Universidad Nacional de La Plata

Facultad de Ciencias Exactas

Trabajo Final de Laboratorio de Procesos Biotecnológicos

“Caracterización de la diversidad de la relación estructura-función en proteínas humanas”

Juan Mac Donagh

Director: Dr. Gustavo Parisi, Grupo de Bioinformática Estructural, UNQ.

Agradecimientos:

A mi papá y mi mamá, Pato y Eli,
a mis hermanos, Mili y Santi,
a mis amigos, Tomi, Juli, Nico y Joaquina, y muchos más,
a Gustavo, Julia y Nico,
y a la Universidad Pública.

All that is gold does not glitter,
Not all those who wander are lost;
The old that is strong does not wither,
Deep roots are not reached by the frost.

J.R.R Tolkien

ÍNDICE

Resumen del trabajo:

El proteoma humano consenso está formado por alrededor de 20000 proteínas cuyas formas, tamaños, comportamiento dinámico, composición y relación estructura-función pueden ser parámetros extremadamente variados. En este Trabajo Final de Laboratorios de Procesos Biotecnológicos hemos caracterizado la diversidad estructural del proteoma humano accesible al utilizar métodos computacionales.
Para este fin utilizamos distintas herramientas bioinformáticas (BLASTP y HHblits), bases de datos (UniProt, PDB, MobiDB, etc) y lenguajes de programación (principalmente Python, R y Julia) para generar un mapeo estructural con su consecuente clasificación en diversas categorías estructurales. El conjunto de estas técnicas, de elección al momento de haber comenzado este trabajo, nos han permitido caracterizar el 65,08 % del proteoma humano. Entre las categorías más importantes en abundancia se encuentran las proteínas bajo clases de CATH (Mayoritariamente Alfa / Beta y Alfa-Beta) seguidas de las proteínas rígidas y las flexibles. Llama la atención que proteínas tan particulares como las proteínas anudadas o incluso las repetitivas constituyen el 3,3% y 2,23% del proteoma caracterizado. Esperamos que esta clasificación de la diversidad estructural en diversas categorías nos ayude a comprender procesos evolutivos, dinámicos y/o funcionales a la luz de las mismas.

Objetivos generales y específicos:

La mayoría de los estudios a gran escala para dilucidar reglas generales sobre el comportamiento de las proteínas (estabilidad a sustituciones, diversidad conformacional, velocidad de evolución, etc) se realizan con conjuntos de proteínas que se asumen homogéneos. Sin embargo, durante los últimos años se han ido caracterizando distintos subconjuntos de proteínas que han logrado adquirir una identidad propia y que se alejan del paradigma clásico de la existencia de proteínas como puramente “globulares” o "fibrilares". Como objetivo general de este proyecto nos proponemos catalogar el proteoma humano de acuerdo a diversas propiedades estructurales que en definitiva puedan reflejar la relación estructura-función de sus proteínas. Esta clasificación nos permitirá, en trabajos futuros, reexaminar la generalidad de determinadas reglas y principios derivados utilizando conjuntos de datos asumidos en principio como homogéneos. Para esto proponemos como objetivos específicos:

- Caracterizar estructuralmente las proteínas que componen el proteoma humano utilizando métodos computacionales basados en propiedades secuenciales y estructurales
- Categorizar las estructuras proteicas mediante el uso de distintas herramientas bioinformáticas y bases de datos.
- Dilucidar la diversidad estructural del proteoma y consolidar un sistema para futuros análisis evolutivos.

1.Introducción:

1.1 Borges y las proteínas

Cualquier persona involucrada en el campo de la Biología, sabe que la palabra “proteína” describe un grupo de macromoléculas esenciales para la vida en la Tierra. Como biólogos moleculares, estamos entrenados desde el principio de nuestros estudios con los conceptos que explican lo que entendemos por proteínas, conceptos que fueron establecidos en el comienzo del siglo XX. Sin embargo, en los últimos años el término “proteína” se convirtió en un eufemismo para describir a un grupo abrumadoramente heterogéneo de estas macromoléculas. Muchos trabajos contemporáneos están apuntados a seleccionar y estudiar exhaustivamente subconjuntos de proteínas para derivar conceptos generales como si las mismas representaran a un todo.

Citando al autor Argentino Jorge Luis Borges (*Otras Inquisiciones*, “El idioma analítico de John Wilkins”, 1952), los animales pueden clasificarse en:

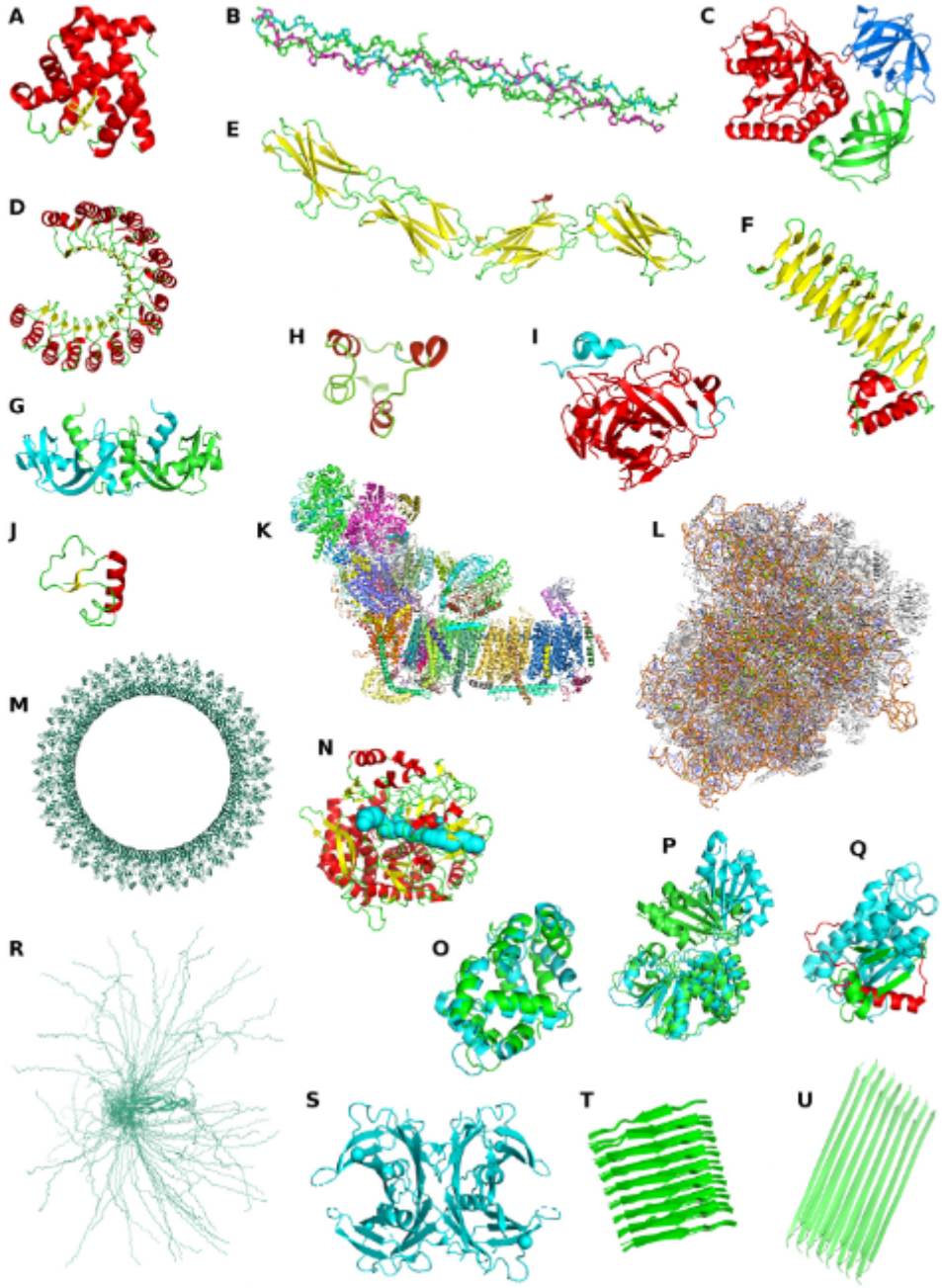
A. pertenecientes al Emperador,
B. embalsamados,
C. amaestrados,
D. lechones,
E. sirenas,
F. fabulosos,
G. perros sueltos,
H. incluidos en esta clasificación,
I. que se agitan como locos,
J. innumerables
K. dibujados con un pincel finísimo de pelo de camello,
L. etcétera,
M. que acaban de romper el jarrón,
N. que de lejos parecen moscas

Borges usa esta particular clasificación para decir que: “(...) notoriamente no hay clasificación del universo que no sea arbitraria y conjetural. La razón es muy simple: no sabemos qué cosa es el universo”. Para Borges, todo esquema de clasificación va a tener errores intrínsecos debido a nuestra ignorancia de potenciales nuevas categorías o a la arbitrariedad de distinguir entidades que pertenecen a distintos grupos.

En los últimos años, se hizo más evidente que es inevitable apreciar a la palabra *proteína* como análoga a la famosa clasificación de animales de Borges. A pesar de nuestras infinitas limitaciones comparados con Borges, podríamos dividir a las proteínas en la figura [1]:

A. Similares a la Mioglobina,
B. Sin forma,
C. Pequeñas,
D. Con nudos,
E. Repetitivas,
F. Provenientes de meteoritos,
G. Caminantes,
H. Que nunca existieron en la Tierra,
I. Resucitadas,
J. Diseñadas artificialmente,
K. Circulares,
L. Que contengan la secuencia “IADAPTEDASDIDMYCHANCES”,
M. Sin descubrir

En esta tesis, buscamos revelar la verdadera diversidad y heterogeneidad presente en las proteínas del proteoma humano.



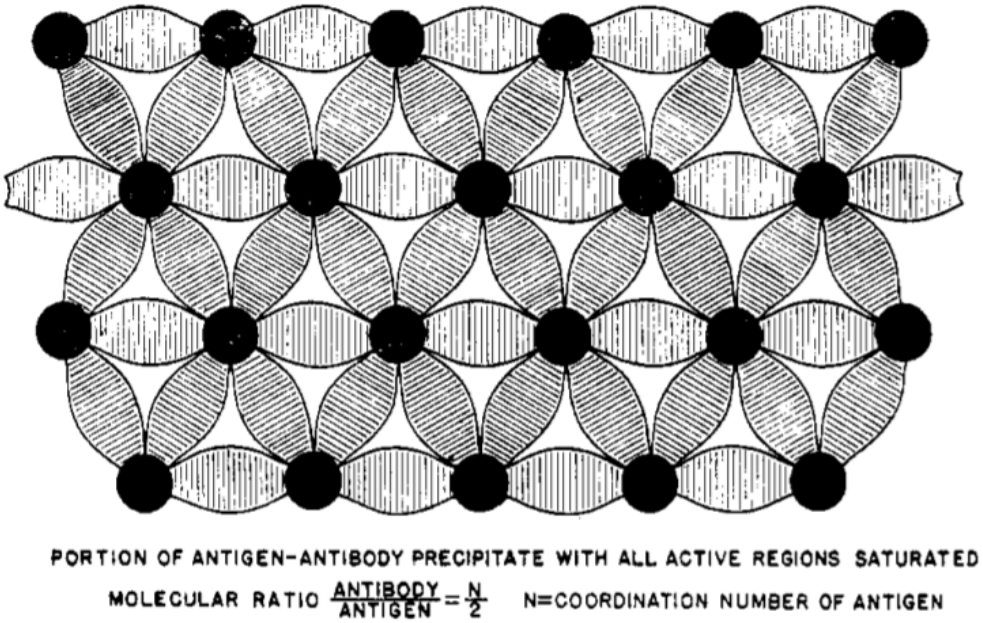
Distintos tipos de proteínas (según su PDB ID y tipo de estructura): A) 5iks, globular. B) 3dmw, fibrilar. C) 2c78, proteína multi-dominio. D) 3tsr, αβ solenoide. E) 3t1w, estructura repetitiva. F) 1lxa, hoja β plegada en sentido contrario. G) 1FS3, "domain swap" protein. H) 2mwr, proteína circular. I) 6y74, estructura nudo. J) 3nir, proteína corta. K) 7a23, complejo multiproteico. L) 6ek0, estructura supramacromolecular. M) 6cb8, estructura altamente simétrica. N) 1f9d, proteína rígida. O) 1niw/1lin, proteína con alta diversidad conformacional. P) 1k20/1k23 estructuras de canal abierto/cerrado ("hinge") Q) 1rk4/1k0n, estructura con "pliegues". R) 2k8p proteína desordenada S y T) 4mrb/2m5n tetrámero y fibra amiloide U) 6nzn, amiloide, [1].

1.2 Breve repaso histórico de la relación estructura-función

Existen muchos modelos que intentan explicar la relación estructura-función que caracteriza a las proteínas. El primero data de 1894, cuando Emil Fisher (Fisher, 1894) propuso el modelo de la "llave y cerradura", que intentaba explicar la relación sustrato-enzima por medio de una analogía denominada *complementariedad geométrica*: El sustrato encaja cual llave, de forma perfecta, en la cerradura (específicamente en su sitio activo). Este modelo fue una excelente primera aproximación (más teniendo en cuenta que se formuló cuando aún no se conocía como estaban unidos los aminoácidos), pero claramente es limitado a la hora de explicar diversos comportamientos de las proteínas que requieren de la existencia de movimientos en la estructura (por ejemplo, la promiscuidad o el cooperativismo).

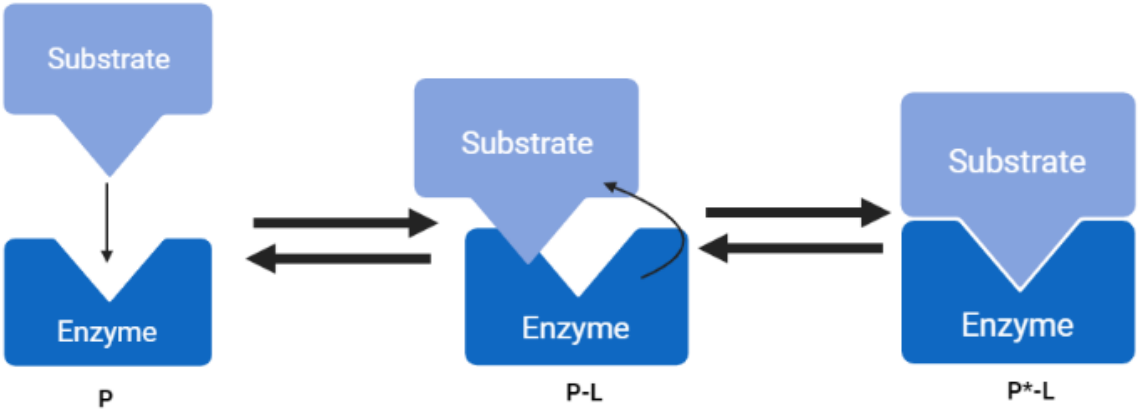
Representación del modelo de complementariedad geométrica de Fischer, [2].

En 1936, Pauling y Mirsky describen un segundo modelo, en el que describen que el estado nativo de las proteínas consiste en el plegamiento de una cadena polipeptídica que le atribuye las características funcionales a la misma, el cual se corresponde con un mínimo de energía libre (Mirsky and Pauling 1936). Luego continúa en esta línea, proponiendo que las proteínas existen como un conjunto de distintas estructuras formadas por los plegamientos posibles de la cadena peptídica, y dentro de este conjunto existía uno que era la estructura nativa (funcionalmente activa), de esta forma explicando cómo los anticuerpos (que comparten una alta similitud secuencial) podían unirse a una gran variedad de antígenos (Pauling 1940).



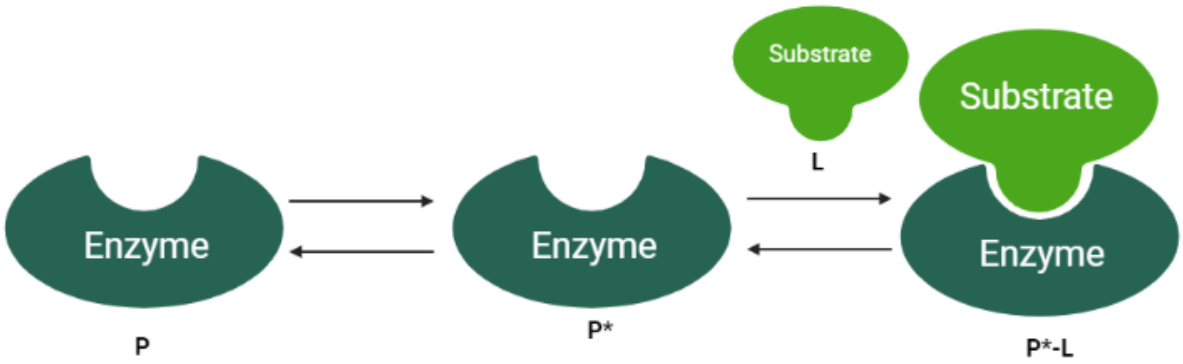
Recorte de la publicación de Pauling, 1940, "A Theory of the Structure and Process of Formation of Antibodies". [3]

Más adelante, en 1950, Karush (Karush, 1950) propone la idea de *adaptabilidad configuracional*, donde explica la razón detrás de que un sitio activo (en su caso estudió a la albúmina bovina) pueda unirse con alta afinidad a distintas moléculas (todas de no gran tamaño y no polares), gracias a pequeñas variaciones estructurales dentro del sitio activo, siempre y cuando estas estuviesen dentro del mismo rango de energía. De este modo encontramos la primera descripción del paisaje enérgico libre, que describe a una proteína con confórmeros distintos como un conjunto de mínimos de energía, permitiéndole una flexibilidad a la hora de la unión con un sustrato. Posteriormente, el descubrimiento de diversas evidencias recolectadas al principio del siglo XX (como la promiscuidad de enzimas, y el efecto cooperativo de la Hemoglobina) promovieron la postulación de Koshland (16) de la teoría del *ajuste inducido*, donde propone que existe un equilibrio de ajuste entre las enzimas y sus sustratos, donde estos pueden lograr un ajuste para que se unan. Este modelo deja de lado las restricciones geométricas y asume la idea actual de entender a las estructuras de una forma más dinámica y no rígidas y estáticas.



Representación del ajuste inducido según Koshland. [4]

Los siguientes aportes al campo fueron por parte de Monod, Wyman y Changeux (Monod et al. 1963), que postularon el modelo del *pre-equilibrio*: este modelo se basa en enfocar a las estructuras de proteínas en marco *alostérico*: las proteínas forman un oligómero (formado por un número finito de monómeros) de forma cooperativa con sus subunidades. También introduce las transiciones conformacionales como dos (o más) estados posibles de estas estructuras, que se pueden adoptar reversiblemente pero que al momento de interactuar con un ligando se ve (el equilibro) afectado favoreciendo uno de estos dos confórmeros (Changeux 2012). En ausencia de ligando, los dos confórmeros (históricamente llamados Tenso y Relajado, T_0 *y* R_0, respectivamente) asumen un equilibrio espontáneo que determina una *constante alostérica* ($L_0 = T_0/R_0$) y el ligando se une distintivamente a uno de estos dos confórmeros, desplazando de esta forma el equilibrio.



Pre-equilibrio, Monod et al. (1965). [5]

Actualmente, entendemos a las proteínas como ensambles dinámicos con confórmeros, y estos deben ser explicados estadísticamente, no estáticamente, debido a las constantes interconversiones de uno a otro, explorando constantemente el paisaje energético libre.
Esto es debido a que las modificaciones que sufren las proteínas, ya sea por la unión de otras moléculas, interacciones con otras proteínas, etc, están constantemente generando cambios en la población de confórmeros. Esto, sumado a la idea del efecto alostérico (que la interacción de una proteína con un sustrato en una parte de su estructura genere un

cambio en otra parte de la proteína) genera los cambios en en las estabilidades relativas de los confórmeros, que dispara el cambio poblacional (Motlagh et al. 2014).
En este sentido, también podemos brevemente hablar de algunas de las técnicas que nos permiten identificar a estos ensembles. Debido a la alta sensibilidad de las modulaciones alostéricas, puede que estos estados solo existan en un porcentaje muy bajo, y durante una minúscula cantidad de tiempo. Es por esto que técnicas como Resonancia Magnética Nuclear (NMR, del inglés Nuclear Magnetic Resonance), o más recientemente, Criomicroscopía electrónica (Cryogenic Electron Microscopy, Cryo-EM), han probado ser inmensamente útiles a la hora de caracterizar estructuras proteicas (Nussinov 2016). También, recientemente, han habido avances muy interesantes en la caracterización de estructuras por medio de métodos de Machine Learning, que amplían inmensamente los horizontes de este campo (Wei 2019, AlQuraishi 2019).

1.3 Plegamiento proteico

Un tema no menor para abordar a la hora de hablar de proteínas es el puente entre la secuencia y la estructura, el plegamiento: en toda secuencia aminoacídica se encuentra la clave para sus estructuras superiores (secundaria, terciaria y cuaternaria). Esto fue explicado por Anfinsen, en su excelente experimento donde demuestra que al desnaturalizar ribonucleasa bovina (purificada) para luego someterla a condiciones óptimas para su plegamiento, esta recupera completamente su actividad, debido a que la proteína recorre, termodinámicamente, un camino hasta llegar a un mínimo de energía, denominado “estado nativo”, en el cual está es funcional.

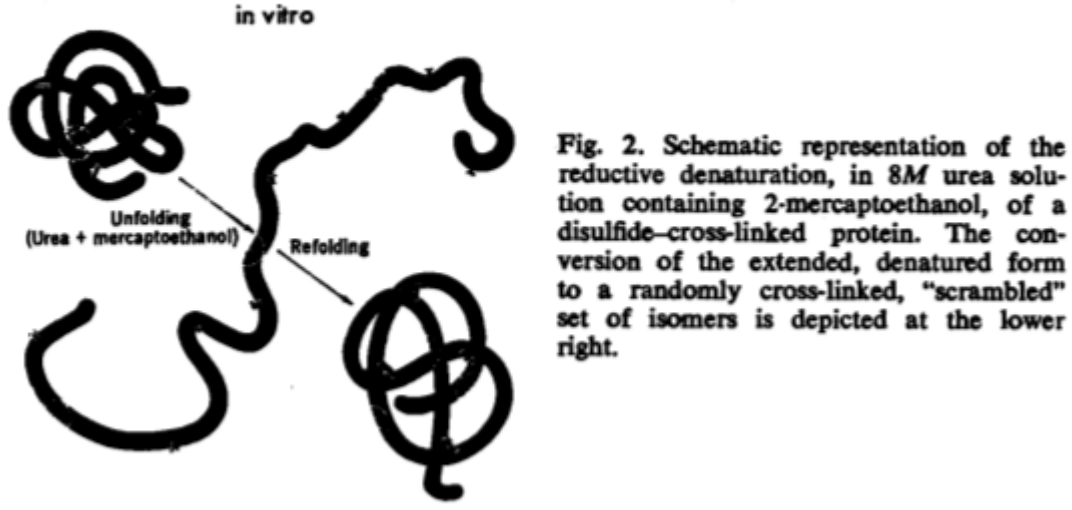
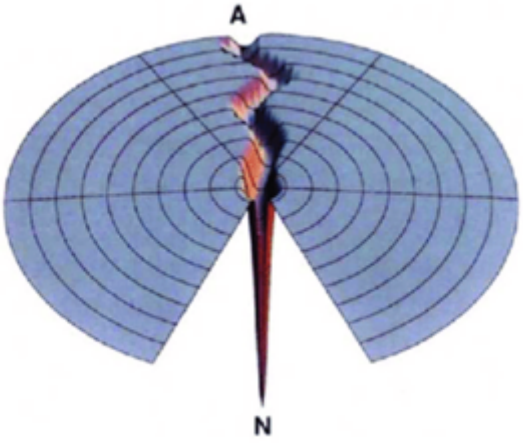


Fig. 2. Schematic representation of the reductive denaturation, in 8*M* urea solution containing 2-mercaptoethanol, of a disulfide–cross-linked protein. The conversion of the extended, denatured form to a randomly cross-linked, “scrambled” set of isomers is depicted at the lower right.

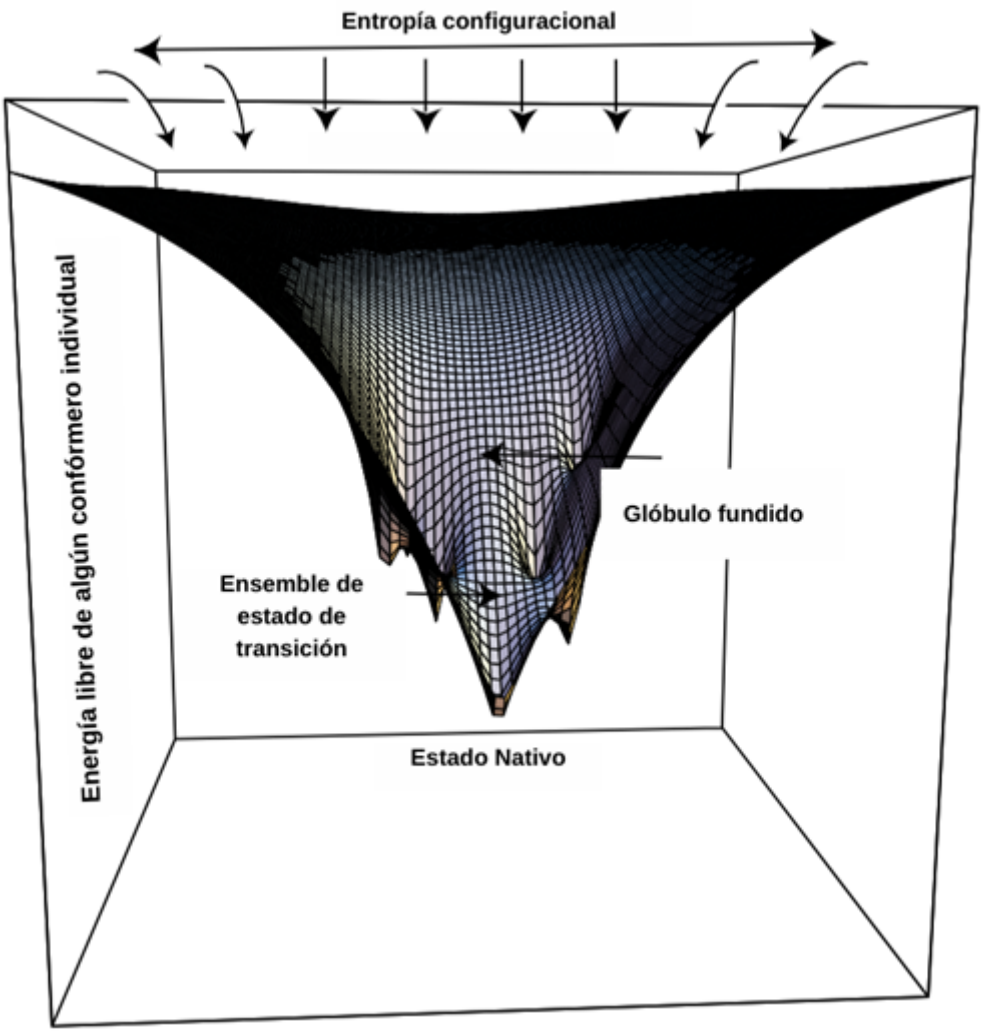
Recorte del paper de Anfinsen, 1973 (Science). [6]

La forma por la cual las proteínas pueden adoptar estos plegamientos es un gran tema de estudio,que fue abordado primero por Levinthal: este proponía que la forma por la que una secuencia de aminoácidos puede adoptar el plegamiento funcional sin pasar infinitamente por plegamientos no funcionales, Si se mira la estadística cruda, una proteína de 100 aminoácidos, donde cada uno puede adoptar 2 conformaciones tiene un total de $1.3e^{30}$ conformaciones distintas. Esto sería perjudicial para cualquier intento de supervivencia celular, y se sabe que una estructura proteica puede plegarse en unos pocos microsegundos. Es por esto que se propone que las secuencias siguen ciertos “caminos” para su correcto seguimiento.



