlex Bravo,Ferran Sanz, Laura I. Furlong. (2016). DisGeNET-RDF: Harnessing the Innovative Power of the Semantic Web to Explore the Genetic Basis of Diseases. Bioinformatics. doi:10.1093/bioinformatics/btw214

[44] Janet Piñero, Josep Saüch, Ferran Sanz, Laura I. Furlong. (2021). The DisGeNET cytoscape app: exploring and visualizing disease genomics data, Computational and Structural Biotechnology Journal. doi.org/10.1016/j.csbj.2021.05.015

[45] Anna Bauer-Mehren, Markus Bundschus, Michael Rautschka, Miguel A. Mayer, Ferran Sanz, Laura I. Furlong. (2011). Gene-disease network analysis reveals functional modules in mendelian, complex and environmental diseases. PLoS ONE doi:10.1371/journal.pone.0020284

[46] Bauer-Mehren A, Rautschka M, Sanz F, Furlong LI. (2010). DisGeNET: a Cytoscape plugin to visualize, integrate, search and analyze gene-disease networks. Bioinformatics. doi: 10.1093/bioinformatics/btq538

[47] Ochoa, D. et al. (2021). Open Targets Platform: supporting systematic drug–target identification and prioritisation. Nucleic Acids Research.

[48] Manjeet Kumar, Marc Gouw, Sushama Michael, Hugo Sámano-Sánchez, Rita Pancsa, Juliana Glavina, Athina Diakogianni, Jesús Alvarado Valverde, Dayana Bukirova, Jelena Čalyševa, Nicolas Palopoli, Norman E Davey, Lucía B Chemes, Toby J Gibson. (2020). ELM—the eukaryotic linear motif resource in 2020, Nucleic Acids Research, Volume 48, Issue D1, 08 January 2020, Pages D296–D306, https://doi.org/10.1093/nar/gkz1030.

[49] GanttProject (2003). https://www.ganttproject.biz/

[50] Turnbull, P. C. B. (2013). Introduction: Anthrax history, disease and ecology. Koehler, Theresa, ed. Anthrax. Springer Science & Business Media. ISBN 3662057670.

[51] Turnbull, Peter C. B.; Shadomy, Sean V. (2011). Anthrax from BC 5000 BC to AD 2010. En Bergman, Nicholas H., ed. Bacillus anthracis and Anthrax. John Wiley & Sons. ISBN 1118148088.

[52] Koch R. (1876). The etiology of anthrax, based on the life history of Bacillus anthracis. Beitrage zer Biologie der Ppanzen 1876;2.2:277-310.

[53] Madigan M, Martinko J (editors). (2005). Brock Biology of Microorganisms (11th ed.). Prentice Hall. ISBN 0-13-144329-1.

[54] Turnbull PCB (1996). Bacillus. In: Barron's Medical Microbiology (Baron S et al., eds.) (4th ed. edición). Univ of Texas Medical Branch. ISBN 0-9631172-1-1.

[55] Alberto J.L. Macario, M.D., and Everly Conway de Macario, Ph.D. (2005). Sick chaperones, cellular stress, and disease. Wadsworth Center, Division of Molecular Medicine, New York State Department of Health, Albany, NY 12201-0509, USA. N Engl J Med. 2005 Oct 6;353(14):1489-50

[56] Valpuesta, J.M., Chaperoninas para plegar proteínas. (2011). SEBBM Divulgación, Acércate a nuestros científicos. DOI: http://dx.doi.org/10.18567/sebbmdiv_ANC.2011.06.1. https://web2020.sebbm.es/web/es/divulgacion/acercate-nuestros-cientificos/227-jose-maria-valpuesta-junio-2011-chaperoninas-para-plegar-proteinas

[57] Hahn, J.S.. (2009). The Hsp90 chaperone machinery: from structure to drug development. BMB Rep, 2009. 42(10): p. 623-30

[58] Blondel D, Regad T, Poisson N, Pavie B, Harper F, Pandolfi PP, De Thé H, Chelbi-Alix MK. Rabies virus P and small P products interact directly with PML and reorganize PML nuclear bodies. Oncogene. 2002 Nov 14;21(52):7957-70. doi: 10.1038/sj.onc.1205931. PMID: 12439746.

[59] Regad T, Chelbi-Alix MK. (2001). Role and fate of PML nuclear bodies in response to interferon and viral infections. Oncogene. 2001 Oct 29;20(49):7274-86. doi: 10.1038/sj.onc.1204854. PMID: 11704856.

[60] Barczyk M, Carracedo S, Gullberg D. Integrins. (2009). Cell Tissue Res. 2010 Jan;339(1):269-80. doi: 10.1007/s00441-009-0834-6. Epub 2009 Aug 20. PMID: 19693543; PMCID: PMC2784866.

[61] Takada Y, Ye X, Simon S. (2007). The integrins. Genome Biol. 2007;8(5):215. doi: 10.1186/gb-2007-8-5-215. PMID: 17543136; PMCID: PMC1929136.

[62] Campbell ID, Humphries MJ. (2011). Integrin structure, activation, and interactions. Cold Spring Harb Perspect Biol. 2011 Mar 1;3(3):a004994. doi: 10.1101/cshperspect.a004994. PMID: 21421922; PMCID: PMC3039929.

[63] Hynes RO. (2002). Integrins: bidirectional, allosteric signaling machines. Cell. 2002 Sep 20;110(6):673-87. doi: 10.1016/s0092-8674(02)00971-6. PMID: 12297042.