Representación de la propuesta de Levinthal. El camino de desplegado a un plegamiento correcto es el camino entre A y N. [7]

Esta idea fue trabajada por distintos grupos de investigación en el campo, hasta consensuar la idea de que las estructuras pueden pasar por plegamientos intermedios que generan contactos transitorios, que son funcionales para “ayudar” al correcto plegado de ciertas partes de las estructuras.
Esto se puede representar por medio de un cono que presenta distintos plegamientos intermedios posibles para llegar a un solo estado nativo y funcional, el cual se encuentra en un mínimo de energía.



Representación gráfica en 3 dimensiones de los distintos caminos que puede tomar una proteína en su proceso de plegamiento para poder llegar a su estructura energéticamente predeterminada (figura de Onuchic et al. 1996). [8]

1.4 Relación estructura-función en proteínas

La relación estructura-función es una de las ideas fundacionales en Biología. Con algunas dificultades adicionales que veremos a continuación, esta idea se extiende también al ámbito molecular y en particular en el área de la Biología estructural de proteínas. De la misma forma que en ejemplos anatómicos o morfológicos (el pulgar oponible en la mano humana y su función en la motricidad fina, por ejemplo), en proteínas la relación estructura-función sostiene que la estructura define o sustenta la función de la proteína (Orengo et al. 1999). A diferencia de los ejemplos en la Biología macroscópica, dicha relación en proteínas es mucho más esquiva. Primeramente porque distintas estructuras pueden realizar la misma función (proceso denominado evolución convergente (1)) sumado a que la misma estructura puede realizar en distintos organismos distintas funciones

(proceso denominado diversificación funcional (2)). Así, la relación estructura función en proteínas se torna un problema complejo, ya que las proteínas pueden existir en diversas formas, más de las que tradicionalmente se han propuesto. En forma general, y en primer lugar, las proteínas han sido clasificadas como “globulares” o “fibrosas”, dicotomía clásica que proviene de las primeras caracterizaciones de proteínas a principios del siglo XX.

Presentación de la estructura de la lisozima (plegada, a su derecha, y desplegada, a la izquierda) por David Phillips en el Royal Institution, Londres, Noviembre de 1965. [9]

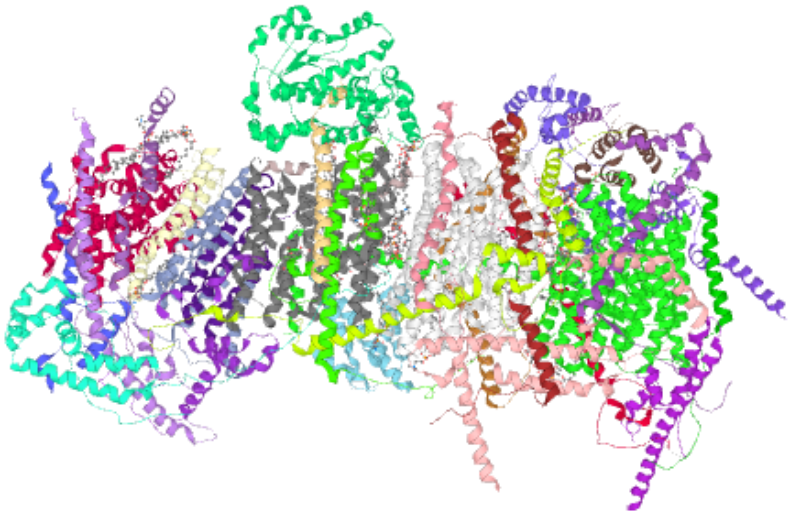
Esta dicotomía sigue siendo utilizada para derivar reglas y principios generales a pesar de los numerosos nuevos tipos de proteínas que se han ido caracterizando en los últimos 20 años. Esta aparente homogeneidad se contrapone con la enorme diversidad de tipos, formas, arreglos estructurales, composiciones y longitudes de proteínas que entran en la categoría “globulares”. Así, y con la intención de no dar una lista exhaustiva, se han caracterizado proteínas repetitivas, aquellas que tienen a nivel secuencial y/o estructural unidades en tandem que se repiten un determinado número de veces (3), proteínas con nudos en su plegamiento (4), proteínas circulares (5), proteínas pequeñas (en general con menos de 50 aminoácidos) (6) o formando enormes complejos supramacromoleculares (7), proteínas con desviaciones de composición (8) o proteínas con estados oligoméricos transitorios. En cada uno de estos casos se suma la variabilidad de la relación estructura-función que cada proteína pueda tener debido a su diversidad conformacional, pudiéndose subclasificar cada categoría en rígidas o altamente flexibles o móviles (9).

Hacia fines de la década del 90 se descubrió otro tipo de proteínas que por su alta flexibilidad no presentan una estructura tridimensional estable en condiciones nativas (de ahí denominadas “desordenadas”), desafiando por esta característica al resto de las proteínas con estructura terciaria u ordenadas (10). Más recientemente, proteínas relacionadas con las proteínas desordenadas pero que pueden cambiar de fase (las denominadas *phase transition proteins (11)*) han puesto nuevamente de manifiesto la enorme diversidad en los distintos tipos de estructura-función que pueden tener las proteínas, reafirmando de esta forma la idea que las proteínas constituyen un grupo

heterogéneo de macromoléculas (Parisi et al. 2021). Adquirir el concepto de heterogeneidad al analizar a las proteínas como un conjunto y evitar de este modo derivar conceptos o reglas globales, quizás redunde en una mejor comprensión de los mecanismos que originaron tal diversidad, su evolución y las propiedades que definen su relación estructura función.

MSSHEGGKKKALKQPKKQAKEMDEEEKAFKQKQKEEQKKLEVLKA KVVGKGPLATGGIKKSGKK

Secuencia de la proteína A0A024R1R8 (Coiled-coil domain-containing protein 72), sin estructura conocida en la Protein Data Base (cuenta con una secuencia con 100% de desorden predicho).

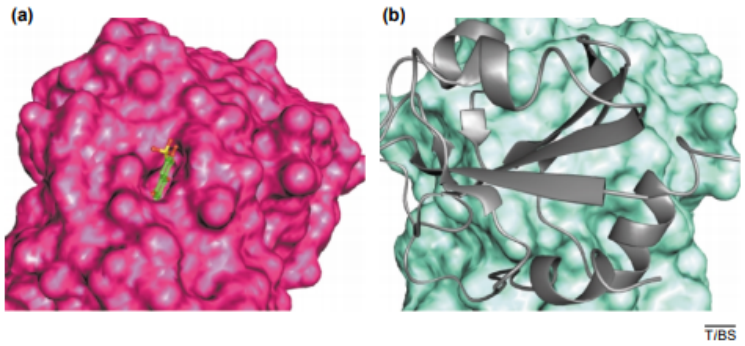


Estructura de la P03886 (NADH-ubiquinone oxidoreductase chain 1), proteína de 318 aminoácidos de longitud, que cuenta con una estructura resuelta que la cubre en su enteridad (no presenta desorden en su secuencia). [10]

1.5 Movimiento

Un parámetro importante a tener en cuenta a la hora de estudiar la correlación estructura-función de las proteínas es la diversidad conformacional. Introducida por Choita y colaboradores en 1994 (Gerstein et al. 1994), la diversidad conformacional nos permite explicar las diferencias estructurales entre confórmeros de una misma proteína, y cómo estos pueden tener distintas funciones entre sí (James et al. 2003). Esto no necesariamente nos habla sobre la relación orden/desorden de las proteínas, ya que una proteína puede tener dos confórmeros muy distintos estructuralmente, pero que sean difícilmente intercambiables. En este caso hablamos de una proteína ordenada con alta diversidad conformacional. Para que la misma sea desordenada, la interconversión de los confórmeros debe darse con alta facilidad, habilitando la rápida conversión de uno a otro.

La diversidad conformacional es la propiedad por la cual se cuantifica la diferencia estructural entre dos (o más) posibles estados estructurales de una proteína: por ejemplo, podemos tener una proteína (ordenada) con dos confórmeros muy distintos entre sí; uno puede ser la estructura unida al ligando (“Holo”) y la otra libre, sin interactuar con el ligando (“Apo”). Esta hipótesis termina de incubar la idea de la flexibilidad proteica, y cómo pueden, por ejemplo, tener distintos ligandos que interactúan con un mismo sitio activo, o cómo puede una proteína con una secuencia fija tener un gran número de interactuantes.



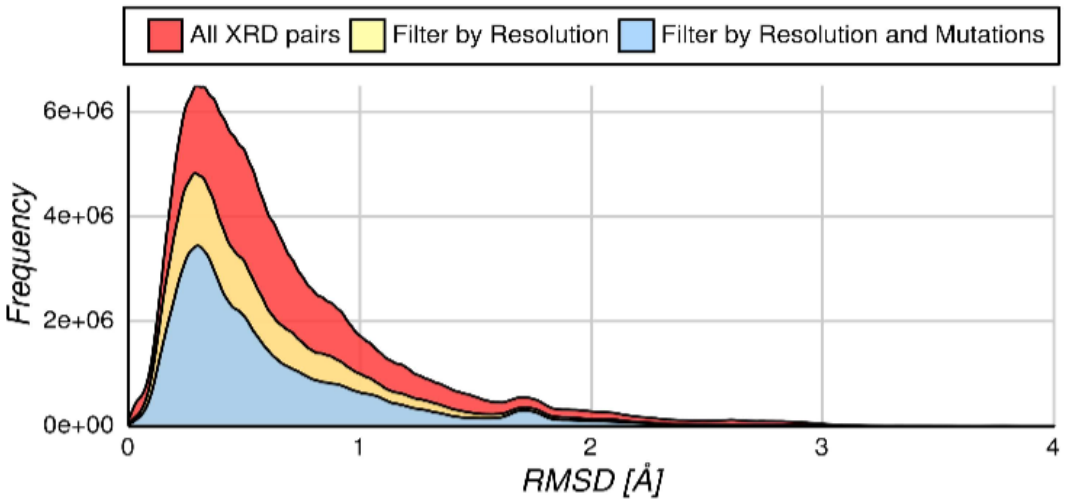
Diversidad conformacional del anticuerpo SPE7. Cada una de los confórmeros le permite unirse a antígenos totalmente distintos. (Figura adaptada de James et al. 2003). [11]

A la hora de cuantificar esta diversidad estructural es común utilizar el RMSD(Root Mean Square Deviation): este es un cálculo que, a través de una superposición de las estructuras de los distintos confórmeros (comenzando con un alineamiento de secuencias), se intenta minimizar las distancias cuadráticas medias (de ahí su nombre, RMSD) de los carbonos alfa equivalentes entre los dos alineamientos (Kufareva and Abagyan 2012, Burra et al. 2009). Esto puede ser cuantificado por la siguiente ecuación:

$$\mathrm{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\delta_i^2}$$

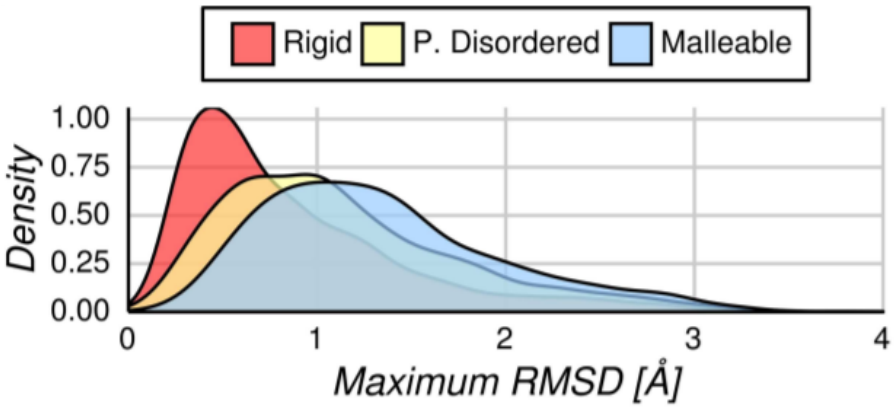
Ecuación para calcular el RMSD

Se han realizado análisis que estudian la diversidad conformacional en grandes datasets de proteínas (Monzon et al. 2017) donde se puede observar como la flexibilidad de las proteínas se correlacionan con su RMSD. En estos, se puede observar como existen 3 grandes subsets de proteínas: proteínas **rígidas,** proteínas **maleables** y proteínas **flexibles** (o desordenadas).
Para llegar a este resultado, primero es necesario filtrar las estructuras (obtenidas de una base de datos como CodNaS) que tengan baja resolución y/o mutaciones, de forma que no generen un bias en los datos. Sumando a esto, se necesita generar una homogeneidad en la información de donde proviene el RMSD (rayos X o NMR), siendo el segundo muchísimo más escaso. Una vez realizado este primer paso, es posible estudiar en más detalle cómo se comportan las distintas proteínas con respecto a su diversidad conformacional. [12]



Distribución de proteínas acorde a su RMSD y grupo al que pertenecen. Figura tomada de (Monzon et al. 2017).. [12]

Luego, es posible hacer una diferenciación dentro de esta distribución entre proteínas rígidas, maleables y parcialmente desordenadas, tomando como parámetro su diversidad conformacional. Esto se debe a que las transiciones entre confórmeros ordenados - desordenados y proteínas totalmente desordenadas tienen una tendencia a tener diversidades conformacionales más altas, siendo esto también evidenciado por el bias en la composición aminoacídica de las proteínas que forman estos dos últimos grupos (existe una tendencia hacia los aminoácidos que promueven el desorden, chequeado gracias a la comparación con la base de datos DisProt (Piovesan et al. 2017).



Grupos de proteínas acorde a su RMSD máximo. Figura tomada de (Monzon et al. 2017). [13]

Estas se ven caracterizadas por medidas experimentales de RMSD, pero también podemos identificarlas por su nivel de desorden (predicho secuencialmente) y/o ausencia de estructura.

1.5.1 La flexibilidad, la relación estructura función y la hemoglobina:

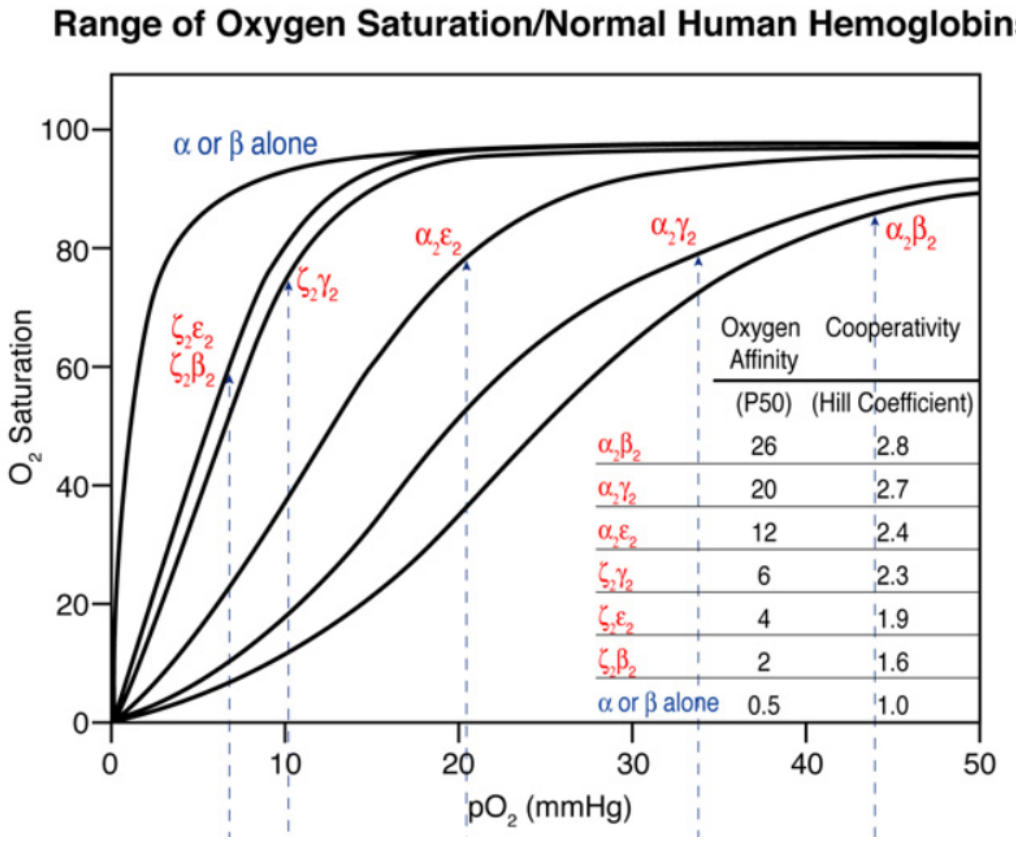
Esta es otra de las caracterizaciones importantes a tener en cuenta para entender la diversidad presente en la gran heterogeneidad que está presente en las proteínas, y es clave para entender la idea de que no todas las proteínas son iguales, y que en este tipo de diferencias conformacionales y estructurales yace la clave para entender la relación entre la alta diversidad de funciones presente en cualquier población de proteínas sintetizadas en una célula.
Todas las hipótesis y teorías anteriores, y los conocimientos de la relación estructura-función, plegamiento y movilidad de proteínas aportan a los modelos actuales: se entiende que una proteína tiene un estado nativo, en el cual es capaz de cumplir su actividad biológica, moviéndose entre una población de confórmeros dentro de un cierto rango, que se encuentra en un mínimo global de energía libre. Estos equilibrios posibles son generados gracias a la flexibilidad de las proteínas para “mover” sus estructuras terciarias y cuaternarias, de modo de facilitar los contactos entre sus cadenas y las interacciones con agentes externos. Las interacciones con sustratos, enhancers e inhibidores pueden facilitar que la población se enriquezca en un cierto confórmero, o disminuya significativamente otro, análogo a uno de los principios más básicos pero más importantes de la química: el principio de Le Châtelier (1884).

A modo de conclusión, podemos hablar brevemente de la hemoglobina humana: la hemoglobina es una proteína conocida desde 1878 (Anon 1878). La proteína está formada por un tetrámero de la forma $2x + 2y \leftrightarrows 2xy \leftrightarrows x_2y_2$, donde se unen dos subunidades idénticas entre sí, para posteriormente unirse a otras dos ya unidas. En humanos encontramos 5 subunidades: α, β, γ, δ, y ε, pero no existen todas las combinaciones de estas, solo 8 son observadas como biológicamente activas (6 de ellas funcionals, y 2 más que están asociadas a la α-talasemia, Origa 2017).
La razón de que existan distintas subunidades y combinaciones está detrás de su capacidad de ligarse/desligarse a moléculas de oxígeno en distintas etapas del desarrollo. Si se estudia la población de hemoglobinas en 3 etapas distintas del desarrollo (estado embrionario, fetal y adulto), se puede encontrar que las poblaciones de hemoglobinas cambian radicalmente:

Subunit Combinations	Name	Developmental Stage
$\zeta_2\beta_2$	Portland-2	α-Thalassemia
$\zeta_2\delta_2$	Portland-3	α-Thalassemia
$\zeta_2\gamma_2$	Portland-1	Embryonic
$\zeta_2\varepsilon_2$	Gower-1	Embryonic
$\alpha_2\varepsilon_2$	Gower-2	Embryonic
$\alpha_2\gamma_2$	F	Fetal
$\alpha_2\delta_2$	A_2	Adult
$\alpha_2\beta_2$	A	Adult

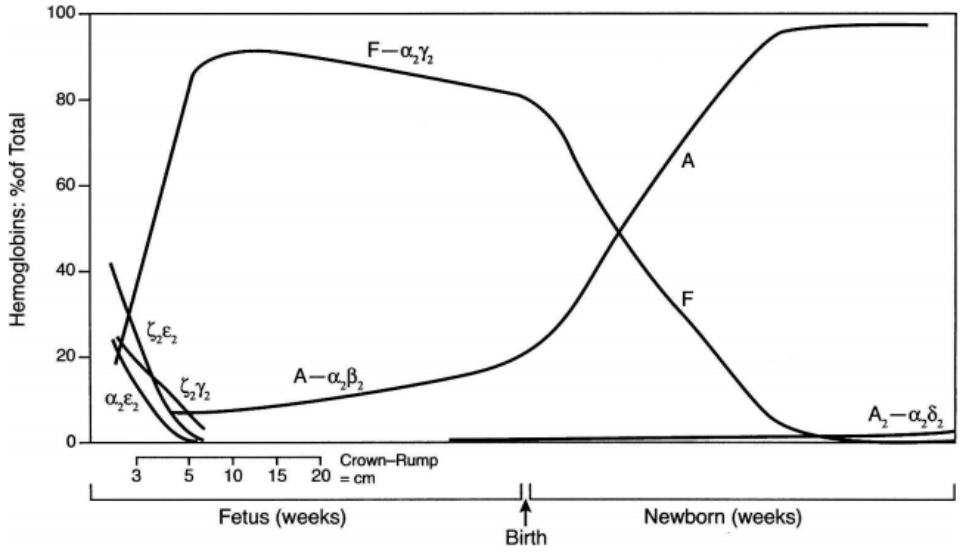
Distintos tetrámeros de hemoglobina. [14]
Imagen extraída de Manning et al. 2009

La razón de ser de estos distintos tetrámeros se debe al volumen de oxígeno disponible en la etapa de desarrollo en la que existen: hay una muchísima más alta accesibilidad al oxígeno en un estado adulto comparado a uno fetal, y una todavía menor en el estado embrionario. Es por esto que al examinar sus curvas de saturación de oxígeno, observamos como todas intentan optimizar el punto donde se da el cooperativismo entre sus estados T (Tenso, no unido al oxígeno) y R (Relajado, unido al oxígeno) son más sensibles a los cambios en la tensión de oxígeno:



Niveles de saturación de oxígeno para distintas hemoglobinas. [15]
Figura de Manning et al. 2017

Lo último a abordar sería qué causa esta diferencia de constantes de unión al oxígeno entre distintos tetrámeros: los cambios en la estabilidad del mismo. Los tetrámeros formados en las etapas embrionaria y fetal forman estructuras más sueltas (flexibles) que son capaces de trabajar con una afinidad al oxígeno más alta, debido a su capacidad de saturarse de oxígeno más rápidamente a presiones más bajas. Esto se da gracias a cambios en la secuencia aminoacídica presentes en los monómeros que los forman, diferencias en la secuencia causan que las estructuras terciarias y cuaternarias varíen entre los distintos tetrámeros y afecten su flexibilidad y cooperativismo. Las diferencias secuenciales entre la cadena beta (β) y gamma (δ) es de 39 aminoácidos (de un total de 143), con solo un cambio en la interface alostérica de las cadenas: un Asp (Ácido aspártico) en la posición 43 de la cadena Gamma en la hemoglobina fetal por un Glu (Ácido Glutámico) en la misma posición en la subunidad Beta de la adulta (Chen et al. 2000 Glu-43(β) → Asp-43(γ)). Esto se debe a que una vez se alcanza la etapa adulta (para estos parámetros se define semanas después del nacimiento), es necesario que las hemoglobinas trabajen a presiones de oxígeno mucho más altas, debido a que aumenta su disponibilidad. Si se siguieran expresando las hemoglobinas fetales, estas se saturarían tan rápido de oxígeno que serían incapaces de liberarlo; Es por esto que los tetrámeros adultos se saturan más lento, permitiendo una respuesta a los niveles del mismo más adaptada a su disponibilidad. Es también por esto que presentan estructuras mucho más rígidas.



% de Hemoglobinas (y tetrámero) respecto a la etapa de desarrollo. [16]
Imagen tomada de Hoeger and Harris 2020, capítulo 11.

1.6 Elección del dataset: el proteoma humano

"All science is either physics or stamp collecting."

"Toda la ciencia es física o filatelia."

Ernest Rutherford

El proteoma humano, del cual hablaremos en más detalle en los próximos capítulos, presenta un excelente sistema blanco para la caracterización de la diversidad estructura-función que explicamos en la sección anterior. Esto es debido no solo al alto interés científico y por lo fructífero que puede ser para el campo contar con un detallado diccionario en el cual se encuentre información secuencial y estructural correspondiente a cada una de las proteínas que lo forman, sino que también debido a que al ser un dataset amplio pero al mismo tiempo muy estudiado, presenta un gran número de secuencias curadas con funciones específicas (en gran parte) conocidas, por lo que su alta heterogeneidad también nos da una excelente forma de obtener una prueba de concepto para la hipótesis de la verdadera diversidad estructural presente en todos los organismos.

Existen varias publicaciones pertinentes a la contraparte genómica y transcriptómica de distintos linajes celulares y tejidos humanos (Shendure et al. 2017, Papalexi and Satija 2018, Hwang et al. 2018) que se ven en alza gracias al avance de tecnologías como

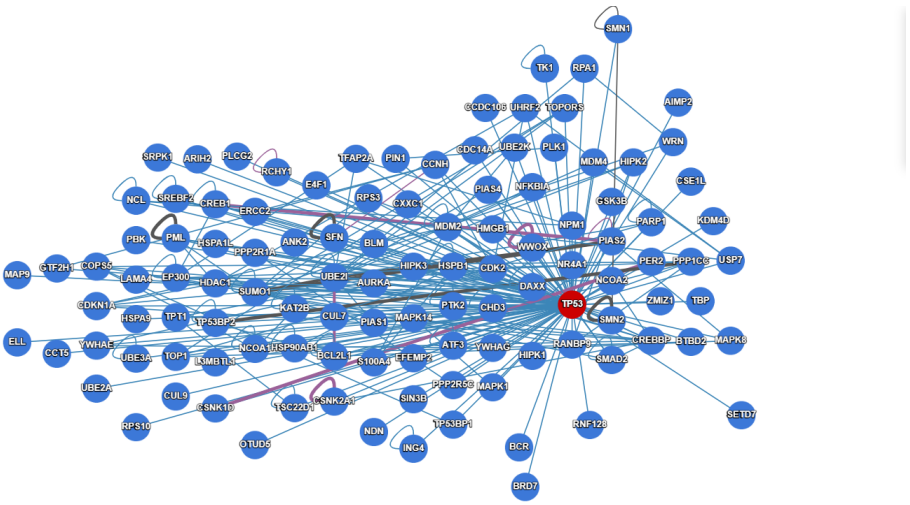
secuenciación de RNA de células individuales (single-cell RNA-Seq) , pero sin embargo no existe (al momento de la escritura) información de rápido acceso que se encargue de estudiar la parte proteica. A pesar de estar optimizado para el proteoma humano, las pipelines y técnicas utilizadas en esta tesis pueden aplicarse a otros proteomas.
Sumado a esto, con la información recolectada, es posible tener una idea más completa que relacione los patrones de expresión de distintos tejidos (18, 19) con las estructuras de las proteínas. Esto es importante debido a varias razones: existen muchas patologías que se generan por mutaciones que terminan afectando la estructura de las proteínas, evitando su plegamiento funcional (Scheckel and Aguzzi 2018), o causando que estas interactúen de manera incorrecta con otras macromoléculas en la célula (o generando nuevas interacciones que terminan siendo maliciosas para las mismas). Además, es importante investigarlo desde un punto de vista evolutivo (Palopoli et al. 2021), ya que la información de la diversidad estructural puede ayudar a entender cómo distintos grupos de proteínas que están representados por estructuras y funciones específicas pueden tener particularidades evolutivas. Para estos tipos de estudios es imperativo contar con una ardua caracterización, no solo del grupo a estudiar, si no también del resto de las proteínas, para contar con puntos de comparación y no causar un bias en el análisis de los datos.

Por último, el análisis del proteoma permite generar nodos dentro del mismo, para poder observar comportamientos similares en proteínas a las cuales no necesariamente se les han realizado estudios en los que se demuestre esto, de modo de dilucidar paralelos estructurales y funcionales dentro del proteoma y generar información de forma emergente entre distintos grupos proteicos. El mismo análisis se puede generar para dominios altamente repetidos en las proteínas (representados por la presencia de estructuras que forman gran parte de la población).

1.6.1 Ejemplos de importancia biológica al estudiar el proteoma humano

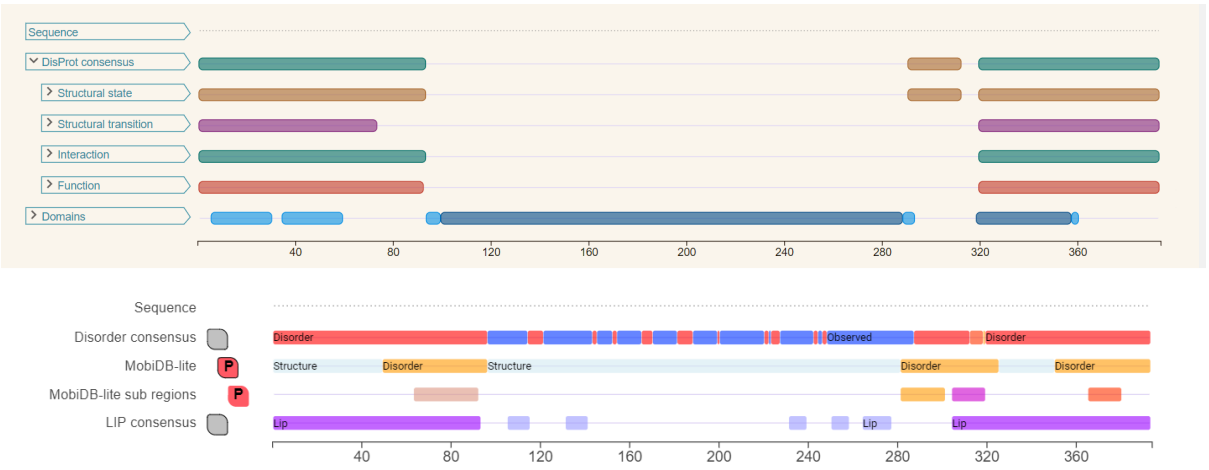
Consecuentemente, podemos comentar algunos ejemplos prácticos de proteínas conocidas para fundamentar nuestro análisis. Como indica la frase de Rutherford al comienzo de esta sección (Cita), no queremos elegir proteínas para solamente caracterizarlas superficialmente (como si fueran estampitas, o en un ejemplo más moderno, “pokemons”) si no que queremos entender la razón de ser estas proteínas, sus estructuras, sus funciones, etc. Para esto necesitamos tener un análisis más profundo y con respaldo estadístico que, idealmente, coincida con nuestras hipótesis:

A la hora de estudiar las funciones de las proteínas más relevantes y clásicas para la biología molecular del momento, como lo es el Antígeno de Tumor Celular p53, es de suma importancia tener en cuenta la estructura que presenta esta proteína: de 393 aminoácidos en longitud, p53 se divide en 3 dominios: los extremos C-terminal y N-terminal, que se encargan de las interacciones con otras proteínas (hay evidencia de interacciones con más de 100 proteínas), y un dominio de unión al DNA (DBD).



Mapa de interactuantes de p53 (fuente: HuRI) [17]

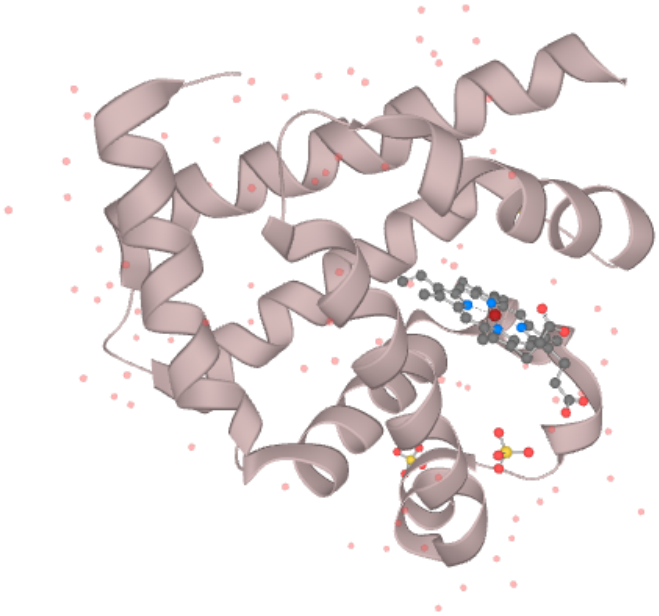
A la hora de estudiar las características estructurales de estos dominios, se puede observar que en los extremos N y C terminal de la proteína (encargados de unirse e interactuar con distintas proteínas) son desordenados, y que el dominio intermedio es altamente ordenado. Esto se puede explicar porque la alta flexibilidad que caracteriza a los dominios desordenados, les permite tener una alta población de interactuantes (113 proteínas anotadas en HuRI y 1,391 proteínas/genes en BioGRID), permitiendo que las proteínas puedan regular múltiples blancos, como es el caso en macromoléculas que se encargan de regular el ciclo celular y otros puntos clave en la vida de la célula. La proteína cuenta con una diversidad conformacional alta, teniendo el promedio en 1.1765 Å.



Mapas estructurales de Disprot (superior) y MobiDB (inferior) de p53. [18]
En ambos puede observarse las zonas con desorden predicho (extremos) y el dominio DBD (en el centro de la secuencia).

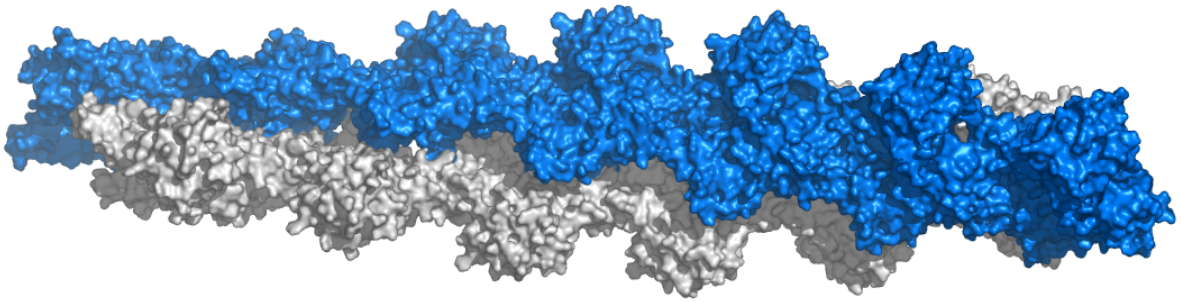
De esta manera, con este ejemplo práctico, se puede entender la intrínseca relación entre la estructura de una proteína en particular y su función.

La mioglobina, en contraste a p53, es una proteína de alta abundancia y de la cual se conoce su existencia y función desde, aproximadamente, 1887, cuando se descubrió su presencia en la orina de soldados después de que estos fueran sometidos a severos ejercicios. La estructura de la proteína, obtenida gracias a Theorell en 1942, es una de las primeras evidencias estructurales con base en cristalización de proteínas globulares (y por lo tanto uno de los culpables de la existencia del bias histórico hacia estas proteínas). Actualmente, de esta proteína se conocen tanto sus funciones moleculares, como su estructura completa: se encarga de unirse a grupos hemo y metales (en este caso, hierro), y, consecuentemente, de ligar oxígeno para actuar como carrier. Con respecto a la estructura, se puede encontrar un cristal que la encubre desde el aminoácido 2 hasta el 154 con una resolución de 1,65 Å (al tener 154 Aa de longitud, está 99,35% mapeado) y cuenta con un RMSD máximo de 2.1 y un promedio de 0.5538. De la mioglobina también se conoce que interactúa con solamente otras 9 proteínas, lo cual puede explicar su estructura, que en comparación con p53, es mucho más rígida.



Estructura resuelta de la mioglobina (3RGK) [19]

También podemos hablar de una de las proteínas más importantes en células eucariontes: la actina (P60709). Su polímero, formado solamente por la unión de monómeros de actina, está encargado de la movilidad celular y asegurar la forma y polaridad de la misma, entre otras funciones. El monómero consta de dos conformaciones: la G-Actina, que tiene estructura globular, y la F-Actina, que tiene estructura fibrilar. No es coincidencia que la población de monómeros de Actina existe como G-Actina, pero una vez que esta se polimeriza, los microfilamentos están compuestos de F-Actina, debido a su mayor rigidez estructural y estabilidad.



Representación de un polímero de 13 subunidades de F-Actina (Atomic model of the actin filament, Holmes et al, 1990). [20]

En comparación a la actina, que tiene la posibilidad de pasar de una conformación globular a una fibrilar, podemos tomar como ejemplo a la Transcortina, o globulina fijadora de corticosteroides (P08185), que es una proteína extremadamente rígida. Esta proteína, de 405 aminoácidos de longitud, cuenta con una fracción de desorden del 0.002% (es decir, solo tiene una posición que presenta desorden en su estructura). Esto es evidente a la hora de observar su RMSD máximo y promedio en, por ejemplo, la base de datos CodNaS (Monzón et al, 2013); estos tienen valores de 1.12 para el máximo, y 0.75 para el promedio. Es de esperar que la función de una proteína tan rígida sea un tanto simple: actuar como una proteína de unión para esteroides e inhibir endopeptidasas. La proteína cuenta con solo 9 interactuantes anotados en BioGRID.

Estos son solo algunos ejemplos de proteínas conocidas, en las cuales podemos apreciar la gran heterogeneidad presente en el proteoma.

1.8 Conclusión

La idea principal que hay que llevarse de este primer capítulo introductorio es que las proteínas son un conjunto de macromoléculas increíblemente complejo, y que esta complejidad reside casi exclusivamente en su amplia variabilidad.
La idea de qué “es” una proteína se contradice a sí misma repetidas veces a través del tiempo y con los avances de la ciencia. Esto se debe principalmente a la naturaleza intrínseca del campo que estudiamos (“*la ciencia sin contradicciones no es ciencia*”), pero también a qué, constantemente, estamos descubriendo que estas macromoléculas no pueden ser clasificadas con simples variables categóricas, sí no qué es posible dedicar años de estudio exclusivamente a una familia de estas y seguir sin entender completamente todo el panorama: cuando, porqué, y cómo se expresan, su función y la relación con sus estructura, sus dominios, la diversidad conformacional de estos y la capacidad de interconvertirse entre unos y otros, son algunas de las características de las cuales podemos hablar, pero, al momento de pasar a estudiar una familia distinta de secuencias (que pueden tranquilamente existir en el gen vecino), seguramente encontraríamos algo que desafía nuestros conceptos establecidos por la experiencia previa. Las bases son las mismas: todas las proteínas se expresan a partir de genes, codificados por DNA (o RNA en el caso de algunos Virus) y son expresados por medio de Ribosomas, pero una vez que su secuencia se encuentra libre, sus características, funciones, y su destino es tan amplio como el número de proteínas que existen.

Citando al poeta Estadounidense Walt Whitman, en “Song of Myself”, encontramos una analogía muy buena para describir a las proteínas:

Do I contradict myself?
Very well then I contradict myself,
(I am large, I contain multitudes.)

¿Qué me contradigo?
Sí, me contradigo. Y ¿qué?
(Yo soy inmenso, contengo multitudes.)

2. Características generales del proteoma humano:

2.1 Introducción

En este capítulo nos vamos a centrar en definir lo que es, según el consenso, el proteoma humano; explicar cómo los datos secuenciales que lo componen fueron obtenidos y caracterizar, según distintos parámetros (longitud, expresión, abundancia, desorden y movimiento), a las proteínas que lo componen.

Es importante realizar estos análisis al principio de la caracterización ya que nos permiten identificar las distintas tendencias de nuestro dataset. Por ejemplo, al analizar el volumen de proteínas desordenadas presentes en el proteoma, podemos estimar cuánto impacto tendrá esto al intentar reclutar estructuras cristalográficas por homología. También nos permite definir máximos y mínimos para con nuestros parámetros.

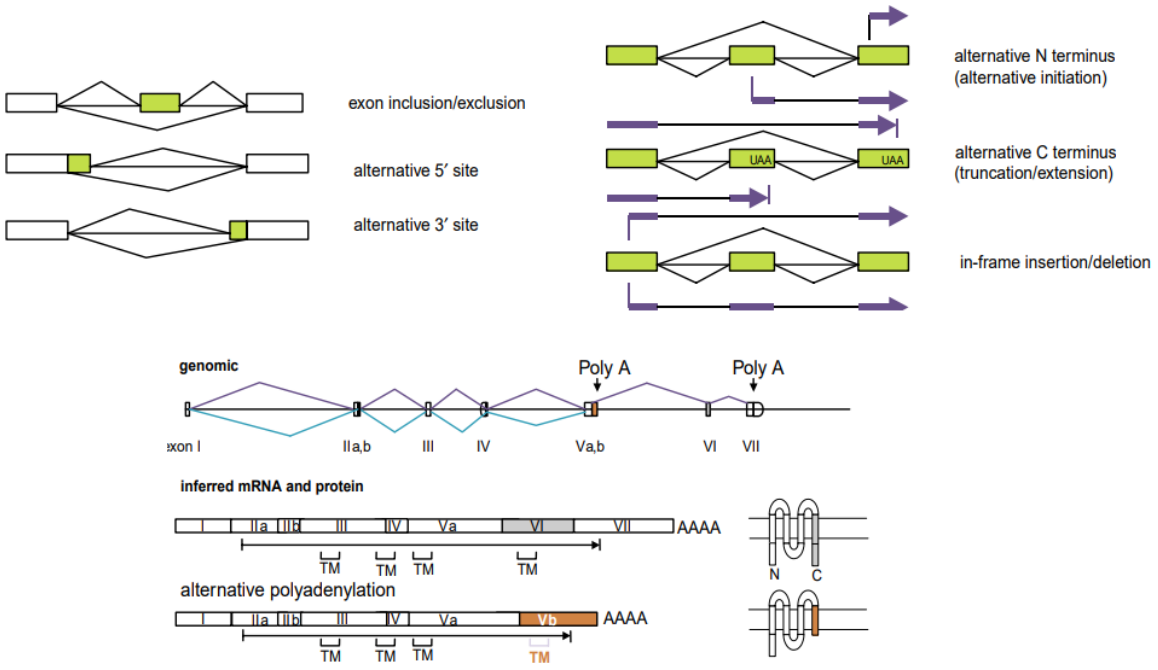
Es prioritario entender en el contexto del proteoma cuales son las proteínas más largas y las más cortas, las más expresadas y menos expresadas, etc, ya que esto nos permite realizar no solo una comparación interna, sino también tener un punto de comparación a la hora de estudiar otros proteomas.

2.2 Obtención del dataset

El primer paso para comenzar este trabajo de investigación fue conseguir el dataset del cual se derivarían todas las proteínas a caracterizar: el proteoma humano. Para esto recurrimos a UniProt, el recurso universal de proteínas(The Universal Protein Resource, UniProt Consortium 2008).
El esquema de anotación para proteínas, en general, sigue la siguiente fórmula: las secuencias son subidas a la base de datos desde fuentes externas (EMBL, GenBank, Ensembl, PDB, etc) a UniParc. En este servidor encontramos secuencias nuevas, revisadas u obsoletas. De UniPrac, las secuencias pasan a UniProtKB (protein Knowledgebase). UniprotKB está compuesto por TrEMBL ,que son secuencias anotadas automáticamente y no revisadas por humanos, y Swiss-Prot, que son secuencias manualmente anotadas y curadas a mano.
Una vez que se tiene un conjunto de proteínas anotadas como provenientes de un mismo organismo, se puede comenzar a construir su proteoma. Según UniProt, un proteoma es "un set de proteínas que se cree que son expresadas por un cierto organismo", y el servidor (UniProt) cuenta con proteomas de organismos completamente secuenciados. Además de esto, para organismos de estudio de alto interés, UniProt cuenta con la calificación de protema de referencia, esto significa que el mismo fue curado (manualmente y con algoritmos) debido a su alta importancia. UniProt también cuenta con un ID propio de cada proteoma, para diferencias entre distintos proteomas anotados sobre un mismo organismo (un Taxonomy ID, varios ID de proteoma).

En el caso del humano (*Homo sapiens* Linnaeus, 1758), Taxonomy ID 9606, cuenta con un proteoma anotado y revisado, de ID UP000005640. La última modificación es del 29 de Febrero de 2021, pero en nuestro caso trabajamos con el dataset que estaba disponible el 1 de Octubre de 2020.

El proteoma puede dividirse en dos datasets: el primero está formado por el "bruto" de proteínas totales (cuenta con un total de 77.027 secuencias proteicas) sin importar que vengan de un mismo gen, como puede ser el caso de una proteína que sufra de splicing alternativo (de un solo gen que codifica para una proteína es posible obtener transcriptos de distintas longitudes, con diferencias secuenciales que, al ser traducido, dan proteínas con diferencias secuenciales y estructurales que determinan distintas funciones, imagen [21]).



Ejemplos de splicing alternativo y su efecto en proteínas (imagen adaptada de Modrek and Lee 2002*), [21].*

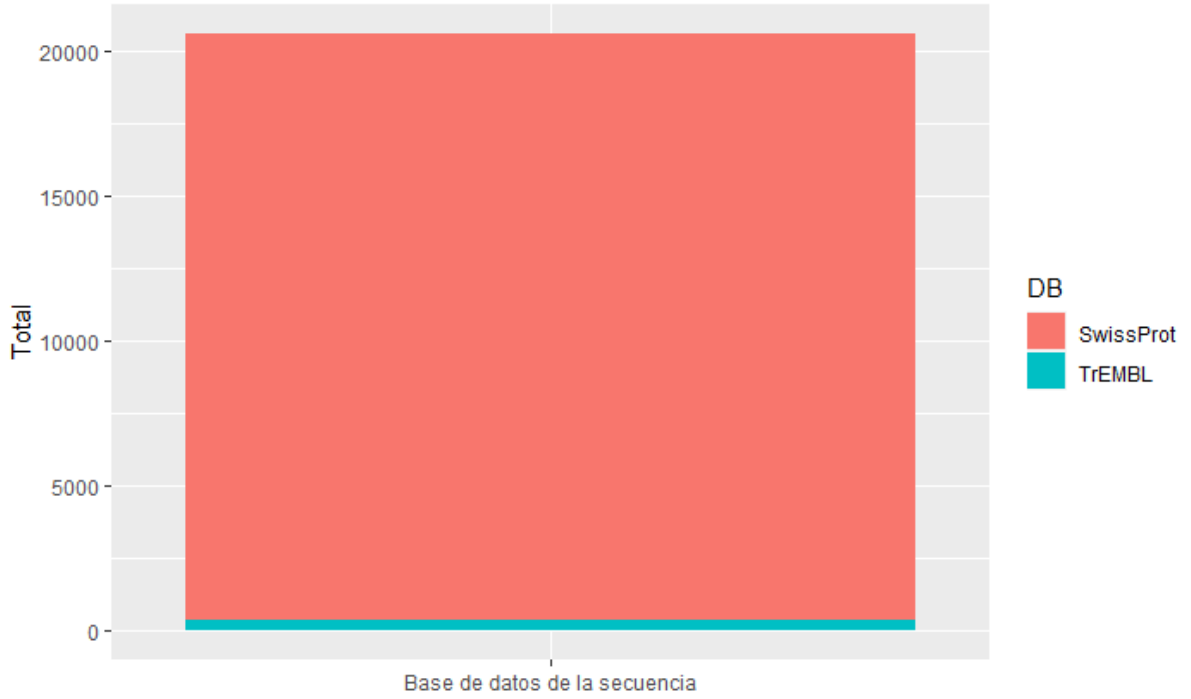
El segundo dataset disponible es llamado “Gene Count”, y provee la información del número total de genes únicos, y de un transcripto (es decir, una proteína) por gen. Esto es calculado algorítmicamente, y la elección de representante por gen se lleva a cabo priorizando la calidad de la entrada (siempre se intenta elegir un representante que sea parte de Swiss-Prot, debido a su alto estándar). Este dataset cuenta con 20.614 entradas al día de la fecha, pero en este trabajo se usó una versión de 20.600 proteínas.

La elección entre el dataset que contiene el total de proteínas y el dataset que contiene genes únicos se basó en los siguientes puntos: la alta redundancia secuencial, el ruido que probablemente se agregase a los datos debido a la alta repetición de varios genes, y a la posible repetibilidad de ciertos dominios estructurales en los mismos transcriptos provenientes de un solo gen, sumado a la posibilidad de que muchas de las entradas de proteínas no estén curadas manualmente, fueron algunas de las razones por las cuales se optó por usar el dataset más chico, aunque igual de representativo.
Esto nos deja con un dataset de 20.600 entradas de UniProt a caracterizar. El listado de los códigos Uniprot de estas proteínas se encuentra en el Anexo II.

2.3 Proteoma humano de referencia

El primer paso para el análisis de los datos fue un análisis de la información contenida en las secuencias obtenidas de UniProt. Estas entradas contienen la siguiente información:

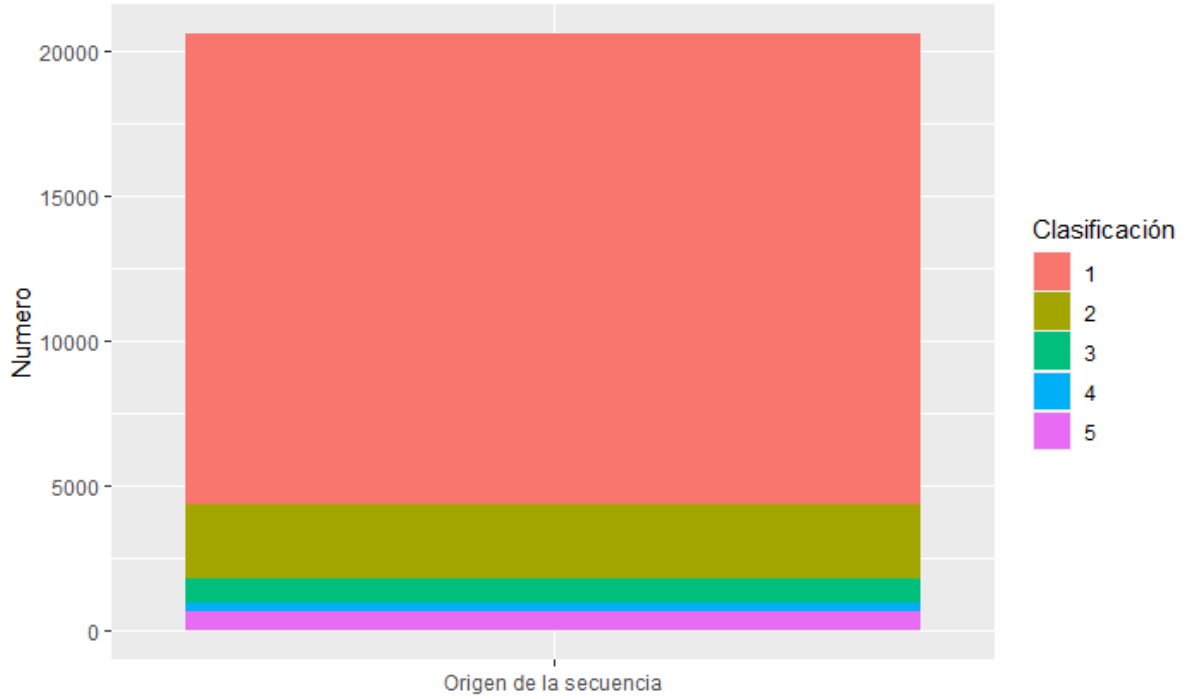
a. La base de datos de la cual proviene la secuencia (tr para TreMBL, sp para SwissProt) el uniprot ID,
b. un entry name correspondiente a la proteína, y el nombre de la misma,
c. el organismo del cual proviene y su identificador,
d. el gen del cual proviene,
e. Evidencias de su existencia (que varía en un rango de 1 a 5: 1. Evidencia experimental a nivel proteico 2. Idem a nivel de transcriptos 3. Proteína inferida por homología 4. Proteína predicha 5. Proteína incierta)
f. La versión de la secuencia.



Distribución de proveniencia de las secuencias en el proteoma de referencia. [22]

Como se observa en el gráfico [22], la gran mayoría de las secuencias anotadas en el proteoma de referencia provienen de SwissProt (20286, un 98,45%) con solo unas pocas secuencias provenientes de TrEMBL (314, solo un 1,52%).