[64] Donner L, Fälker K, Gremer L, Klinker S, Pagani G, Ljungberg LU, Lothmann K, Rizzi F, Schaller M, Gohlke H, Willbold D, Grenegard M, Elvers M. (2016). Platelets contribute to amyloid-β aggregation in cerebral vessels through integrin αIIbβ3-induced outside-in signaling and clusterin release. Sci Signal. 2016 May 24;9(429):ra52. doi: 10.1126/scisignal.aaf6240. PMID: 27221710.

[65] Reed NI, Jo H, Chen C, Tsujino K, Arnold TD, DeGrado WF, Sheppard D. (2015). The αvβ1 integrin plays a critical in vivo role in tissue fibrosis. Sci Transl Med. 2015 May 20;7(288):288ra79. doi: 10.1126/scitranslmed.aaa5094. PMID: 25995225; PMCID: PMC4461057.

[66] Lilja J, Ivaska J. (2018). Integrin activity in neuronal connectivity. J Cell Sci. 2018 Jun 15;131(12):jcs212803. doi: 10.1242/jcs.212803. PMID: 29907643.

[67] Seguin L, Desgrosellier JS, Weis SM, Cheresh DA. (2015). Integrins and cancer: regulators of cancer stemness, metastasis, and drug resistance. Trends Cell Biol. 2015 Apr;25(4):234-40. doi: 10.1016/j.tcb.2014.12.006. Epub 2015 Jan 5. PMID: 25572304; PMCID: PMC4380531.

[68] Hussein HA, Walker LR, Abdel-Raouf UM, Desouky SA, Montasser AK, Akula SM. (2015). Beyond RGD: virus interactions with integrins. Arch Virol. 2015 Nov;160(11):2669-81. doi: 10.1007/s00705-015-2579-8. Epub 2015 Sep 1. PMID:26321473; PMCID: PMC7086847.

[69] Asokan A, Hamra JB, Govindasamy L, Agbandje-McKenna M, Samulski RJ. (2006): Adeno-associated virus type 2 contains an integrin alpha5beta1 binding domain essential for viral cell entry. J Virol. 2006 Sep;80(18):8961-9. doi: 10.1128/JVI.00843-06. PMID: 16940508; PMCID: PMC1563945.

9. Anexos

Es necesario acceder con la cuenta de la UOC para poder visualizar los archivos y carpetas de todos los enlaces de los anexos.

TFM: Carpeta con todos los archivos incluídos en el TFM

- Bases de Datos:
  - *Moon.xlsx*: Base de datos de las proteínas multitarea analizadas.
  - *Epítopos.xlsx*: Base de datos con todos los epítopos predichos.
  - *PDB.xlsx*: Base de datos con la lista de proteínas análogas a las proteínas multitarea a las que se han predicho sus epítopos.
  - *1iok_A.xlsx*: Base de datos con las anotaciones de proteínas humanas similares a la secuencia de mimotopos predicha.
  - *1iok_A_Disease.xlsx*: Base de datos con la lista de patologías asociadas a las proteínas de la base de datos *1iok_A.xlsx*.
  - *Mimotopos.xlsx*: Base de datos con las anotaciones de proteínas humanas con motivos similares a los tres mimotopos analizados individualmente.
  - *IKFXZB_Disease.xlsx*: Base de datos con la lista de patologías asociadas a las proteínas con motivos análogos al mimotopo IKFXZB.

- Archivos:
  - Alineaciones:
    - *1iok_A FASTA.txt:* Archivos con los códigos FASTA de las seis proteínas multitarea análogas a 1iok_A.
    - *1iok_A FASTM.txt*: Archivo con los resultados del análisis de secuencias análogas a 1iok_A realizado con FASTM.
    - *Mimotop_1iok_A.txt*: Archivo con la secuencia de mimotopos obtenida a partir de Tomtom.
    - *Motivos.docx*: Archivo con las secuencias de epítopos con los cuales se han generado los mimotopos y la lista de motivos proteicos análogos a cada uno de los mimotopos.
  - Bepipred: Carpeta con todos los archivos de epítopos generados con bepipred 2.0.
  - Ellipro: Carpeta con todos los archivos de patologías asociadas a las proteínas humanas análogas a los mimotopos analizados.
  - FASTA: Carpeta con todas las estructuras primarias extraídas en formato FASTA.
  - Disease: Carpeta con todos los archivos de epítopos generados con Ellipro.

- Planificación:
  - Planificacion.gan: Archivo de planificación del TFM.
  - Planificacion.png: Archivo en formato de imagen de la planificación del TFM.

- R:
  - TFM.Rmd: Archivo en formato Rmarkdown a partir del cual se ha realizado el análisis estadístico de la base de datos de epítopos.
  - TFM.R: Archivo en formato R con el código en bruto utilizado para realizar el análisis de la base de datos de epítopos.
  - TFM.html: Archivo de resultados del análisis con R en formato web.
  - TFM.pdf: Archivo de resultados del análisis con R en formato pdf.
  - TFM.docx: Archivo de resultados del análisis con R en formato de documento de texto.
  - Imágenes: Carpeta con las imágenes utilizadas con Rmarkdown (4 imágenes) para realizar el análisis de la base de datos de epítopos.
  - Data: Carpeta con las bases de datos utilizadas en el análisis estadístico con *R*.