Con respecto a la evidencia del origen de las secuencias, encontramos la siguiente distribución:



Origen del total de secuencias en el proteoma de referencia. [23]

Como podemos observar en la figura [23], la gran mayoría de las secuencias pertenecen al grupo 1, siendo este el que agrupa proteínas de las cuales existe evidencia experimental a nivel protético (un total de 16297 secuencias). El siguiente grupo es el que agrupa secuencias con evidencia a nivel de transcrito (2558 secuencias), seguidas por los grupos 3 y 5 (proteínas inferidas por homología, con un total de 790, y proteínas inciertas, con un total de 610) siendo el último grupo el 4 (proteínas predichas, con solo 345 pertenecientes a este).

```
>sp|A0A075B6I1|LV460_HUMAN Immunoglobulin lambda variable 4-60 OS=Homo sapiens (Human) OX=9606 GN=IGLV4-60 PE=3 SV=1
MAWTPLLLLFPLLLHCTGSLSQPVLTQSSSASASLGSSVKLTCTLSSGHSSYIIAWHQQQ
PGKAPRYLMKLEGSGSYNKGSGVPDRFSGSSSGADRYLTISNLQFEDEADYYCETWDSNT
```

```
>sp|O60928|KCJ13_HUMAN Inward rectifier potassium channel 13 OS=Homo sapiens (Human) OX=9606 GN=KCNJ13 PE=1 SV=1
MDSSNCKVIAPLLSQRYRRMVTKDGHSTLQMDGAQRGLAYLRDAWGILMDMRWRWMMLVF
SASFVVHWLVFAVLWYVLAEMNGDLELDHDAPPENHTICVKYITSFTAAFSFSLETQLTI
GYGTMFPSGDCPSAIALLAIQMLLGLMLEAFITGAFVAKIARPKNRAFSIRFTDTAVVAH
MDGKPNLIFQVANTRPSPLTSVRVSAVLYQERENGKLYQTSVDFHLDGISSDECPFFIFP
LTYYHSITPSSPLATLLQHENPSHFELVVFLSAMQEGTGEICQRRTSYLPSEIMLHHCFA
SLLTRGSKGEYQIKMENFDKTVPEFPTPLVSKSPNRTDLDIHINGQSIDNFQISETGLTE
```

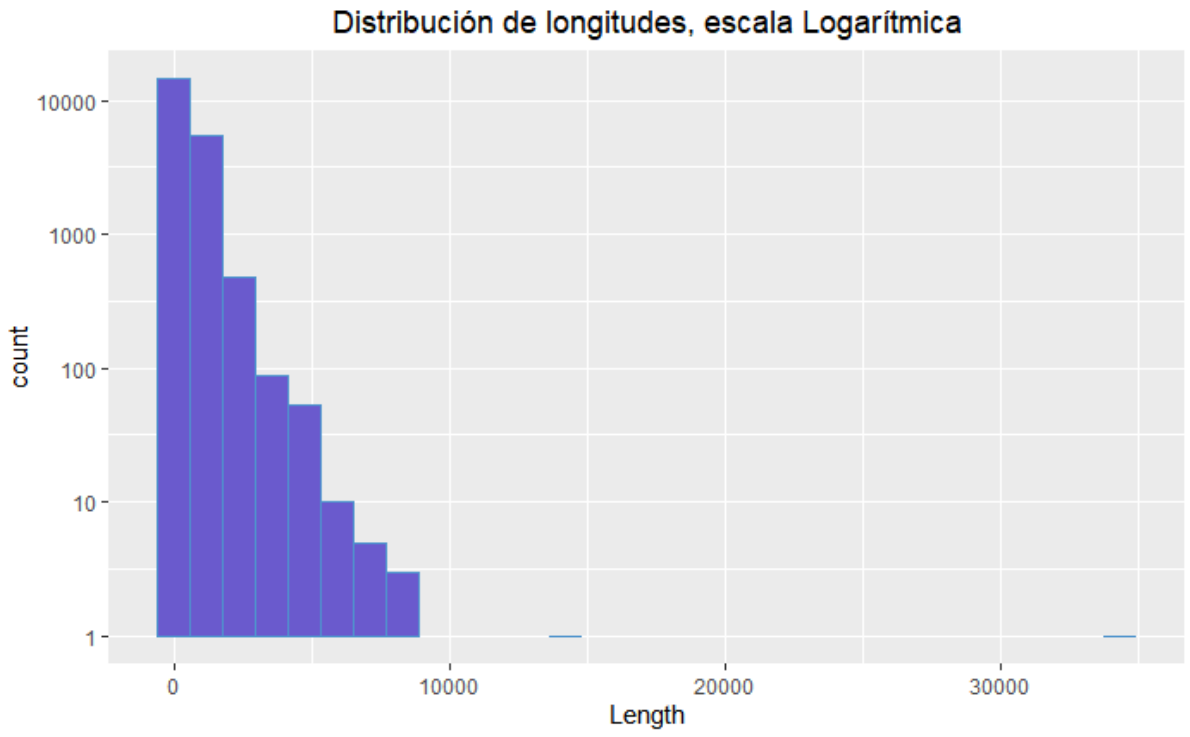
```
>sp|W6CW81|PYDC5_HUMAN Pyrin domain-containing protein 5 OS=Homo sapiens (Human) OX=9606 GN=PYDC5 PE=1 SV=1
MESKYKEILLLTSLDNITDEELDRFKCFLPDEFNIATGKLHTLNSTSSQLDLKRWHGVCS
EEDRIFQKLNYMLVAKCLREEQETGICGSPSSARSVSQSRLGLSFHGISGNAC
```

Ejemplos de entradas en formato FASTA del proteoma.

2.4 Distribución de longitudes

Como primer análisis, decidimos hacer un análisis de la distribución de longitudes de las proteínas. Esto es importante ya que de la longitud secuencial de las proteínas se pueden derivar varios aspectos relevantes de estas, como la variabilidad de funciones presentes en la misma (Brocchieri and Karlin 2005). Proteínas más largas son capaces de tener distintos dominios, permitiendo que esta cumpla, posiblemente, múltiples funciones. Esto puede ser muy favorable desde un punto de vista llanamente matemático (teniendo en cuenta también la capacidad de ciertos genes de pasar por splicing alternativo): un solo gen capaz de expresar una proteína que cumple distintas funciones puede parecer metabólicamente más favorable que tener distintos genes, cada uno regulado individualmente, que codifican para distintas proteínas. Sin embargo, esta última conjetura es la actualmente establecida. Esto se debe a los condicionamientos adicionales que puede tener un gen y una proteína: es más fácil conservar (evolutivamente hablando, a razón de mutaciones por gen) varios genes cortos que pocos genes largos (Lipman et al. 2002).

Al realizar un histograma que muestre la distribución de longitudes, nos encontramos con este resultado:

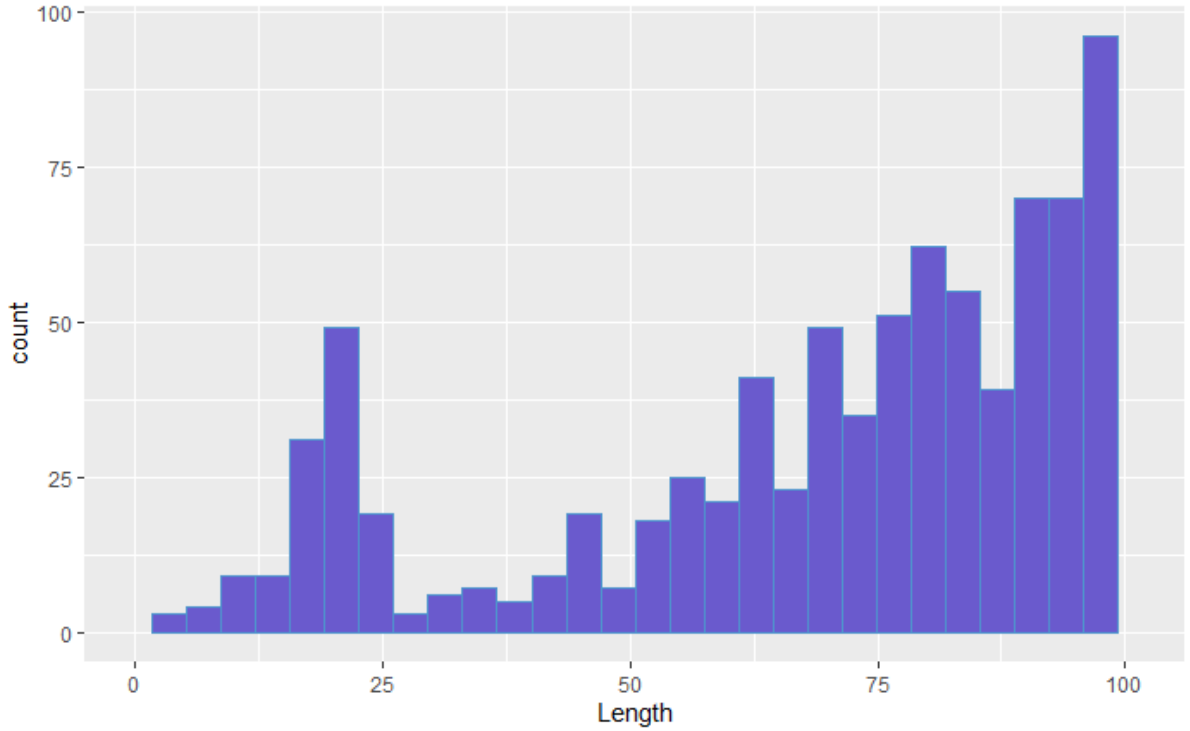


Histograma de longitudes para todo el proteoma, en escala logarítmica [24]

Se puede apreciar en el gráfico [24] que existe un número muy pequeño de proteínas con longitudes muy largas que corren la distribución hacia la izquierda, disminuyendo la resolución, pero que la mayor densidad se encuentra en rangos menores a 5000 Aa.

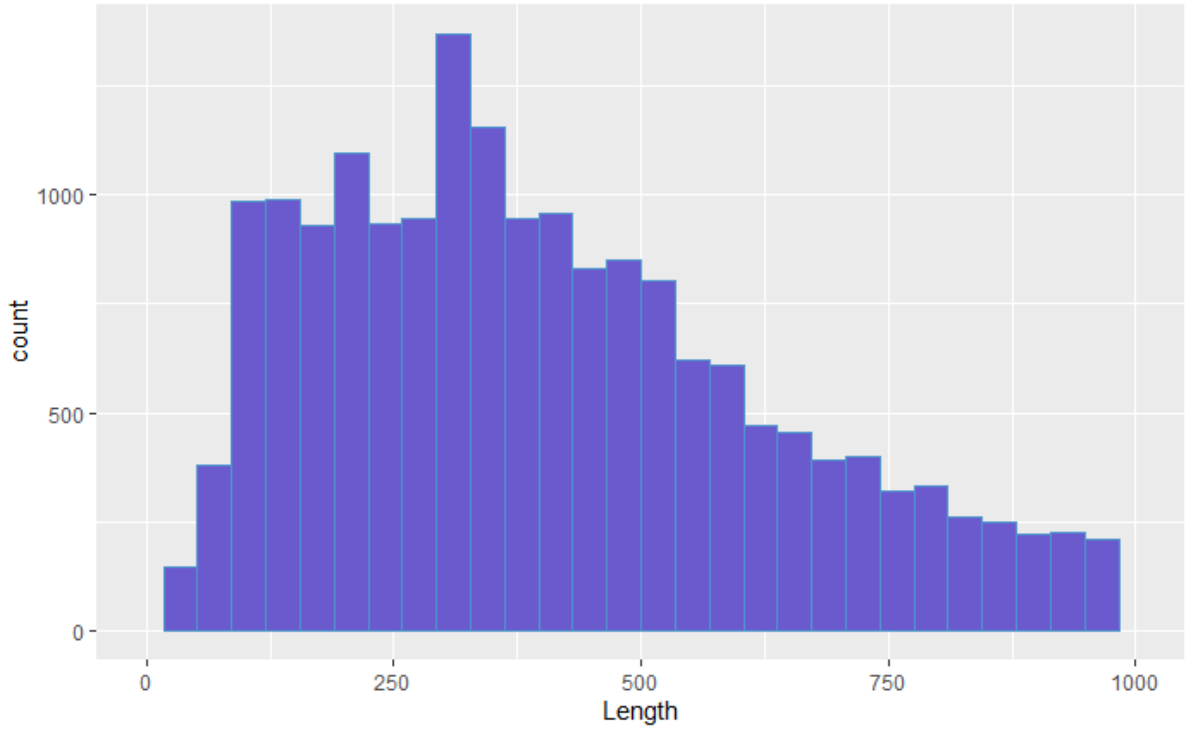
Al dividir este gráfico en detalle, podemos observar las distribuciones de proteínas entre 1 y 100, 1 y 1000, 1 y 5000, y, con el propósito de observar las proteínas más grandes, 5000 y 15000 aminoácidos de longitud:

Centrándonos en rangos más cortos (1 a 100 Aa) y el rango anterior (1 a 1000) encontramos las siguientes distribuciones:



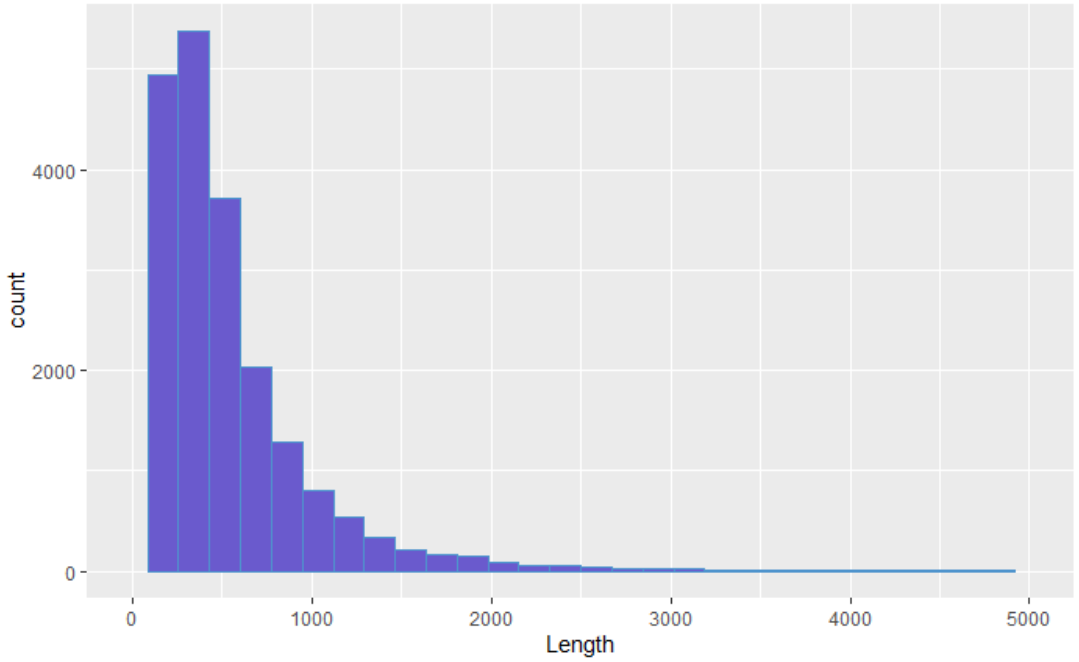
Distribución de longitudes para proteínas cortas, rango 1 a 100 Aa [25]

Al estudiar la distribución centrada en proteínas cortas (secuencias de menores a 100 aminoácidos) encontramos el gráfico [25]. El interés en las proteínas cortas ya fue demostrado por distintos autores (Storz et al. 2014) y parte desde la base de integrar péptidos pequeños (que puede tener unos pocos aminoácidos) a la definición de proteínas funcionales. Por lo pronto, podemos observar que 855 proteínas (alrededor de un 4% del total) existen en esta zona, y que la longitud media es de 78 aminoácidos, pero es de esperar que haya un número mucho más alto de proteínas cortas en el proteoma, que, por la naturaleza de las pipelines utilizadas en distintos métodos bioinformáticos de caracterización y/o por su alta dificultad a la hora de ser purificadas y caracterizadas, no estamos observando a la hora de analizar esta distribución.



Distribución de longitudes, rango 1 a 1000 Aa [26]

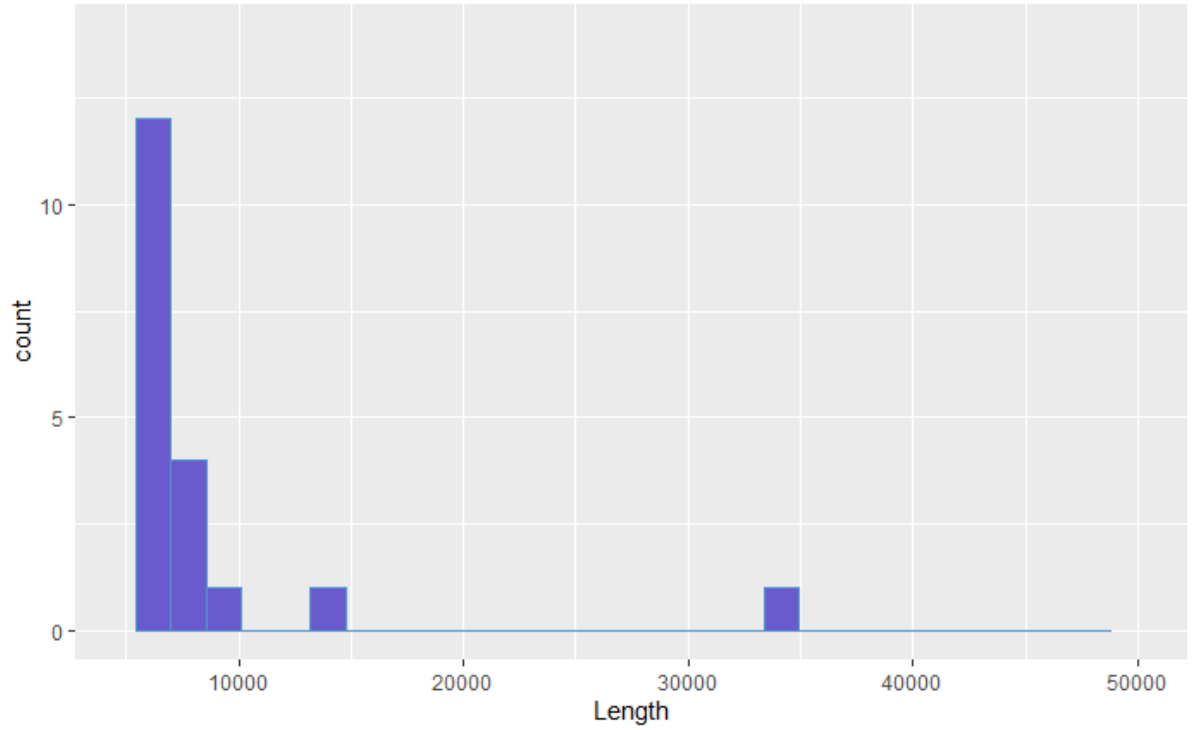
Estudiando las proteínas en un rango más “estándar” [26] (alrededor de los 300 Aa, Zhang 2000) nos encontramos con la distribución esperada: encontramos que la mayoría de la población de proteínas se encuentra en este rango, observando que el promedio es de 367 aminoácidos. 18195 secuencias (el 88,21% del dataset) se encuentra en esta zona. Este es un resultado esperado, debido a que diversos estudios (Surkont et al. 2015) demuestran que este es el tamaño óptimo de una proteína.



Distribución de longitudes, rango 1 a 5000 Aa [27]

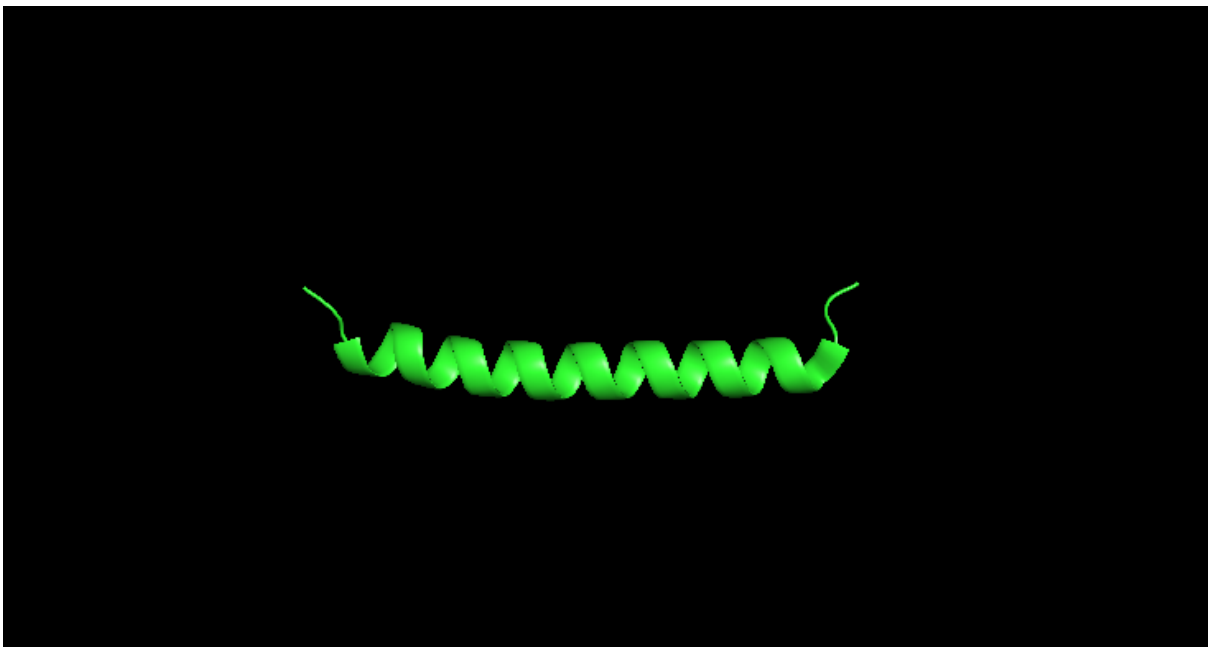
Al aumentar el rango de 1000 a 5000 aminoácidos [27], solo encontramos un total de 20567 secuencias (agregando 2372 proteínas, un aumento del 12%) pero podemos observar como los rangos de longitud empiezan a ser mucho menos poblados, indicando que hay menos proteínas con la misma longitud, siendo que están más dispersas en el rango.

Por último, podemos observar las proteínas de remarcable longitud que encontramos en el proteoma [28]:

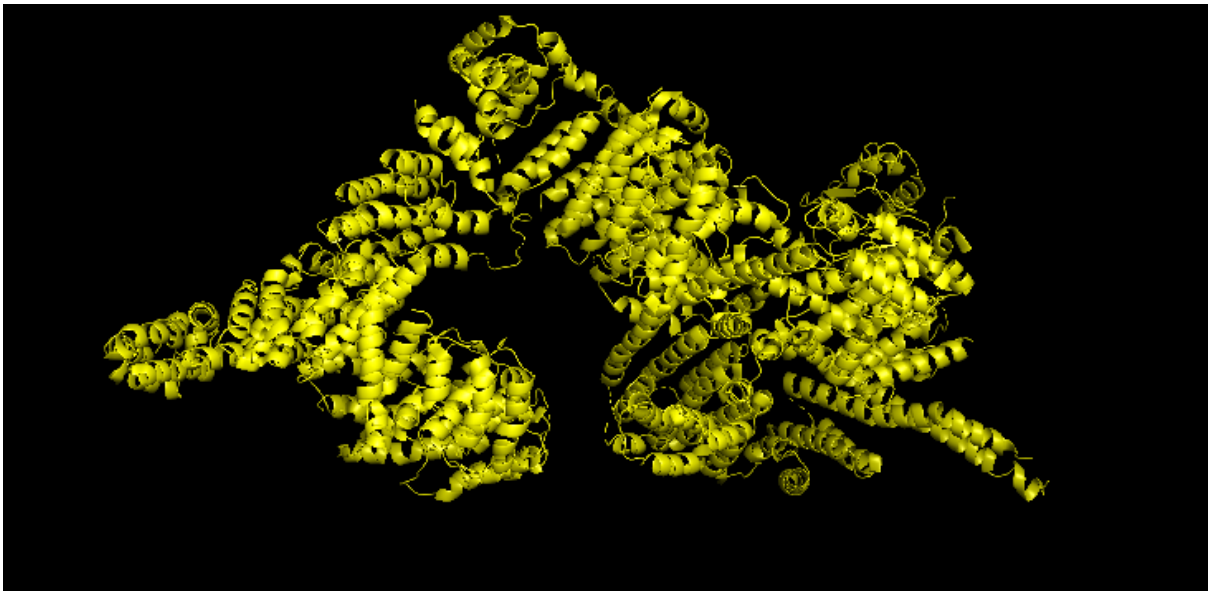


Distribución de longitudes, rango 5000 a 50000 Aa [28]

Es interesante observar que las 36 proteínas que se encuentran en esta zona no exceda los 10.000 aminoácidos, con excepción de Q8WZ42 (TITIN_HUMAN), que tiene 34350 aminoácidos, rompiendo la tendencia que se puede observar en el resto del dataset. Las proteínas que le preceden son Q8WXI7 (MUC16_HUMAN), que tiene 14507 aminoácidos, y Q8NF91 (SYNE1_HUMAN), que tiene 8797 aminoácidos de longitud. De esta forma podemos entender la alta variación presente en la cola de la distribución.
También podemos observar una alta variabilidad a la hora de observar estructuras: podemos encontrar secuencias con estructuras muy pequeñas, como es el caso de 6S70, una estructura de la proteína P0C6T2, que solo tiene 35 Aa de longitud, o estructuras mil veces más grandes, como es el caso de 5NP0, proveniente de la proteína Q13315, que con sus cadenas A y B tiene una longitud total de 3051 Aa de longitud



Estructura de P0C6T2 (6S70, cadena B), de 35 Aa de longitud [29]



Estructura de Q13315 (5NP0, cadenas A y B), 3051 Aa de longitud [30]

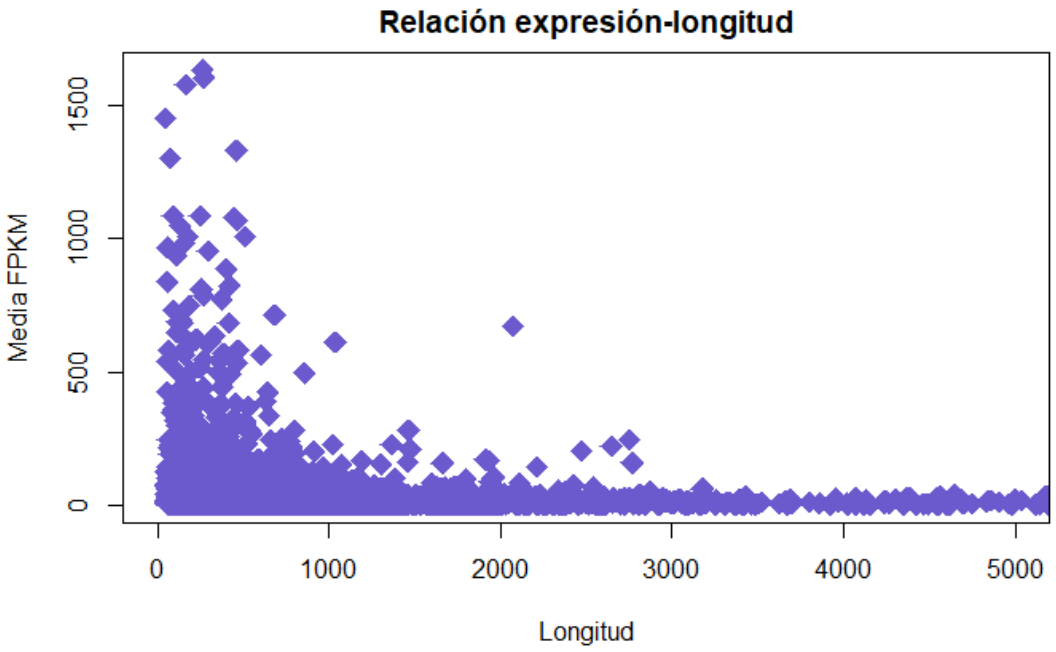
2.5 Expresión y abundancia

A la hora de caracterizar y clasificar estructuras proteicas y proteínas, se puede también incorporar el dato de qué tan presentes están en el organismo: este análisis puede hacerse a nivel de tejidos específicos o de forma general, y es posible medir cantidad de transcriptos (haciendo hincapié en obtener la información adecuada que relacione solamente el transcripto que le corresponde a una proteína dentro del proteoma de referencia) o midiendo abundancia relativa de la proteína.

Sin embargo, es importante entender que estos dos parámetros no están necesariamente correlacionados entre sí, debido a que las vidas medias de los transcriptos no son comparables con las vidas medias de las proteínas: las proteínas, una vez expresadas, pueden sufrir una serie de eventos (desde modificaciones post-traduccionales hasta ser atrapadas en gránulos de estrés) que modifican activamente su disponibilidad, por lo que es importante tener en cuenta este detalle los siguientes análisis: a pesar de que ambos parámetros (expresión y abundancia) se analizan en esta sección, no significa que indiquen lo mismo (Greenbaum et al. 2003).

En este caso, contamos con valores de tanto transcriptos en forma de FPKM (Fragments Per Kilobase of exon Model per million mapped reads, fragmentos por kilobase de modelo de exón por millón de lecturas mapeadas, obtenidos de Panda et al. 2017) y de abundancia, en forma de Partes Por Millón (PPM, tomados de la base de datos PAXDB (Wang et al. 2015)).

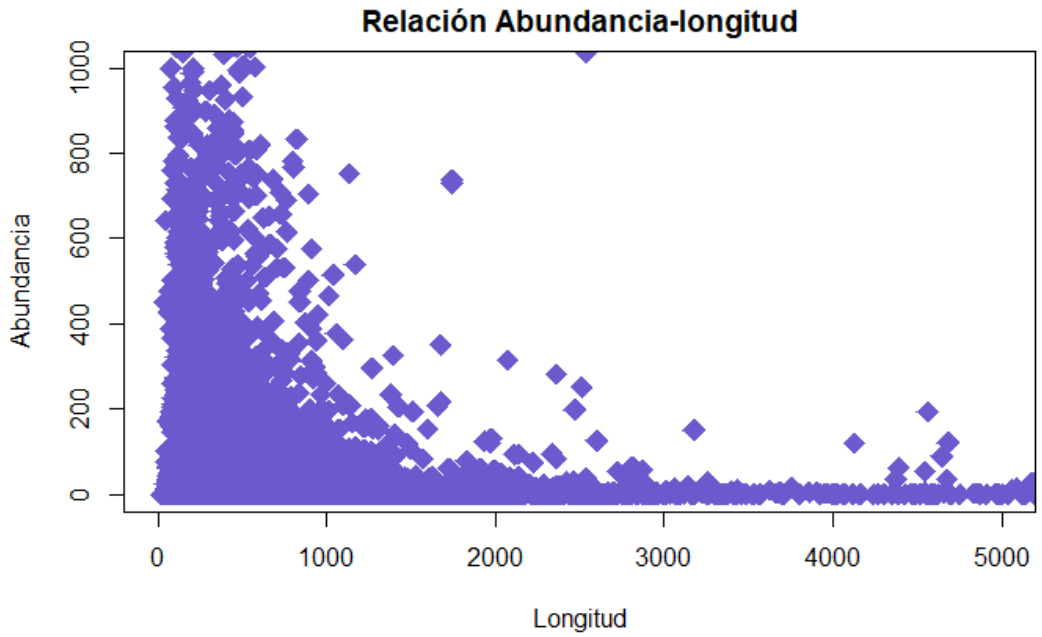
Existe evidencia de que las proteínas de longitudes más cortas tienen una tendencia a ser expresadas a niveles más altos que proteínas más grandes (Urrutia and Hurst 2003). En el caso de nuestros datos, podemos encontrar la siguiente relación al graficar los niveles de FPKM con la longitud de cada proteína:



Distribución por proteína entre la expresión y la longitud. [31]

Sin embargo, al analizar la correlación entre estos datos, no encontramos una estadística muy fuerte: un test de Pearson arroja un valor de -0.07461 con un P-value <e-09, que parece respaldar a la bibliografía, pero con un valores muy bajos.

Con respecto a la abundancia, encontramos un resultado y un análisis estadístico que muestra un comportamiento algo similar, pero con mayor libertad dentro de los datos:



Distribución por proteína entre la abundancia y la longitud. [32]

Donde la estadística nos indica una correlación negativa de -0.05897 con un P-value de <1e-06. Esto parece respaldar nuestras hipótesis y a la bibliografía, pero no son datos suficientemente sólidos para poder hacer afirmaciones.

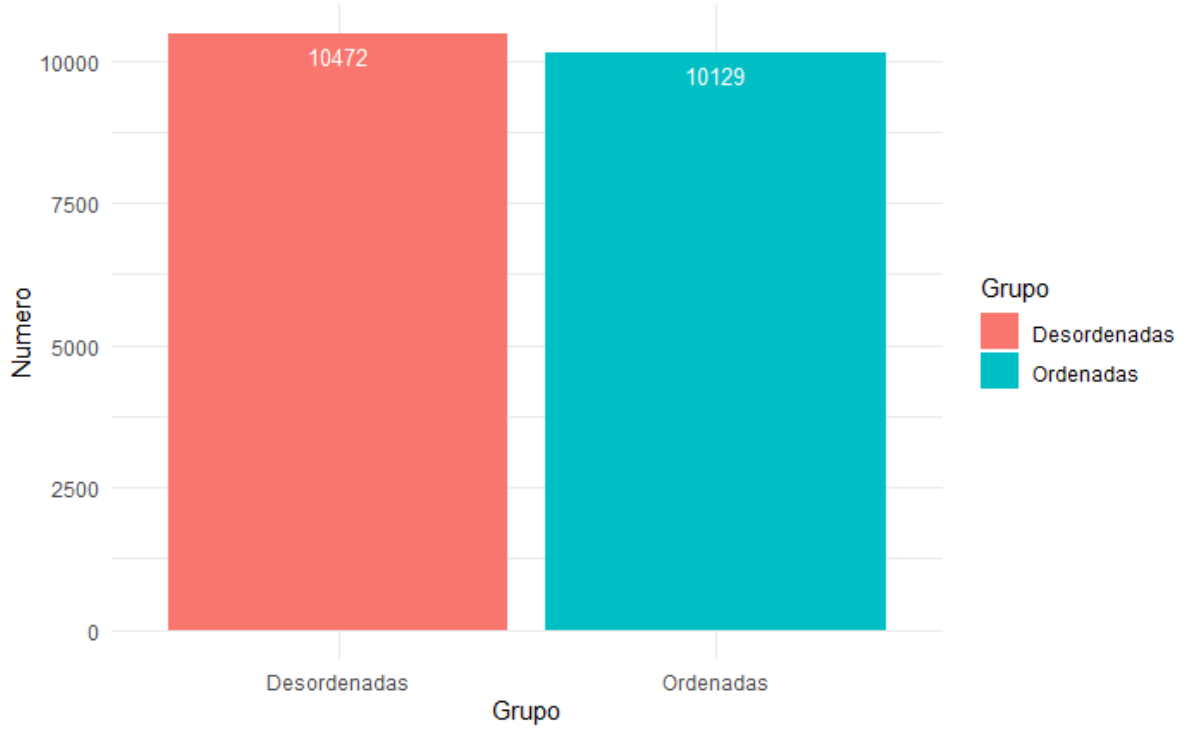
2.6 Análisis del contenido de desorden

La idea de que ciertas proteínas (o regiones dentro de las mismas) puedan desarrollar su actividad sin tener un plegamiento fijo (unfolded state), no es nueva. De una forma u otra, esta idea se remonta a fines de los 90, cuando se encontraron evidencias de proteínas con segmentos desordenados que cumplían roles funcionales en la célula (Dunker et al. 2001, Kriwacki et al. 1996). Básicamente, la idea del desorden en el contexto proteico significa que una región (o toda una proteína) no se encuentra en una estructura fija, sino que tiene una estructura con alta flexibilidad que le permite desarrollar su función biológica. Esto es una contradicción a la idea de proteínas globulares, que tienen una o dos estructuras fijas con las cuales realizan todas sus funciones, y son capaces de intercambiarse entre una y otra de forma ordenada. Las primeras evidencias del desorden surgen de estudiar cristales de proteínas, en los cuales se encontraban regiones de las cuales no se obtenian densidades electrónicas discernibles, pero aún así estas regiones eran esenciales para su función (Bode et al. 1978, Bloomer et al. 1978). Las primeras estrategias que fueron utilizadas para entender qué parte(s) de una proteína era desordenada se basaban en métodos físicos (Cristalografía de Rayos X, NRM, Dicroísmo circular y determinación del radio de giro de Stokes), mientras que los métodos más modernos usan predictores algorítmicos (que pueden tomar información adicional de las estructuras definidas de las proteínas, si las tienen) para definir qué posiciones de la secuencias proteica es o no desordenada.

Desde esa época han habido numerosos avances en materia de desorden, que involucran desde predictores en base a secuencia hasta intentar entender su origen y las ventajas evolutivas que aporta a las células (Dunker et al. 2008).

Existen diversos estudios y publicaciones científicas que demuestran el potencial de conocer la distribución del desorden en las proteínas debido a sus diversas funciones: entre estos podemos encontrar algunos que intentan explicar la correlación entre organismos que involucran una mayor cantidad de líneas celular con distintos patrones de expresión con proteomas más grandes y el nivel de desorden presente en las proteínas que los forman (Schad et al. 2011), y otros estudios que se centran en conocer los cambios estructurales generados en proteínas causantes de patologías provocadas por mutaciones en sus regiones desordenadas (Uversky et al. 2014), por nombrar algunos.

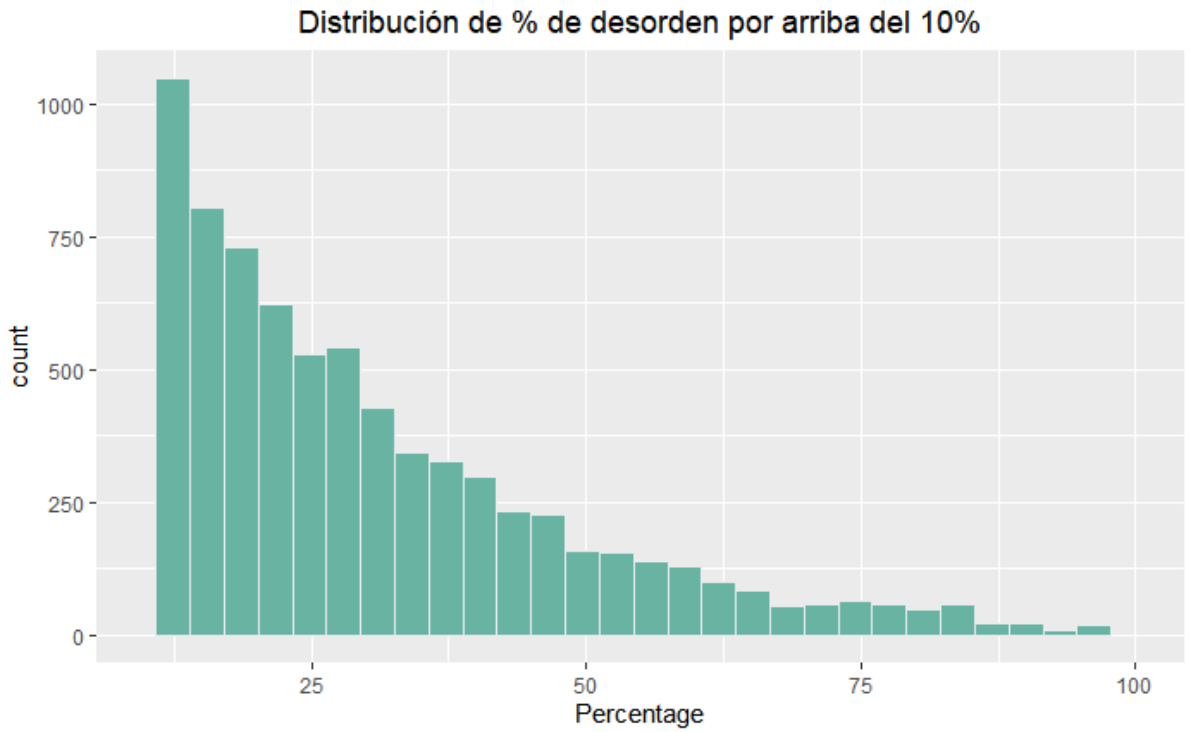
Para nuestro dataset, vamos a hacer un análisis en más profundidad y con más información en el capítulo 3, pero como primer acercamiento, usando la información obtenida de la base de datos MobiDB-Lite (Necci et al. 2017), podemos realizar un análisis exploratorio de esta característica.



Número de proteínas con al menos una posición desordenada vs proteínas sin desorden alguno. [33]

En este gráfico [33] se cuantifica el porcentaje de proteínas que presentan, al menos, un solo aminoácido desordenado (predicho por MobiDB-Lite) contra las que no presentan desorden. Se observa que el 50.82 % del proteoma presenta algún porcentaje de desorden, mientras que en el 49.16% no observamos desorden. Más adelante encontraremos resultados que contradicen estos números.

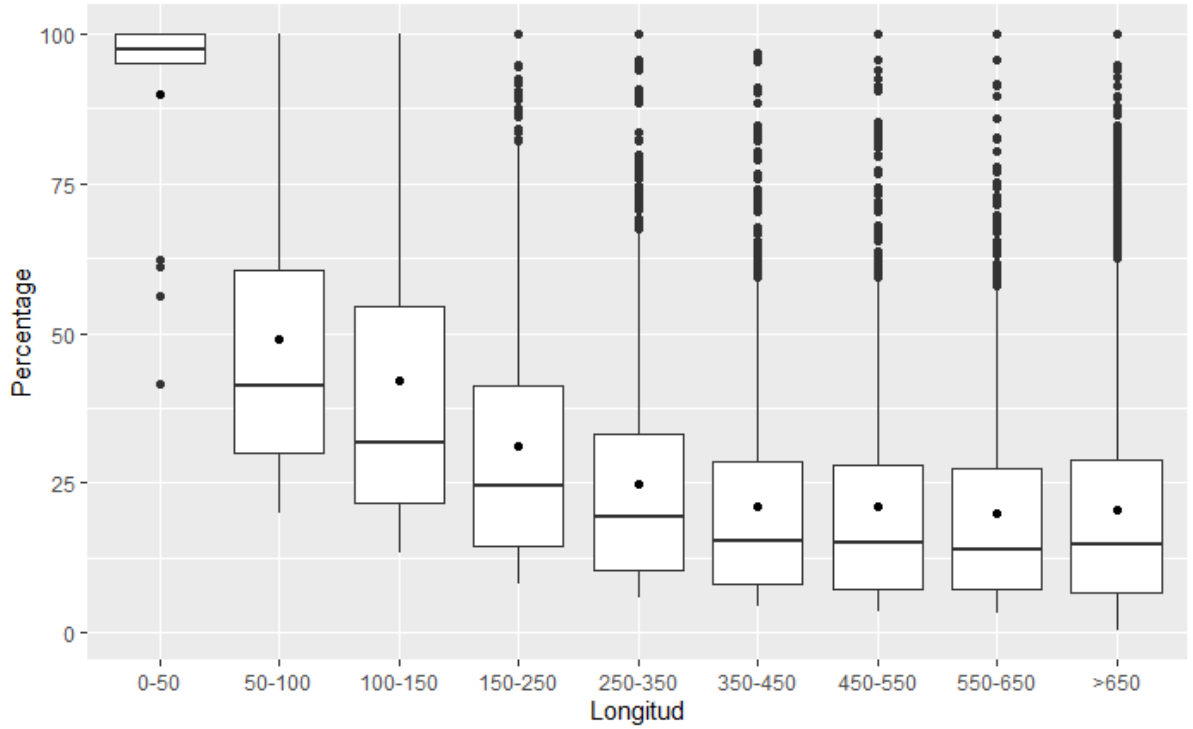
El promedio (con estos datos) cae en 23,86% de desorden dentro de las proteínas que, de nuevo, presentan desorden.
Si nos centramos en proteínas que tengan un 10% de desorden o más, observamos lo siguiente:



Histograma del % de desorden, scores de MobiDB-Lite. [34]

En esta figura, [34], encontramos que a pesar de que más de la mitad de las proteínas en el dataset presentan desorden, este está presente en porcentajes muy bajos (menores al 25%). Esto ayuda a retomar la idea inicial de la heterogeneidad del proteoma: el desorden está presente, aunque sea en un porcentaje pequeño, en muchas proteínas.

Encontramos también resultados al observar las proteínas que son 100% desordenadas: 92 secuencias caen dentro de este percentil, con una longitud promedio de 144 aminoácidos, marcando una tendencia hacia proteínas más cortas.
Haciendo un análisis más detallado, podemos estudiar el % de desorden respecto a la longitud de la proteína, de esta forma podemos estudiar cómo la estructura y los plegamientos más estables se ven relacionados con el tamaño de la proteína:



Porcentaje de desorden en función de rangos de longitudes. [35]

En este gráfico [35] vemos en el eje Y el % de desorden de la proteína y en el eje X boxes de rangos de longitud de estas (para evitar redundancia, las proteínas de más de 650 Aa fueron agrupadas). Podemos señalar como la tendencia parece marcarse de la siguiente forma: mientras más largas son las proteínas, tienen más posibilidades de plegarse establemente y adoptar estructuras fijas, de modo que el desorden parece estar favorecido en secuencias de tamaños menores, o la predicción de desorden de MobiDB-Lite tiene a ser más exacta para proteínas de rangos de longitud más cortos.

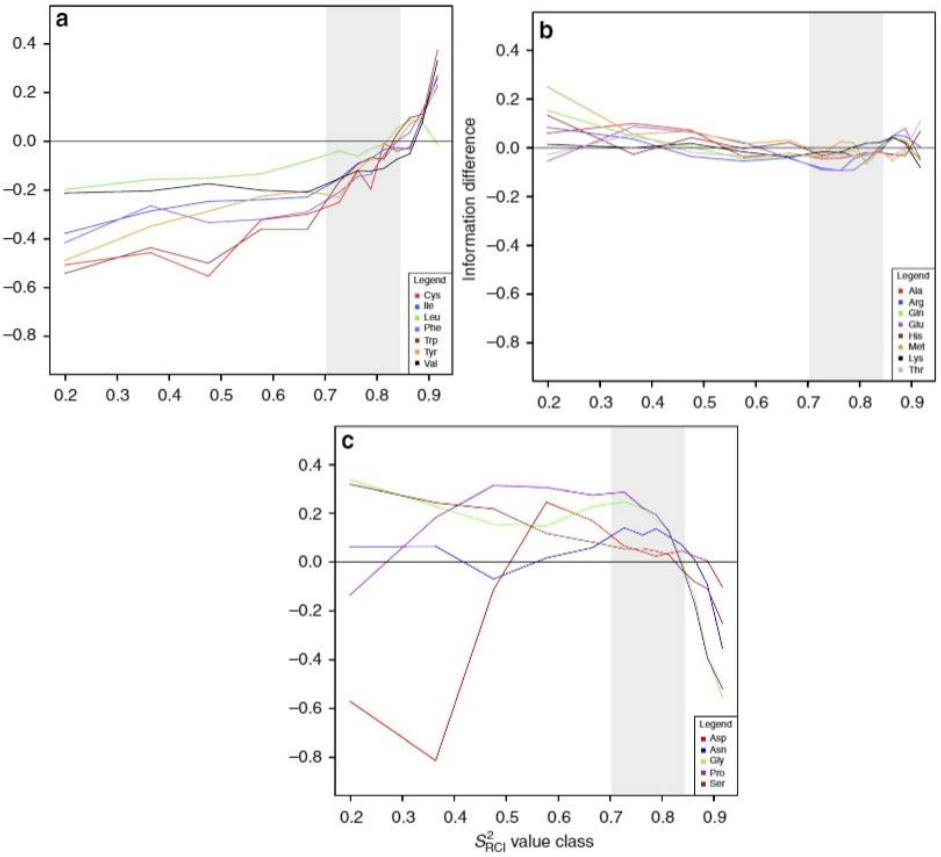
Encontramos qué hay una correlación de Pearson negativa de **-0.1242** con un P-value de **<e-38**, por lo que podemos afirmar (adhiriendo a lo descrito en la figura 35 en la sección 2.6 de este capítulo) que las proteínas más cortas parecen estar enriquecidas en segmentos desordenados.

2.7 Dinámica proteica

Por último, y para cerrar este capítulo, vamos a hablar del rango de flexibilidad que encontramos en el proteoma. Como ya hablamos en la sección 1.5, la flexibilidad de las proteínas es algo crucial para entender su funcionamiento. A la hora de estudiar este fenómeno, existen distintos tipos de estrategias: si contáramos con un set de estructuras distintas para cada proteína, podríamos obtener los RMSD de cada una de estas para estudiar la distribución de movilidades en el proteoma humano. Si bien determinar la flexibilidad contando con varias estructuras por proteína sería lo ideal, también sería muy costoso desde el punto de vista computacional y/o experimental.

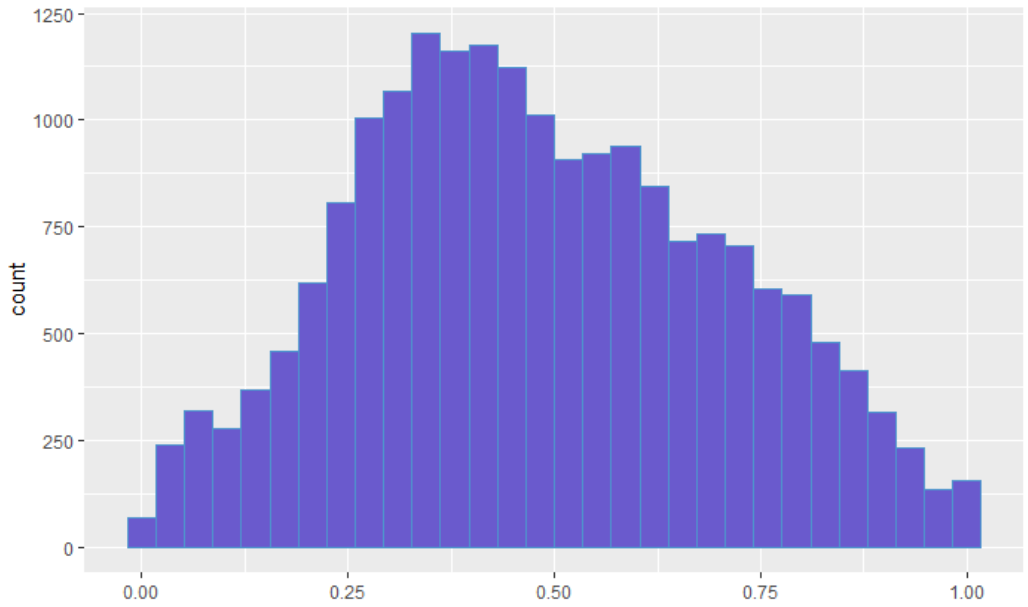
Sin embargo, podemos estimar dicha distribución utilizando métodos que usan información secuencial. En nuestro caso, usaremos datos del predictor DynaMine (Cilia et al. 2013). El método por el cual el algoritmo predice dichas regiones consiste en estimar la propensidad al orden, desorden o neutralidad mediante parámetros de N-H S^2, estimados por valores de shift químico gracias al software Random Coil Index (RCI, Berjanskii and Wishart 2008). Esta predicción se hace a nivel de residuo y se obtiene un score de movilidad por posición de cada proteína.

La tendencia de los aminoácidos tiene a obedecer el siguiente orden: Cys, Phe, Ile, Leu, Val, Trp y Tyr son clasificados como ordenados, Ala, Glu, Lys, Met, Gln, Arg y Thr son neutrales, y Asp,Gly,His,Asn,Pro y Ser son desordenados.



Valores de S^2 para un dataset fijo de proteínas utilizadas para clasificar proteínas, figura adaptada de Berjanskii and Wishart 2008. [36]

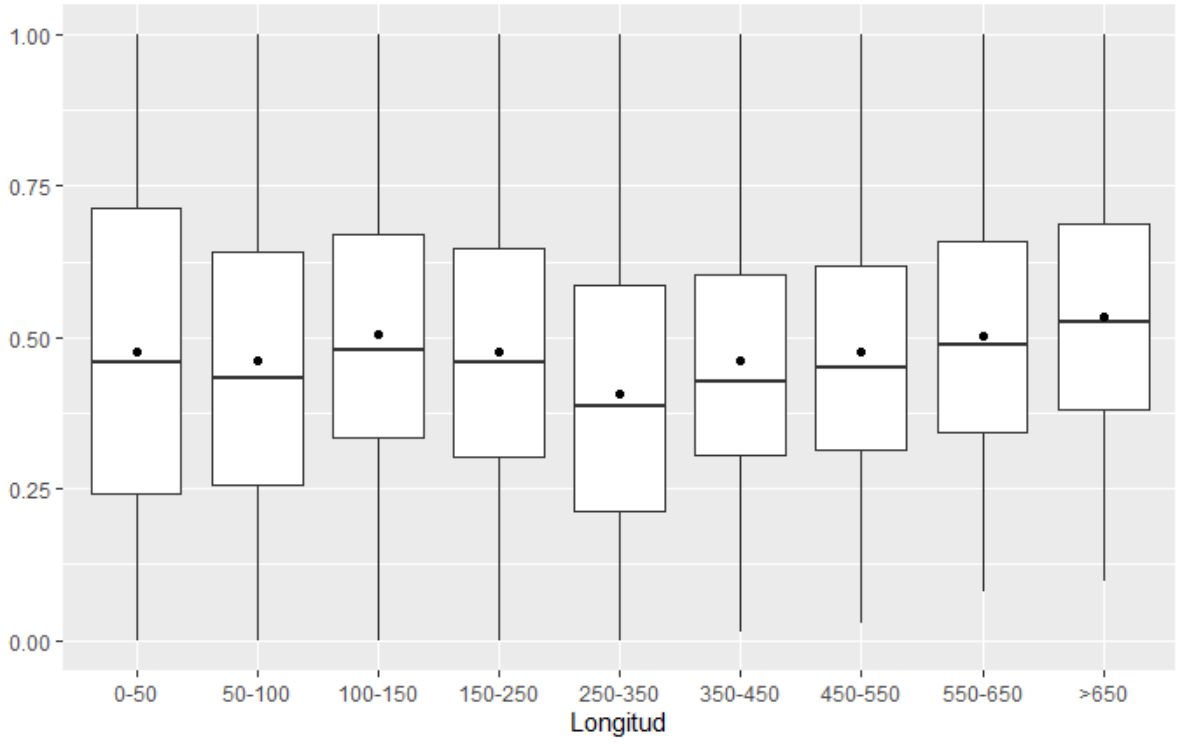
Utilizando las predicciones para todas las proteínas del proteoma humano, procesamos los valores individuales de DynaMine por posición de cada proteína sumando todas las posiciones que tengan un resultado mayor a un cierto umbral (0.200 en nuestro caso, llegando a un máximo de 1) como un valor de 1, y analizando la relación entre este y la longitud de la proteína. De esta forma podemos identificar los sitios más desestructurados en el contexto de la proteína.
Siguiendo este esquema, obtenemos la siguiente distribución:



Número de secuencias vs el % de posiciones con mayor a 0.200 valor de DynaMine sobre la longitud. [37]

Que nos muestra una distribución semejante a una normal [37], observando una tendencia a que la gran mayoría de las proteínas contengan scores por arriba del cut-off en el 50% de sus aminoácidos.

Si combinamos este análisis en boxes de longitud, podemos ver que tiene una tendencia a mantenerse en una línea cerca del 50%:



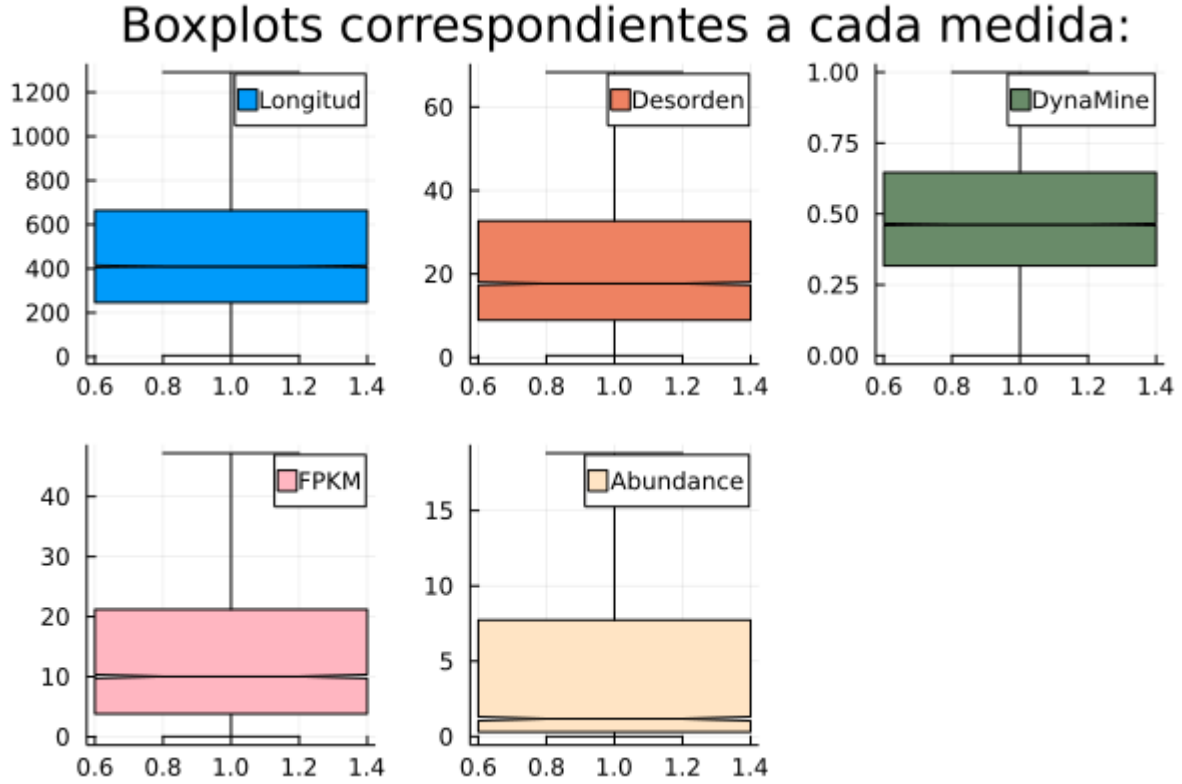
% de posiciones con score > 0.200 en DynaMine / la longitud en boxes discretos. [38]

Lo que, en principio, indica que las proteínas tienden a tener una flexibilidad que se compensa con estructura bastante distribuida, sin importar la longitud de la proteína.

En este caso, tenemos una correlación de Pearson con un valor de **0.7077** y un **P value** de **<e-99**, que arroja una fuerte correlación entre proteínas desordenadas y con altos scores de DynaMine. Esta correlación era de esperarse y nos habla de la calidad de ambas medidas, al obtener el resultado esperado podemos confirmar que la calidad de las secuencias es buena: no tenemos un desbalance entre estructuras rígidas o flexibles.

2.8 Conclusión

En este capítulo estudiamos características categóricas para una proteína: expresión, longitud, abundancia, desorden y dinámica. Además, nos permiten definir en base a estos estándares una proteína “promedio”: usando estos datos podemos afirmar que lo más estadísticamente probable es que una secuencia dada del proteoma tenga una media de 533 aminoácidos, un porcentaje de desorden del 24% (145 aminoácidos con scores de desorden positivos) y un score de DynaMine del 48%. También, que tiene un nivel de expresión de 21 FPKM, y una abundancia de 33,87 (PPM) :



Distribuciones características del proteoma humano. [39]

Estas variables servirán para contextualizar a los distintos tipos estructurales de los cuales hablaremos más adelante, y nos van ayudar a entender qué tan distintas son estás estructuras con respecto a la proteína “tipo”, pero es importante tener en cuenta que estas clasificaciones se basan casi exclusivamente en análisis secuenciales de las proteínas (debido a qué el desorden cuenta como motivo estructural). Por lo tanto, de esta forma generamos herramientas para poder hacer un análisis más profundo en relación a los distintos tipos de esculturas: saber si estos dominios ocupan toda la secuencia, si están caracterizados por tener un número alto de confórmero o si solo tienen muy pocas o una sola conformación posible, si están expresadas diferencialmente, etc, y también cómo se correlacionan estas variables con el tipo de estructura: en esta sección pudimos estudiar en detalle los scores de desorden el predictor MobiDB-Lite y observar como existe una correlación positiva y cercana a 1 con los scores de DynaMine, y como también parece

haber una anti-correlación con la longitud de las proteínas y el desorden. Este tipo de análisis son de suma importancia, debido a que nos ayudan a caracterizar y personalizar a los distintos tipos de estructuras, entendiéndolas como una variable compleja y no solamente como una etiqueta que acompaña a la secuencia.

Este pequeño análisis nos familiariza, no solo con el dataset, si no también con el procedimiento estadístico que vamos a llevar a cabo una vez tengamos un panorama estructural más completo del proteoma, y hayamos asignado distintos tipos de estructuras a estas.

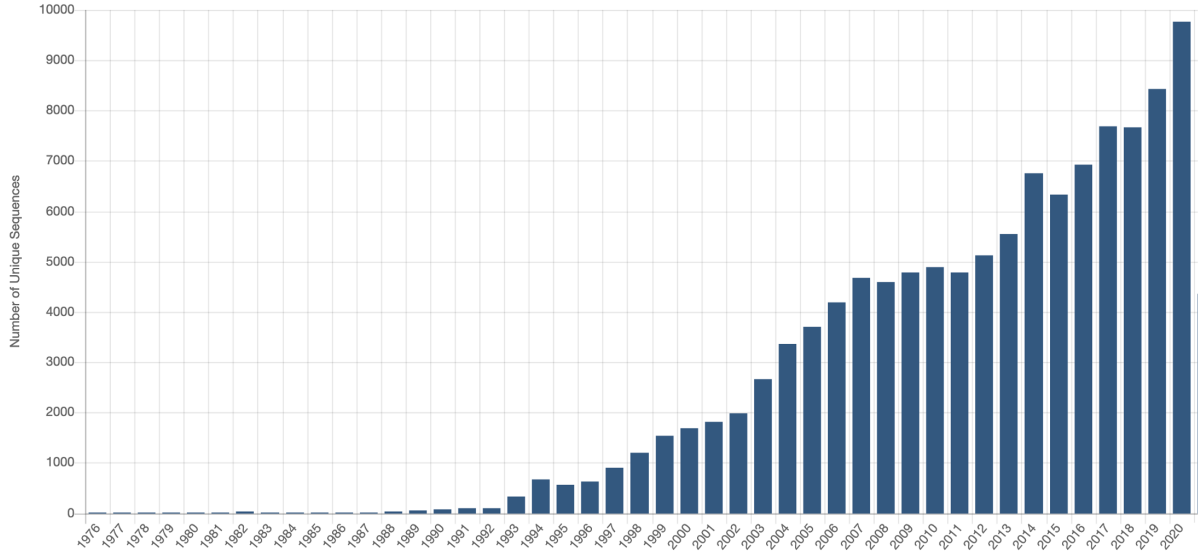
3. Estudio del estructuroma humano:

3.1 Introducción:

Cuando hablamos de estructuroma nos referimos a la colección de estructuras asignadas a sus distintas proteínas de un determinado organismo, en nuestro caso *Homo Sapiens*. Estas entradas en la PDB pueden provenir de diversos experimentos tales como RMN (Resonancia Magnética Nuclear), Difracción de Rayos X, Cryo Em (Criomicroscopía electrónica), etc. Estas distintas representaciones ofrecen distinto detalle de la relación estructura-función de las proteínas. Por ejemplo, la determinación estructural utilizando RMN ofrece aspectos dinámicos en solución que sólo lo ofrece la comparación de distintas cristalizaciones de una misma proteína en distintas condiciones utilizando difracción de rayos X. Otro ejemplo es que la Cryoelectro-microscopía ofrece la posibilidad de caracterizar la estructura de grandes complejos proteicos difíciles de estimar utilizando NMR o cristalografía de rayos X (Hebert 2019).
El conjunto de estas estructuras ofrece una gran cantidad de posibilidades de estudio para profundizar nuestro conocimiento de la Biología de las proteínas. De esta forma, contar con estas representaciones estructurales y estimaciones dinámicas nos habilita a conocer, por ejemplo, los sitios de unión o sitios activos de las proteínas, enzimas o transportadores; definir mecanismos de reacción catalíticos; modos de acción de moduladores a alostéricos; activación o inhibición por modificaciones post-traduccionales; diseñar computacionalmente inhibidores selectivos de determinadas proteínas; explicar el mecanismo de mutaciones somáticas que conducen a patologías; dar las bases estructurales de la herencia y sus determinantes fenotípicos, etc. Estos pocos ejemplos de un enorme abanico de posibles estudios pretende mostrar la enorme importancia de contar con el estructuroma humano o de cualquier otra especie.

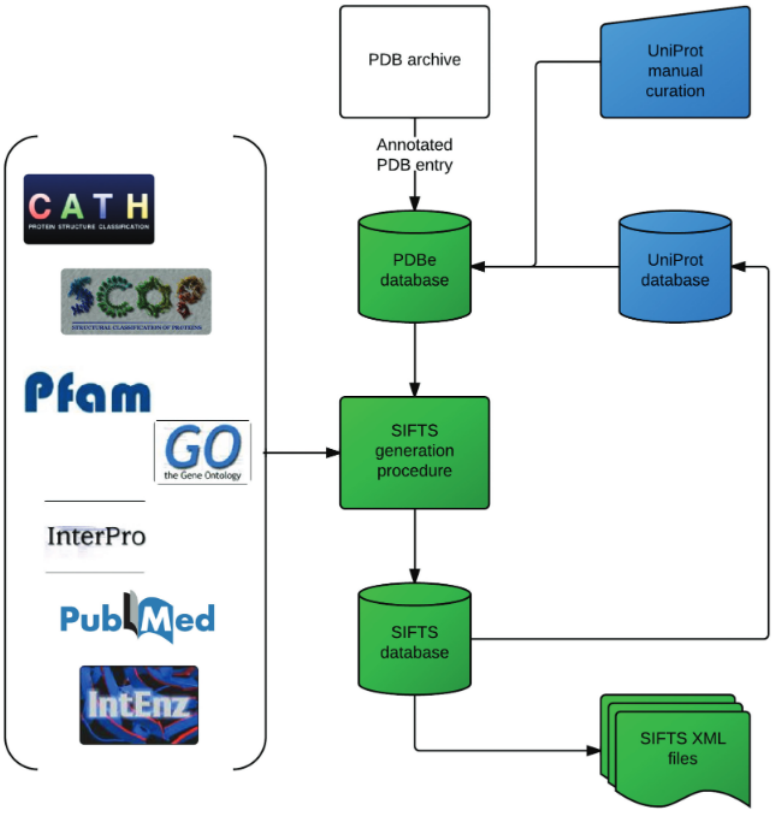
3.2 Estructuras basadas en evidencia previa

En esta parte del trabajo estamos interesados en detectar cuántas de las proteínas del proteoma humano tienen al menos un representante total o parcial con estructura conocida y depositada en la base de datos PDB. Si bien la PDB contiene alrededor de 180 mil estructuras, muchas de ellas pertenecen a la misma proteína. Utilizando los análisis estadísticos de PDB tenemos que en la actualidad esas 180 mil estructuras sólo representan ~10000 secuencias proteicas únicas.



Número de secuencias únicas en función del año (extraído de PDB). [40]

Igualmente, de las estructuras disponibles, aproximadamente 2700 corresponden a *Homo Sapiens*. Sin embargo, debemos aclarar que muchas de estas estructuras corresponden a fragmentos o dominios de proteínas, pudiendo derivar en este sentido una representación parcializada de la estructura de una proteína. De esta forma, para comenzar a caracterizar el estructuroma humano, se procedió a una búsqueda de bases de datos estructurales que contengan información específica del proteoma humano y su relación con las entradas PDB. En un primer intento, utilizamos la información otorgada por la base de datos SIFTS (Dana et al. 2019 ,Velankar et al. 2013). En esta base de datos, procedente del EMBL (European Molecular Biology Lab), en principio encontramos información completa sobre estructuras asignadas a secuencias de UniProt ID's, con sus coverages y los residuos que no están mapeados. Además SIFTS anota estructuras provenientes de secuencias sólo si tienen un 90% de identidad, y, de no provenir del mismo organismo, tienen que tener un ancestro común no menos de dos niveles de separación entre especies en un árbol taxonómico.



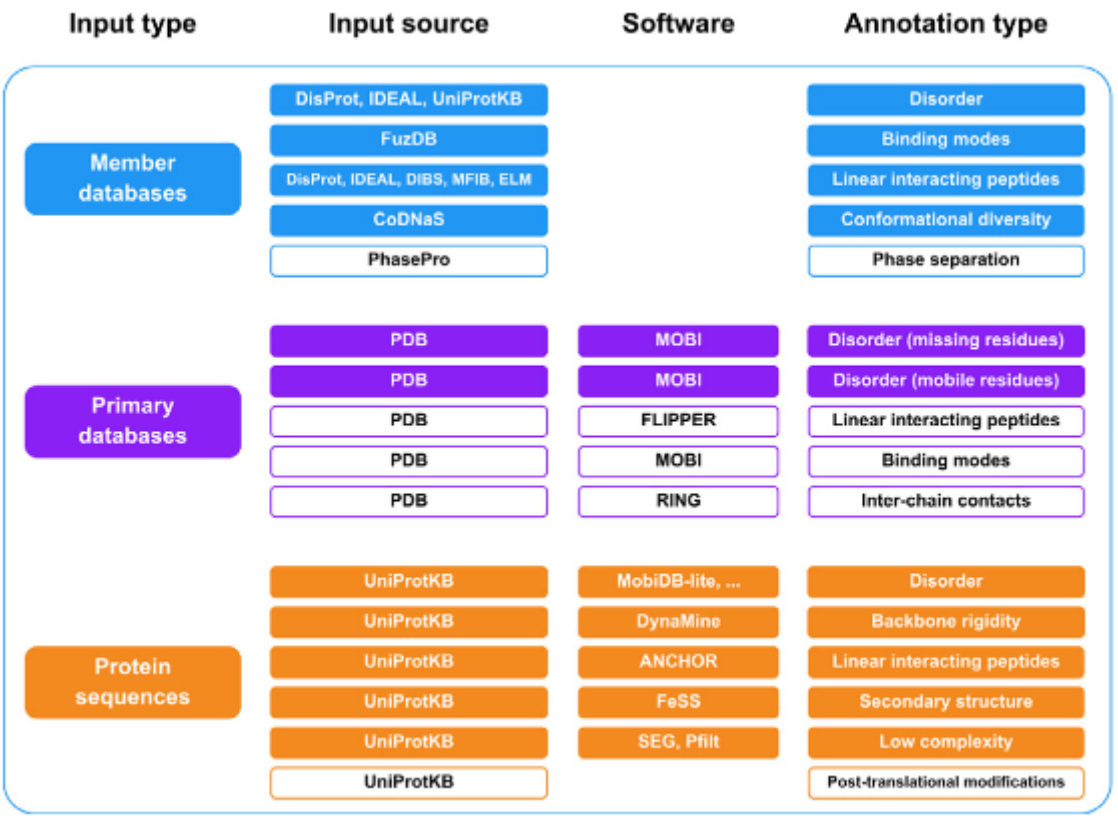
Esquema de trabajo para la anotación de estructuras / secuencias en SIFTS. [41]

A pesar de todo esto, la información de SIFTS no fue de mucha utilidad. La información que proveía era escasa y mal organizada, por lo cual a pesar de haber aportado mucho tiempo para poder dilucidar qué podíamos extraer de esta base de datos, los resultados fueron peores de lo esperado. En lugar de proveer datos particulares para cada proteína con cada una de sus estructuras informada particularmente, SIFTS provee información para estructuras (sin informar a qué proteína le pertenece) y sin información sobre la ocurrencia de los denominados *missing residues* (los aminoácidos ausentes en la cristalización). A pesar de esto, era posible recuperar cierta información, pero el tiempo necesario para reunir todos los datos necesarios era mucho mayor comparado con optar por analizar otras bases de datos más recientes que cumplen el mismo propósito.

Por estas razones, decidimos utilizar otra base de datos que nos aporta información estructural (y secuencial): MobiDB (Piovesan et al. 2021, Di Domenico et al. 2012).
MobiDB, a pesar de ser una base de datos orientada hacia desorden y flexibilidad, contiene toda la información que en principio necesitábamos: agrupa información tanto para el proteoma de referencia (21k proteínas canónicas humanas) como para el proteoma extendido (77k proteínas no revisadas). La diferencia entre estos dos grupos es que el proteoma entero, con las 77 mil secuencias, es el set entero de proteínas que se considera que es expresado por un organismo. La mayoría de las secuencias en proteomas de UniProt provienen de estudios de traducción de proteomas enteros secuenciados, y pueden incluir secuencias que provienen de elementos extracromosomales, ya sea de plásmidos o genomas de organelas. También pueden provenir de secuencias basadas en bibliotecas de

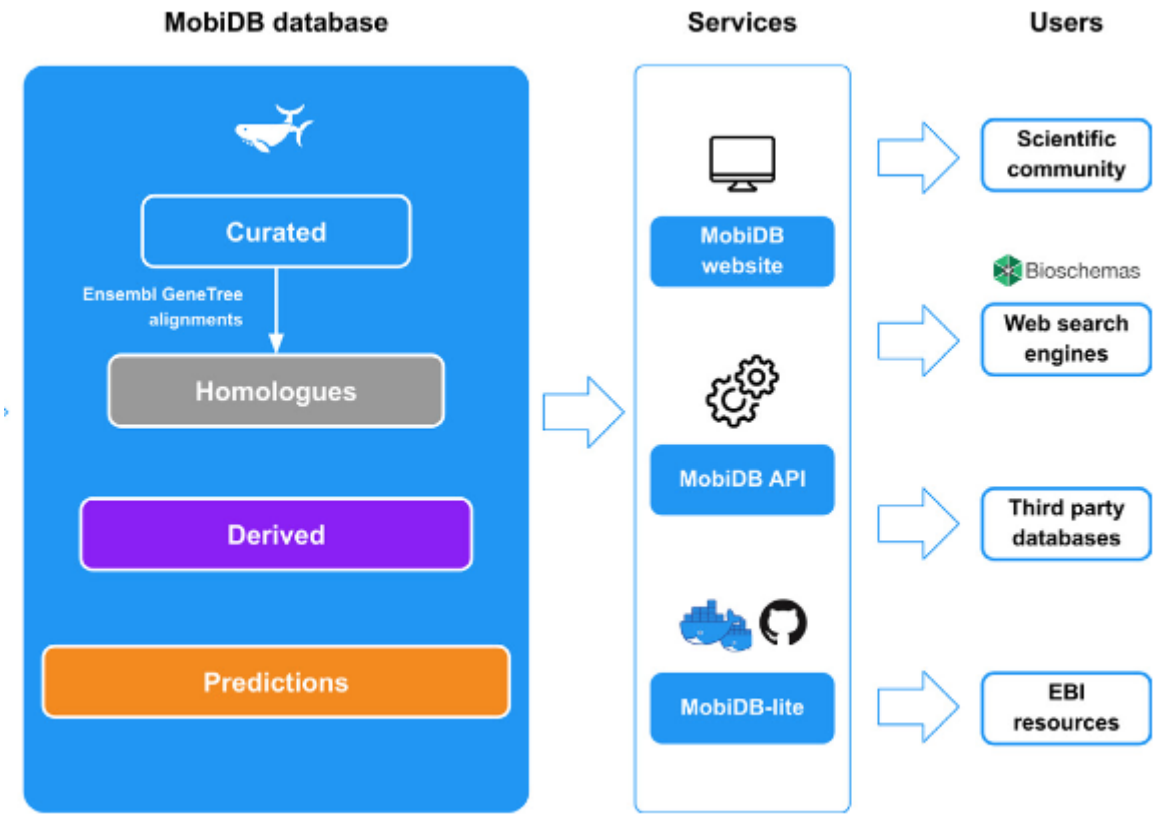
cDNA que no pueden ser mapeadas a los genomas actuales. En cambio, los proteomas de referencia son secuencias seleccionadas específicamente (manual o algorítmicamente) entre todas las del proteoma para poder constituir una representación de la diversidad presente en el mismo.

En el siguiente gráfico podemos encontrar los pasos por los cuales se generan los datasets de MobiDB como el que estudiamos:



Pipeline utilizada por MobiDB. [42.1]

Como primer paso, se recolectan las secuencias input de distintas bases de datos (CoDNaS, PDB, UniProtKB, etc) y se analizan estas proteínas inputs con distintos softwares para tener una caracterización primaria utilizada para anotar el las características de las proteínas.

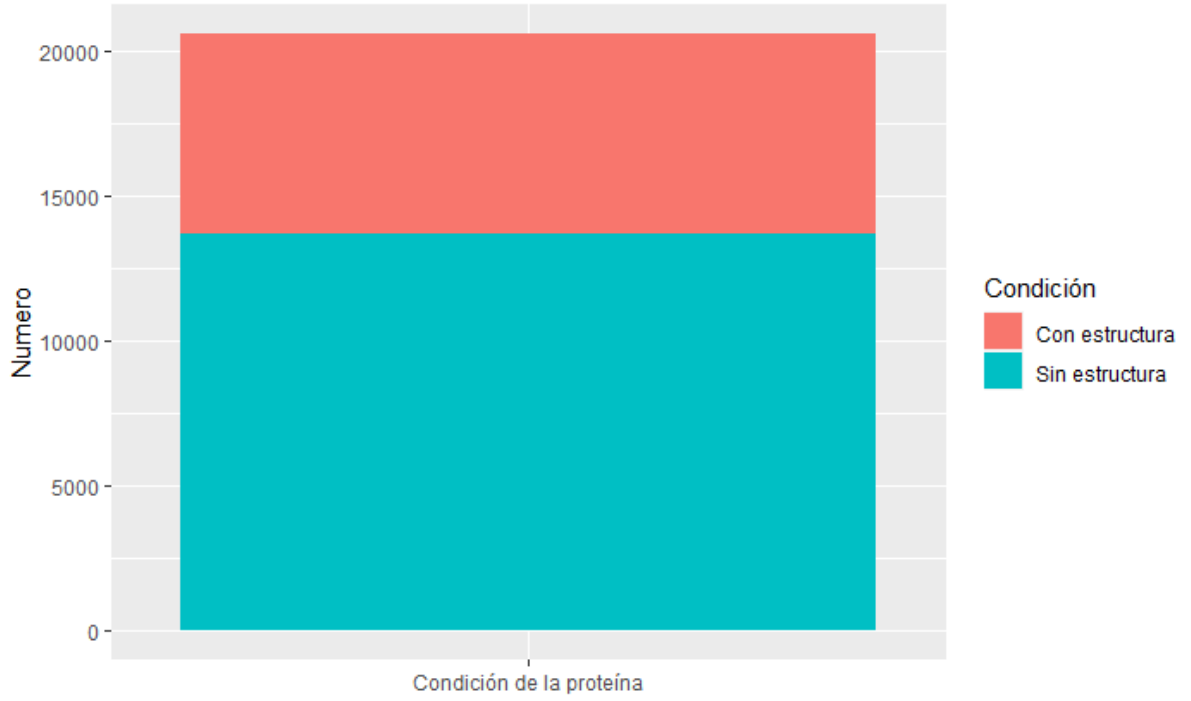


Pipeline utilizada por MobiDB. [42.2]

Luego, estas secuencias son divididas según el tipo de curación que reciben: en base a predictores, derivadas de curaciones previas o por homología. Por último, esta información es alojada en los distintos servidores de MobiDB, para poder ser accedida por distintas fuentes.

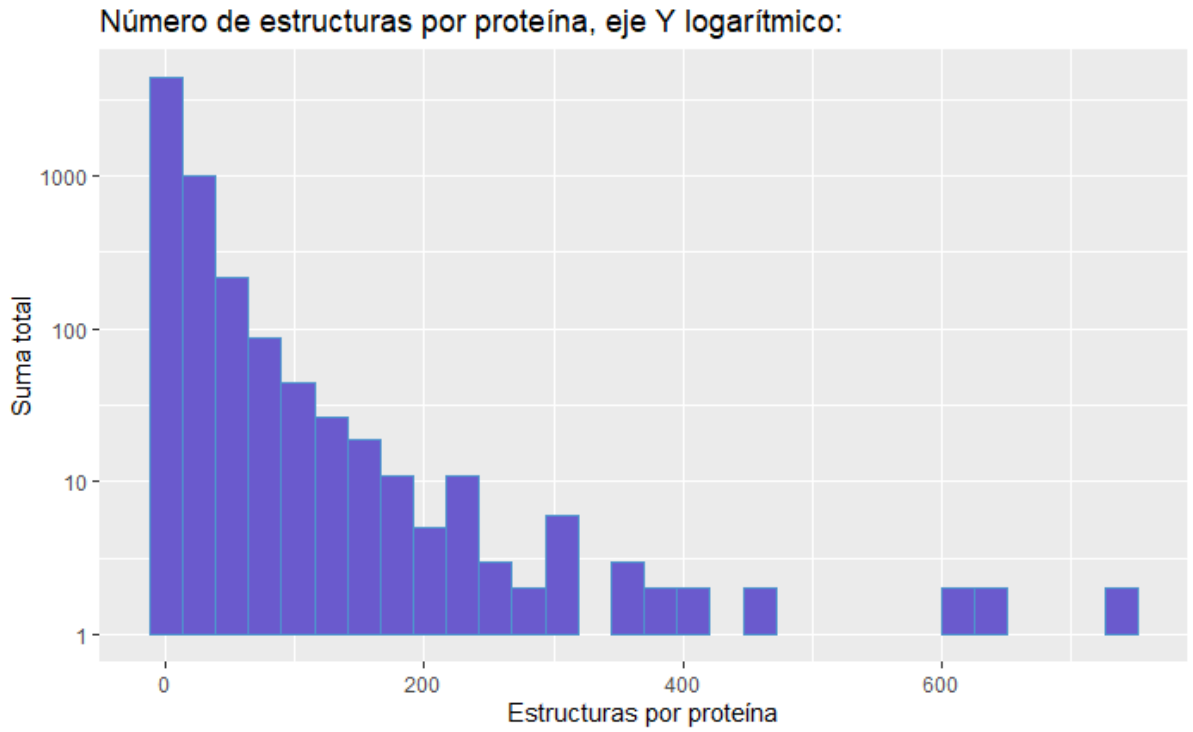
De esta forma, MobiDB contiene información sobre predictores de desorden (y una medida de consenso), scores del DynaMine (que fueron explicados en el capítulo 2, Cilia et al. 2013), predicción de segmentos transmembrana y de baja complejidad, además de información de estructuras de proteínas propias a las secuencias (es decir, solo estructuras provenientes de *Homo sapiens*, no asignadas por homología) con los segmentos mapeados de cada una, los missing residues, y un consenso del fragmento mapeado total de cada proteína. Esto último es de especial importancia, debido a que al albergar tantas estructuras repetidas, muchas veces se genera mucha información redundante. Al tener un consenso, es posible obtener un "mapeo total", proveniente de la suma de todas las estructuras.

En el siguiente gráfico podemos observar el total de proteínas del proteoma humano que tienen estructuras determinadas experimentalmente por cristalografía de rayos X o por NMR (para ser exactos, tienen al menos una estructura). En total, hay 6911 proteínas con al menos una estructura cubriendo la totalidad o algún fragmento de la misma. Esto representa el 33.5% del proteoma, quedando sin caracterización estructural alguna una 13692 proteínas.



Número de proteínas con / sin estructura/s asignadas. [43]

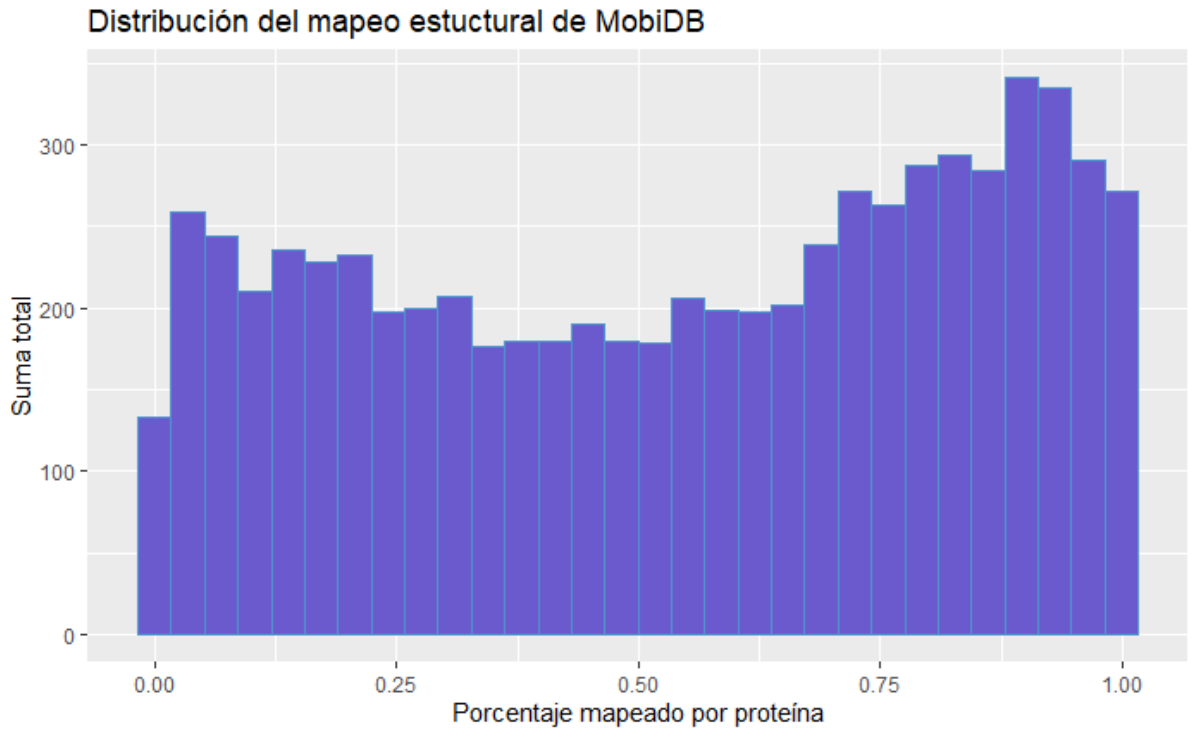
Mencionamos anteriormente que la comparación de distintas estructuras (obtenidas en distintas condiciones) pueden ser utilizadas para estimar la dinámica, diversidad conformacional o flexibilidad de una proteína (Monzon et al. 2013, Monzon et al. 2019, Monzon et al. 2016). De esta forma, también estamos interesados en evaluar la redundancia de estructuras por secuencia. La redundancia de la PDB es bien conocida y una propiedad que atenta contra la caracterización del espacio estructural de las proteínas. La existencia de redundancia produce que distintas estructuras puede que mapeen en las mismas regiones o distintas. Es un análisis preliminar, pero podemos esperar que estructuras más fáciles de cristalizar y de mayor interés científico (asociadas a patologías, de determinadas especies, etc) tengan un mayor nivel de redundancia que estructuras desordenadas, con determinadas particularidades secuenciales que las haga difícil de cristalizar, etc (Marino-Buslje et al. 2019). Estudiando esta distribución, encontramos el siguiente resultado:



Número de estructuras por proteína. [44]

Como podemos observar en el gráfico [44], la gran mayoría de las proteínas tienen asignadas no más de 2 o 3 estructuras, pero existen varias excepciones que introducen un sesgo en esta distribución.

Como existe una elevada redundancia de regiones de mapeo por proteína, MobiDB provee un consenso de mapeo general de proteína: teniendo en cuenta todas las estructuras ,que puede que tengan solapamiento en algunos aminoácidos y no en otros, genera un consenso numérico en base a un cierto número de las estructuras totales de una proteína: si el 90% las estructuras que mapean la región coinciden en que una posición “X” está presente en sus estructuras, esta entra dentro del consenso, que lo expresa como una fracción con respecto a la longitud de la proteína. Al estudiar la distribución de este parámetro, encontramos lo siguiente:



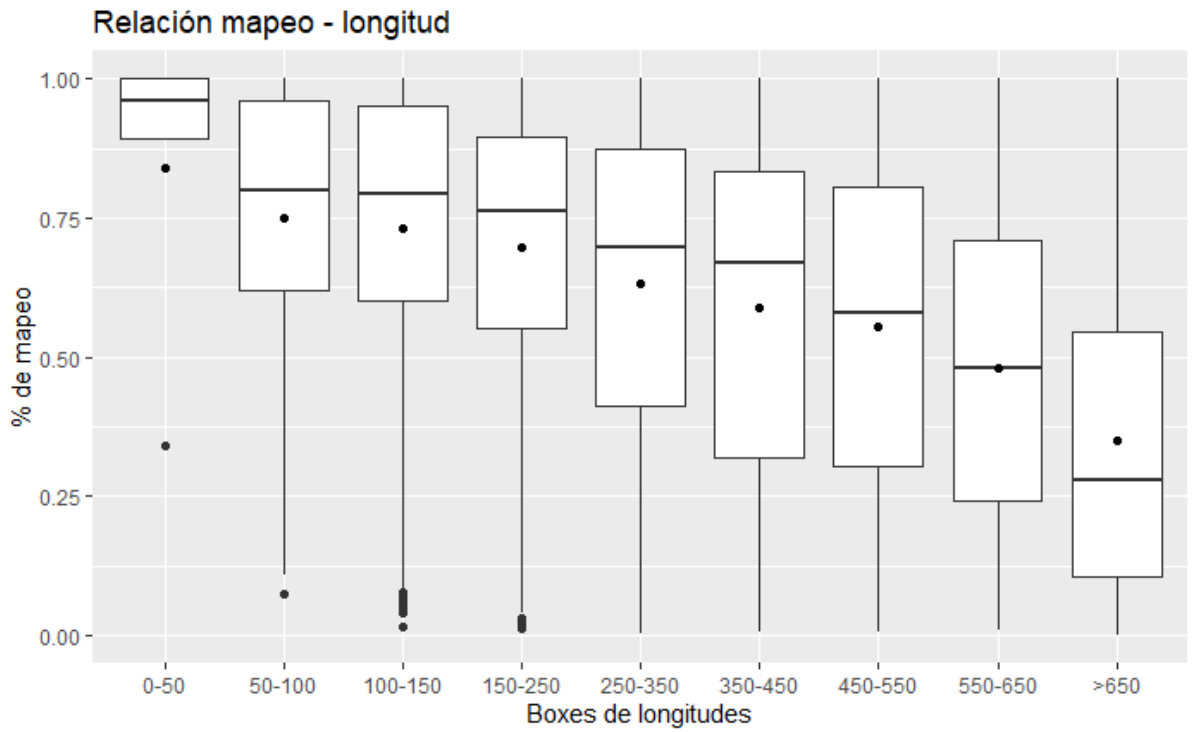
Histograma del porcentaje mapeado por proteína, [45]

Como se puede observar en la figura [45], la población está bastante homogéneamente distribuida, observándose gran cantidad de proteínas con un buen mapeo estructural y casi en la misma proporción sin ese mapeo estructural. Las proteínas que tienen una fracción de mapeo baja fueron apartadas en el presente trabajo para intentar encontrarles otra estructura en esta zona, lo cual analizaremos más adelante. De este gráfico también podemos extraer que sólo un 5,01% de las proteínas del proteoma humano tienen más del 90% de su secuencia canónica cubierta por una estructura obtenida en forma experimental.

Como conclusión podemos afirmar que, por ahora, gran parte del proteoma (66,5%) permanece sin mapear. Esto no significa que no podamos estudiar a estas proteínas desde un punto de vista estructural, más adelante aplicaremos dos métodos de búsqueda y asignación de estructuras para poder dilucidar y estudiar el mayor volumen posible de estructuras posibles. Además, existen ciertas caracterizaciones para las cuales no necesitamos más información que la secuencia aminoacídica (como el desorden y la flexibilidad), pero sumamos muchísima robustez a nuestro análisis si este se ve complementado con el mapeo estructural.

3.2.1 Mapeo estructural y longitud:

Al estudiar la relación del mapeo con la longitud, con el objetivo de analizar que tan bien cubiertas están las proteínas por las estructuras asignadas, encontramos es siguiente resultado:



Boxplots de longitud y su correlación con el % de mapeo. [46]

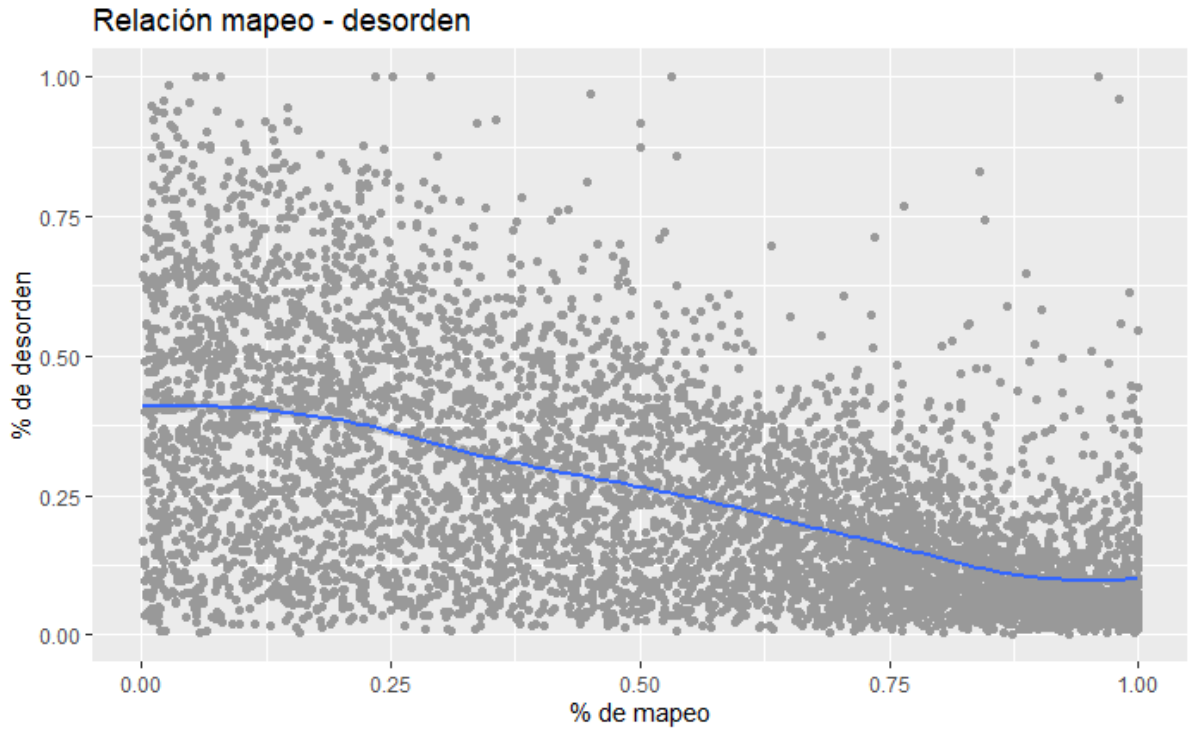
Las proteínas más cortas tienen más tendencia a tener un porcentaje de mapeo más alto que las proteínas de mayores longitudes. Esto se puede deber a que es más sencillo obtener la estructura completa de una proteína siempre que ésta esté compuesta por un solo dominio, que pueda ser cristalizado. Este razonamiento coincide con lo observado en el gráfico [47].

3.2.2 Proteoma humano y desorden:

MobiDB, además de otorgar información estructural de proteínas humanas, ofrece información sobre la predicción de desorden de estas: para cada secuencia otorga los resultados de distintos predictores de desorden (MobiDB-lite, ESpritz-DisProt, ESpritz-NMR, ESpritz-Xray, IUPred-Long, IUPred-Sort, VSL2b, DisEMBL-465, DisEMBL-HotLoops, GlobPlot, JRONN), indicando qué regiones de la proteína son predichas individualmente como desordenadas. Además de esto, provee información sobre el consenso de desorden para el 50% de los predictores aplicados a una secuencia, es decir, si una secuencia tiene

desorden predicho por 10 de los 11 predictores, y 5 de estos predicen que una región entre el aminoácido 25 y el número 100 es desordenada, se informa este resultado en el archivo.

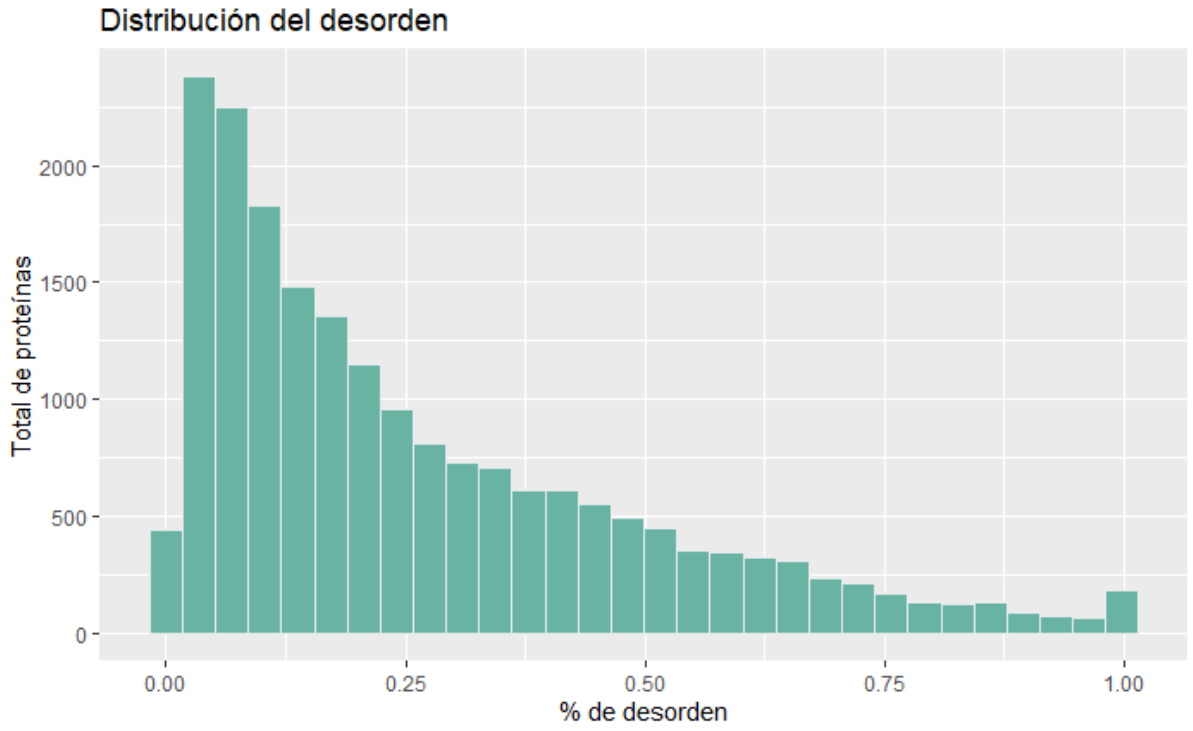
Al obtener esta información y graficarla contra la fracción mapeada de las proteínas, encontramos lo siguiente:



Distribución de los porcentajes de desorden y de mapeo consensos. [47]

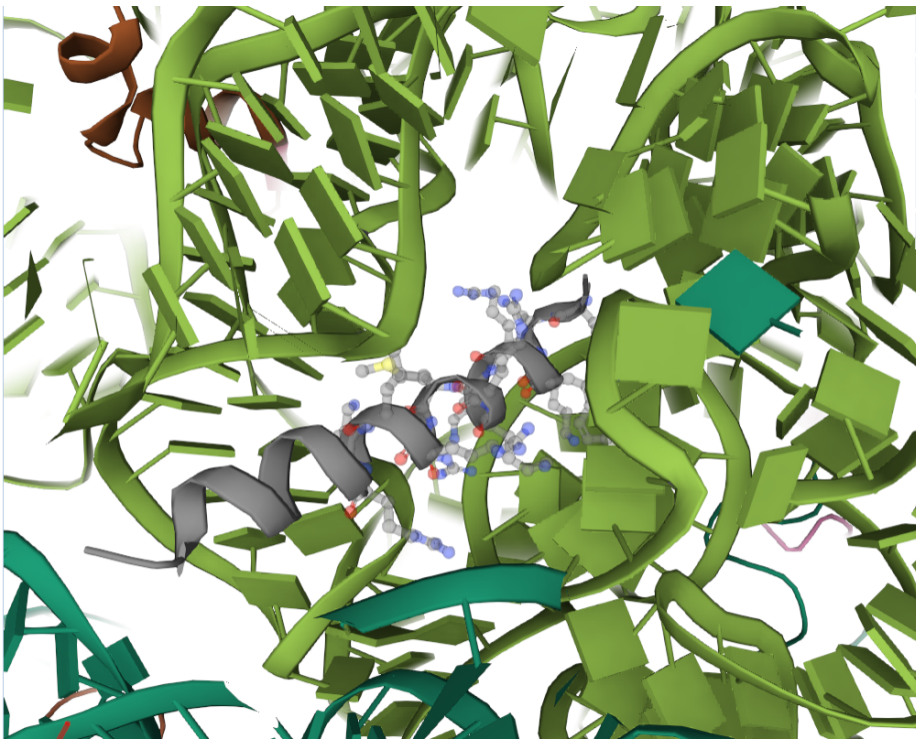
De este gráfico [48], podemos entender la relación entre estructura y desorden: mientras más desordenadas sean las proteínas, menos tendencia a tener una estructura tienen. Esto es un resultado algo esperable, vimos comportamientos similares cuando hablamos de flexibilidad y desorden (específicamente proveniente de MobiDB-lite) en el capítulo 2, solo que en esta instancia lo estamos confirmando con un consenso de predictores y con estructuras, en lugar de plantearlo con flexibilidad y desorden.
Con este consenso estudiamos entonces la distribución de proteínas desordenadas en el proteoma humano.

En este caso, 19546 secuencias tienen al menos una región desordenada (definida como un segmento que componga al menos un 10% de la longitud de la proteína), y solo 1057 quedan fuera de esta lista (94,86% con desorden predicho con consenso vs 5,13% sin ningún tipo de desorden predicho).
De forma análoga, si observamos cómo es la distribución del porcentaje de desorden por secuencia, encontramos:



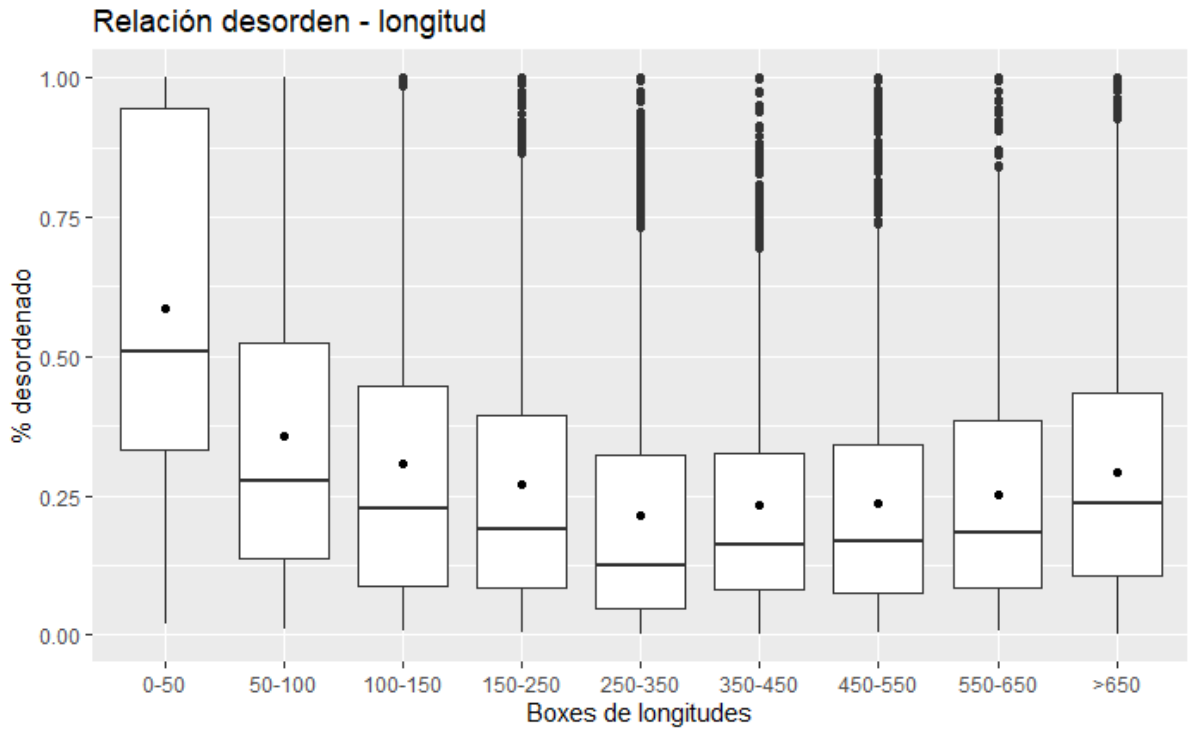
Histograma del % desordenado por número de proteínas, [48]

Que, a diferencia del planteado en el capítulo 2, presenta poblaciones más grandes de proteínas con niveles de desorden más altos, generando una curva mucho más suave.
De forma análoga al análisis anterior, es interesante estudiar las secuencias con 100%de desorden en su longitud, ya que estás de alguna forma rompen con la tendencia observada y demuestran ser un subgrupo de nuestro dataset que es importante de estudiar desde un punto de vista biológico: es un grupo de 164 proteínas que rompen con la tendencia observada, con una longitud promedio de 122 aminoácidos. Aún más interesante, pero lógico debido a la dificultad de cristalizar proteínas desordenadas (Keen and Goodwin 2015), es que de estas 164 proteínas, solo 8 cuentan con alguna estructura, y la fracción de mapeo promedio es de 30,8%. Es remarcable el caso de P62945 (60S ribosomal protein L41), una proteína corta, de 25 aminoácidos, con un consenso de desorden del 100%, pero con una estructura asignada que mapea el 95% de su longitud. Hablaremos de la relación mapeo/desorden un poco más adelante.



Estructura de P62945 (PDB: 4UG0) dentro del Ribosoma 80S humano. [49]
Fuente: PDB

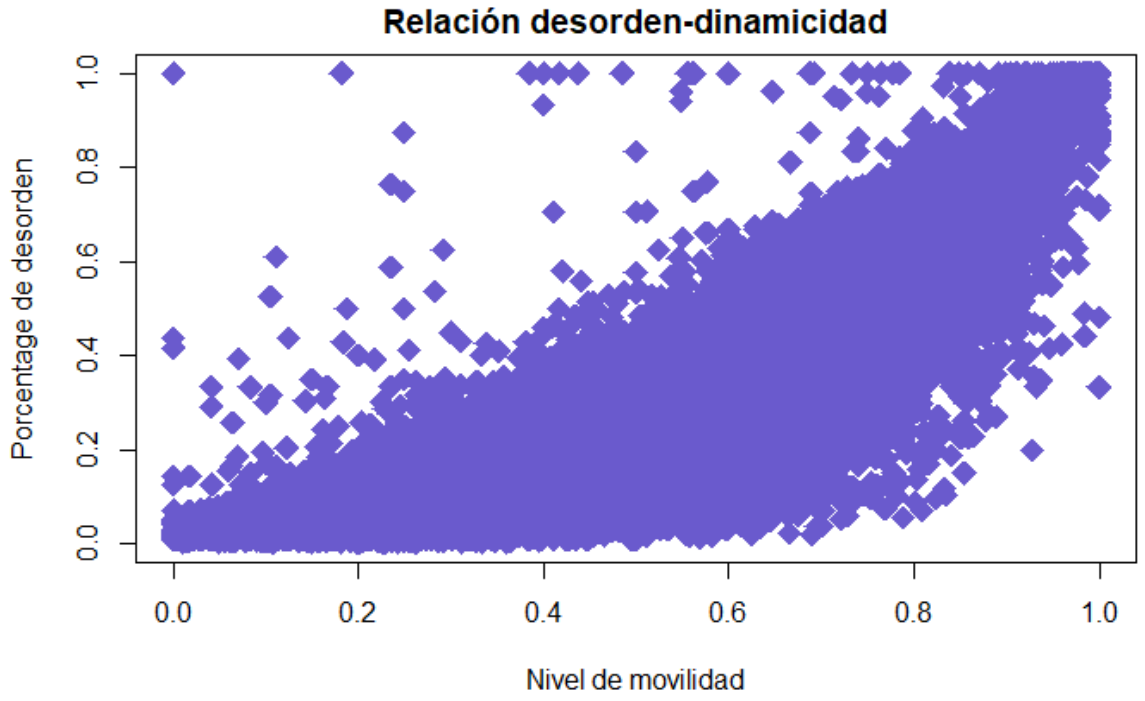
A manera de repasar los resultados del capítulo 2, podemos observar cómo cambian los histogramas de longitud, desorden y scores de DynaMine con desorden al tratarlos con los resultados consensuados, en lugar de usar solo la información de MobiDB-lite. De esta forma, al contar con muchas más secuencias, veríamos de forma más enriquecida ciertas tendencias al estudiar la longitud con respecto al porcentaje de desorden:



Porcentaje de desorden y su relación con la longitud. [50]

Este resultado, a pesar de haber sido obtenido con el proteoma completo, es similar al encontrado anteriormente en la sección 2.6 (el desorden parece estar favorecido en proteínas cortas).
De todas formas, el resultado sigue la misma tendencia que el gráfico previamente explicado [50], pero, en líneas generales, al tener un set de datos mucho más amplio, existe más dispersión y ruido en los datos. Es posible reproducir el mismo esquema para cada uno de los predictores individualmente, pero sería redundante para con el análisis.

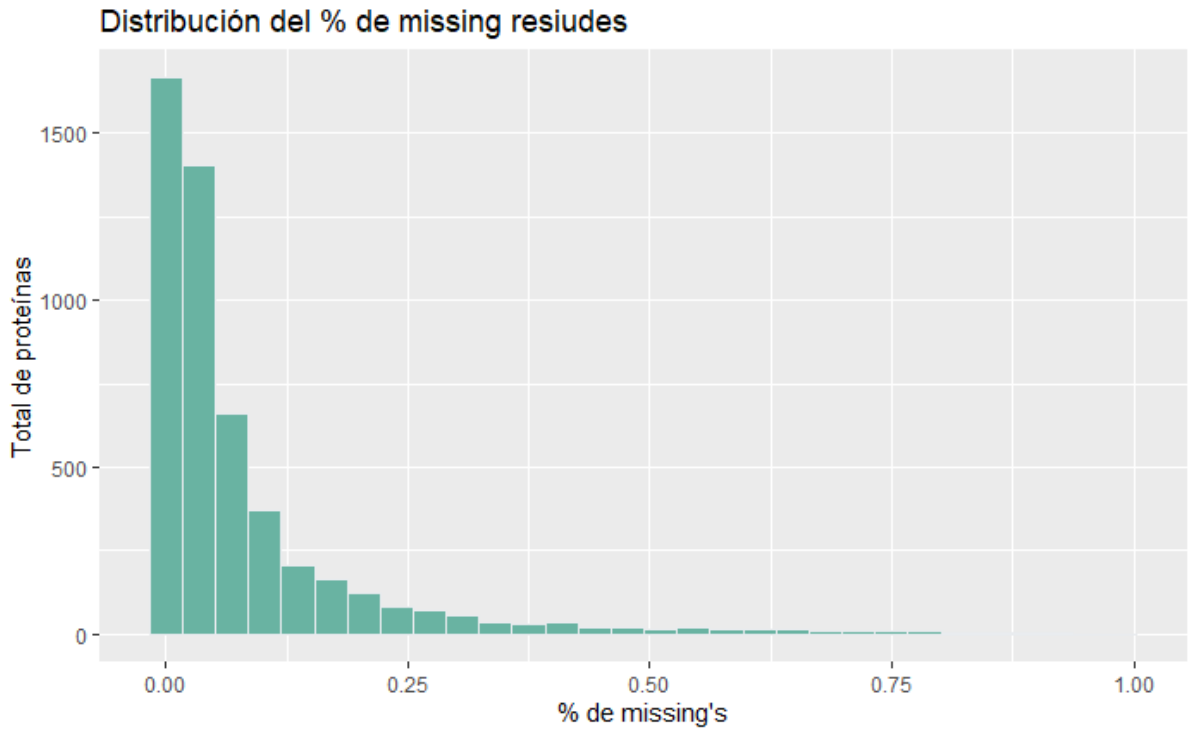
Por último, podemos ver que la tendencia entre el score de DynaMine y el porcentaje consenso de desorden también sigue cumpliendo la misma tendencia, pero encontramos un nivel mucho más alto de outliers dentro de la población. Esto puede deberse a que las proteínas con altos scores de DynaMine tienen predicciones de desorden consenso relativamente pobres, ó que se tratan de falsos positivos:



Distribución de puntos entre los porcentajes de desorden y scores de DynaMine. [51]

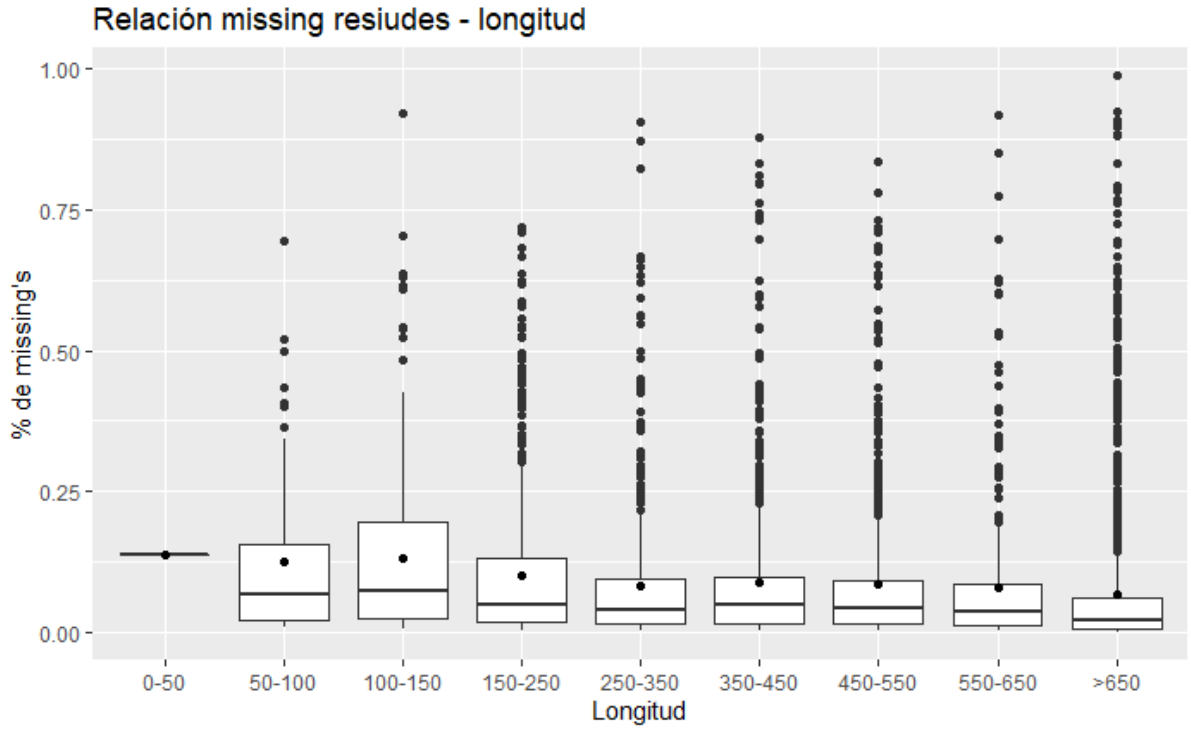
Para cerrar esta parte del capítulo, hablaremos brevemente de otro dato que nos otorgan las estructuras humanas almacenadas en MobiDB: los Missing Residues.
Ya habíamos mencionado que los missing residuos son residuos o regiones que por su alta movilidad no pueden ser estimadas sus coordenadas espaciales utilizando difracción de rayos X. También vimos que por representar regiones desordenadas poseen una composición diferencial (Djinovic-Carugo and Carugo 2015).

En nuestro dataset, analizando las proteínas humanas en MobiDB, un total de 5083 proteínas con estructura cuentan con *missings residues* (MR, Djinovic-Carugo and Carugo 2015). Esto representa a un 73,55% de las proteínas con estructuras humanas asignadas.
Estudiando su histograma y su relación con la longitud y el desorden, encontramos lo siguiente:



Histograma de la distribución de los Missing Residues. [52]

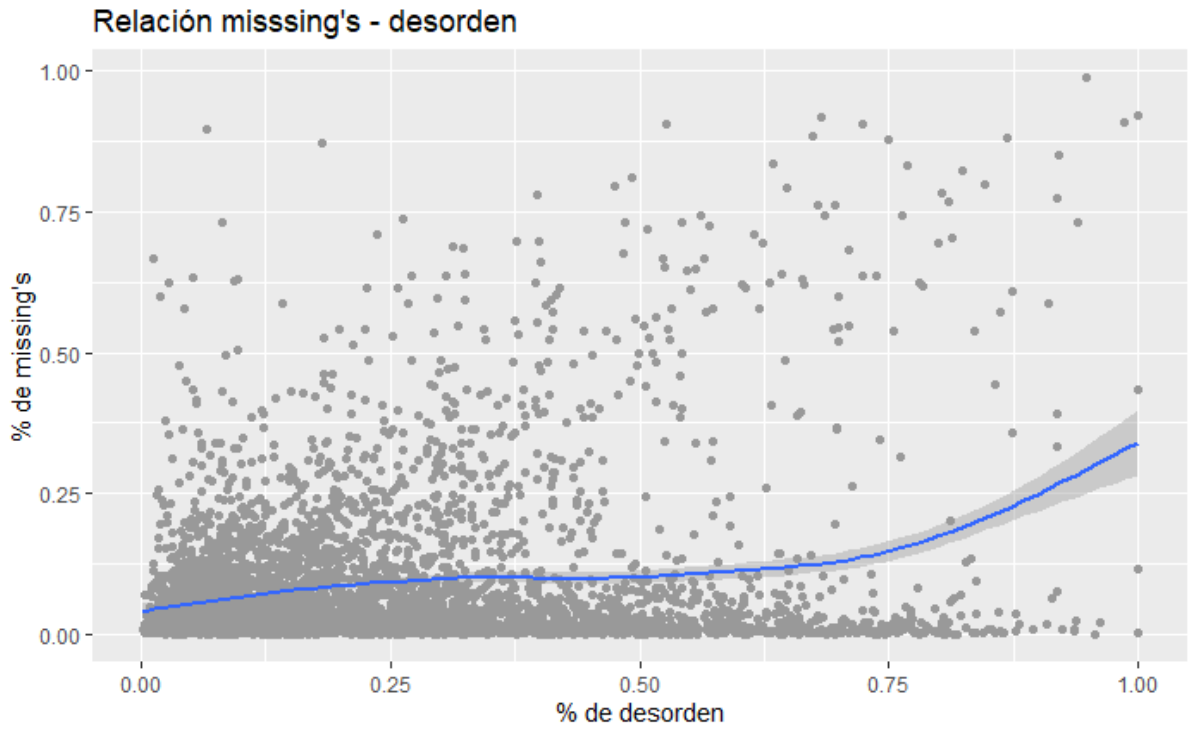
En este histograma [52], podemos observar como la gran mayoría de las proteínas con estructuras no superan un número mayor al 10% de missings residues con respecto a su longitud, sin marcar ninguna clara tendencia.



Boxplots sobre la relación de los MR y la longitud. [53]

Al examinar la relación con la longitud [53], podemos ver que tampoco existe una correlación muy marcada.

Por último, y probablemente de mayor interés, podemos examinar si existe una correlación entre el porcentaje de los aminoácidos missings y el porcentaje de desorden predicho. Es importante notar que en este gráfico no estamos examinando si las posiciones que en la cristalización fueron anotadas como missings también fueron predichas como desordenadas, sino si existe relación entre un alto porcentaje de missings y un alto desorden:



Distribución entre los missing residues y el desorden. [54]

Como podemos observar en el gráfico [54], tampoco encontramos una correlación clara en este caso.

3.3 Inferencia de estructuras basada en homología

En la sección anterior utilizamos la base de datos MobiDB para extraer información sobre el estructuroma humano ya caracterizado por técnicas de cristalografía de RX, NMR y predicciones de desorden. Como quedó demostrado sólo un 5,01% de las proteínas del proteoma humano tienen más del 90% de su secuencia canónica cubierta por una estructura obtenida en forma experimental. Evidentemente el repertorio de estructuras de proteínas humanas depositadas en PDB es limitado. Sin embargo, podemos ampliar la cantidad de estructuras estimadas para el proteoma humano utilizando estructuras estimadas usando métodos predictivos como el comúnmente usado por homología. Este método, enormemente utilizado en biología estructural en los últimos 40 años, consiste en asignar un molde (proteína con estructura conocida) a la secuencia problema a partir del cual se construirá un modelo tridimensional de la misma. Este molde idealmente debería ser lo más cercano posible en términos evolutivos.
En la década del 80, Chothia y Lesk establecieron las bases del modelado por homología determinando que las proteínas secuencialmente similares lo son también desde el punto de vista estructural (Chothia and Lesk 1986). Asimismo establecieron que las estructuras proteicas se conservan más en la evolución que las secuencias que las codifican. Esto último permite buscar moldes para proteínas muy alejadas evolutivamente, aunque con la consecuente disminución en la calidad del modelo obtenido. Utilizar el modelado por homología nos va a permitir extender la caracterización del estructuroma humano cubierta actualmente por las estructuras depositadas en la PDB, enriqueciendo de esta forma nuestro conocimiento de la diversidad de estructuras y relaciones estructura-función en el proteoma humano. De esta forma para caracterizar el resto del estructuroma necesitamos realizar inferencias de modelos estructurales utilizando información secuencial. Entre los métodos de elección para tal propósito encontramos la asignación de estructura por homología es un método ampliamente utilizado (Henikoff and Henikoff 1992).

3.3.1 Fundamentos de los métodos de inferencia de modelos 3D por homología:

La base de datos bajo la cual se realiza la búsqueda en nuestro caso es la PDB, debido a que de esta forma los matches que encontramos son estructuras ya conocidas, y usando BLAST P, qué es el algoritmo de BLAST optimizado para el alineamiento y búsqueda de matches proteína - proteína (o en nuestro caso, proteína-estructura). Los resultados de esta asignación dependen exclusivamente de qué tan idénticas son las secuencias de la proteína y la secuencia de la estructura homóloga (Chothia and Lesk 1986).

```
# /bin/sh
```

```
for i in `cat lista`
  do
   blastp -db pdb_seqres.txt -query $i -evalue 1e-5 -out $i.out
  done
```

Script básico en Bash para ejecutar el BLASTP. [55]

El algoritmo sirve para, al dar una secuencia query a la cual queremos encontrarle una secuencia supuestamente homóloga, realizar alineamientos (lo cual enfrenta la secuencia contra una base de datos) mediante una matriz de sustitución (como la matriz PAM-250 de Dayhoff, 1978) [56], (Altschul et al. 1990, Altschul et al. 1997).

Table I. The 250 PAM PET91 matrix ($\log_{10}$ relatedness odds), based on 59 190 accepted point mutations found in 16 130 protein sequences

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2	247	216	386	106	208	600	1183	46	173	257	200	100	51	901	2413	2440	11	41	1766
R	-1	5	116	48	125	750	119	614	446	76	205	2348	61	16	217	413	230	109	46	69
N	0	0	3	1433	32	159	180	291	466	130	63	758	39	15	31	1738	693	2	114	55
D	0	-1	2	5	13	130	2914	577	144	37	34	102	27	8	39	244	151	5	89	127
C	-1	-1	-1	-3	11	9	8	98	40	19	36	7	23	66	15	353	66	38	164	99
Q	-1	2	0	1	-3	5	1027	84	635	20	314	858	52	9	365	182	149	12	40	58
E	-1	0	1	4	-4	2	5	610	41	43	65	754	30	13	71	156	142	12	15	226
G	1	0	0	1	-1	-1	0	5	41	25	56	142	27	18	93	1131	164	69	15	276
H	-2	2	1	0	0	2	0	-2	6	26	134	85	21	50	157	138	76	5	514	22
I	0	-3	-2	-3	-2	-3	-3	-3	-3	4	1324	75	704	196	31	172	930	12	61	3938
L	-1	-3	-3	-4	-3	-2	-4	-4	-2	2	5	94	974	1093	578	436	172	82	84	1261
K	-1	4	1	0	-3	2	1	-1	1	-3	-3	5	103	7	77	228	398	9	20	58
M	-1	-2	-2	-3	-2	-2	-3	-3	-2	3	3	-2	6	49	23	54	343	8	17	559
F	-3	-4	-3	-5	0	-4	-5	-5	0	0	2	-5	0	8	36	309	39	37	850	189
P	1	-1	-1	-2	-2	0	-2	-1	0	-2	0	-2	-2	-3	6	1138	412	6	22	84
S	1	-1	1	0	1	-1	-1	1	-1	-1	-2	-1	-1	-2	1	2	2258	36	164	219
T	2	-1	1	-1	-1	-1	-1	-1	-1	1	-1	-1	0	-2	1	1	2	8	45	526
W	-4	0	-5	-5	1	-3	-5	-2	-3	-4	-2	-3	-3	-1	-4	-3	-4	15	41	27
Y	-3	-2	-1	-2	2	-2	-4	-4	4	-2	-1	-3	-2	5	-3	-1	-3	0	9	42
V	1	-3	-2	-2	-2	-3	-2	-2	-3	4	2	-3	2	0	-1	-1	0	-3	-3	4

Matriz de sustitución PAM 250. [56]

3.3.2 Aplicación de inferencia por homología al proteoma humano:

A la hora de realizar la búsqueda para nuestras secuencias, dividimos el dataset en tres:

Partiendo de todas las proteínas sin estructura evidenciada en PDB y proveniente de MobiDB, separamos aquellas que tenían al menos un segmento transmembrana de las que no tenían. A las proteínas con segmentos transmembrana (detectados con información proveniente de MobiDB, que a su vez, fueron recopilados de UniProt) se eliminamos el/los segmentos transmembrana y utilizamos cada uno de los fragmentos restantes si tenían una longitud mayor a 100 aminoácidos. Elegimos este umbral ya que dicha cantidad de aminoácidos representa un dominio pequeño aumentando las chances de mejorar la predicción estructural.

En el proteoma, existen un total de 6.251 proteínas con al menos uno de estos segmentos, pero después de aplicar los filtros (proteínas que no tengan todavía estructuras asignadas y que las secuencias restantes sean de una cierta longitud), quedaron en total de 735 segmentos de secuencias de más de 100 aminoácidos con un solo segmento transmembrana y 466 secciones de proteínas de más de 100 aminoácidos que están flanqueados, también, por regiones transmembrana. Estos dos paquetes fueron sometidos al BLAST, por separado.
Una vez analizado los resultados de la búsqueda de similitud secuencial, se encontraron 497 estructuras para el primer set (proteínas con un solo segmento transmembrana), y 244 estructuras para el segundo (proteínas con más de un segmento transmembrana). Antes de analizar los coverages y E-values, tenemos un total de secuencias con estructuras encontradas por blast (retorno) del 67,16% para el primer set, y un 52,3% para el segundo.

Para el segundo set de datos, aquellas proteínas que no tenían regiones transmembrana y no tenían asignación de estructura, partiendo de un total de 13692 secuencias presentes en el proteoma se encontraron 9149 secuencias posiblemente homólogas con estructura conocida, un retorno del 66,82%.

Este total de proteínas fue sometido a un análisis adicional: las proteínas con un posible homólogo de estructura conocida fueron alineadas con la misma y las secciones por fuera de este alineamiento fueron cortadas y, después de un filtro aplicado a su longitud (solamente buscamos dominios mayores a 100 aminoácidos), fueron sometidas a nuevas búsquedas de similitud secuencial nuevamente. En total, 3204 secuencias volvieron a ser analizadas, pero solo 1559 recuperaron algún hit.

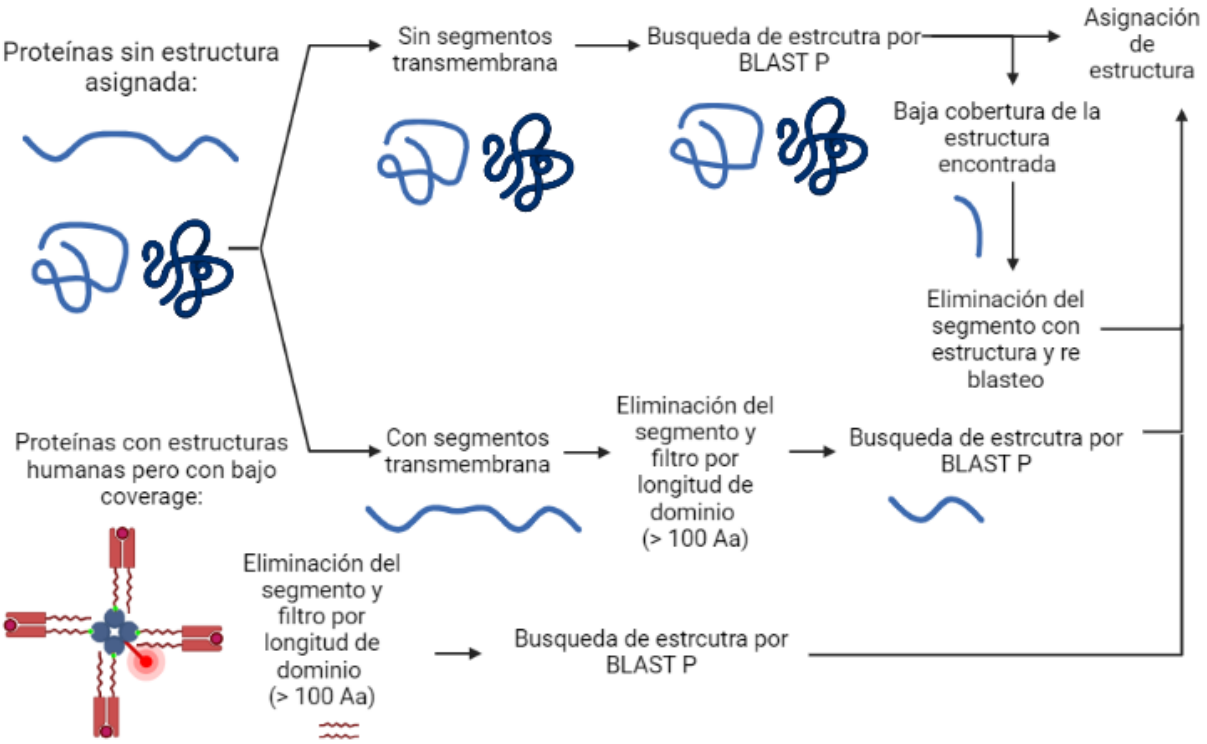
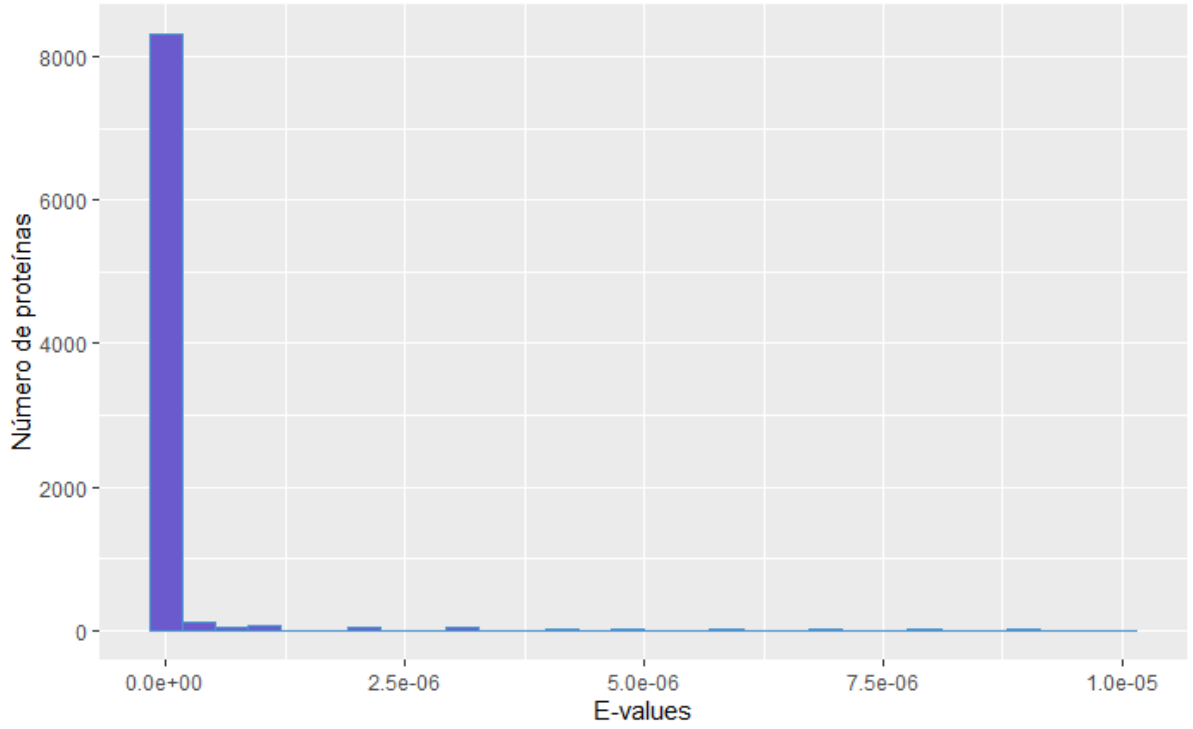


Diagrama de flujo para la asignación de estructura por BLASTP. [57]

3.3.3 Análisis de los resultados:

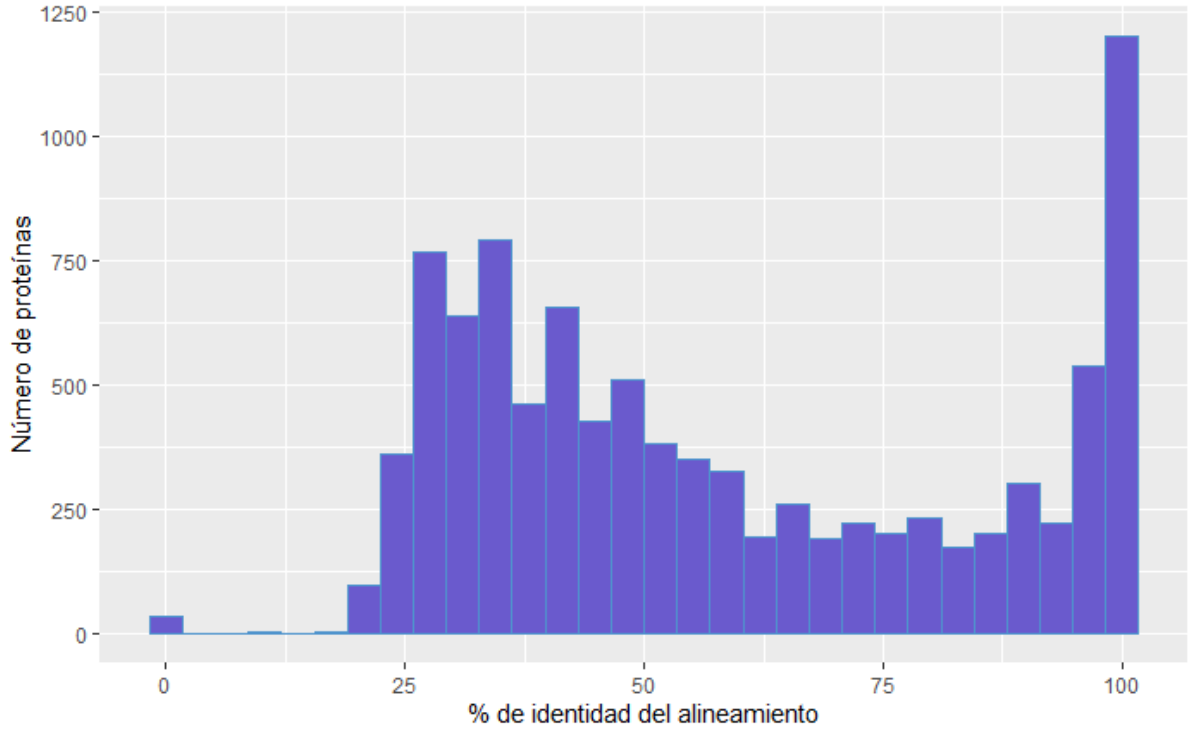
Luego de haber utilizado el BLAST con el paquete de secuencias sin estructura, podemos empezar a analizar los parámetros importantes a la hora de poder afirmar si una estructura se corresponde con una secuencia o no: su identidad, E-value y coverage, e intentar sacar conclusiones de este último con parámetros ya estudiados.



Histograma de la distribución de E-values por proteína. [58]

Comenzamos estudiando los E-values, para poder validar los hits obtenidos [57]. Los E-values más cercanos (o mayores) a 0.01, indicarían una alta probabilidad de haber obtenido un falso positivo. En nuestro caso, podemos observar una muy buena distribución de estos scores. La gran mayoría de las proteínas caen en un valor de 0 o muy cercanas a este (es decir, números con exponentes negativos muy altos), con solo 4 secuencias con valores mayores (es decir, más cercanas a 1) que 1.0e-05.

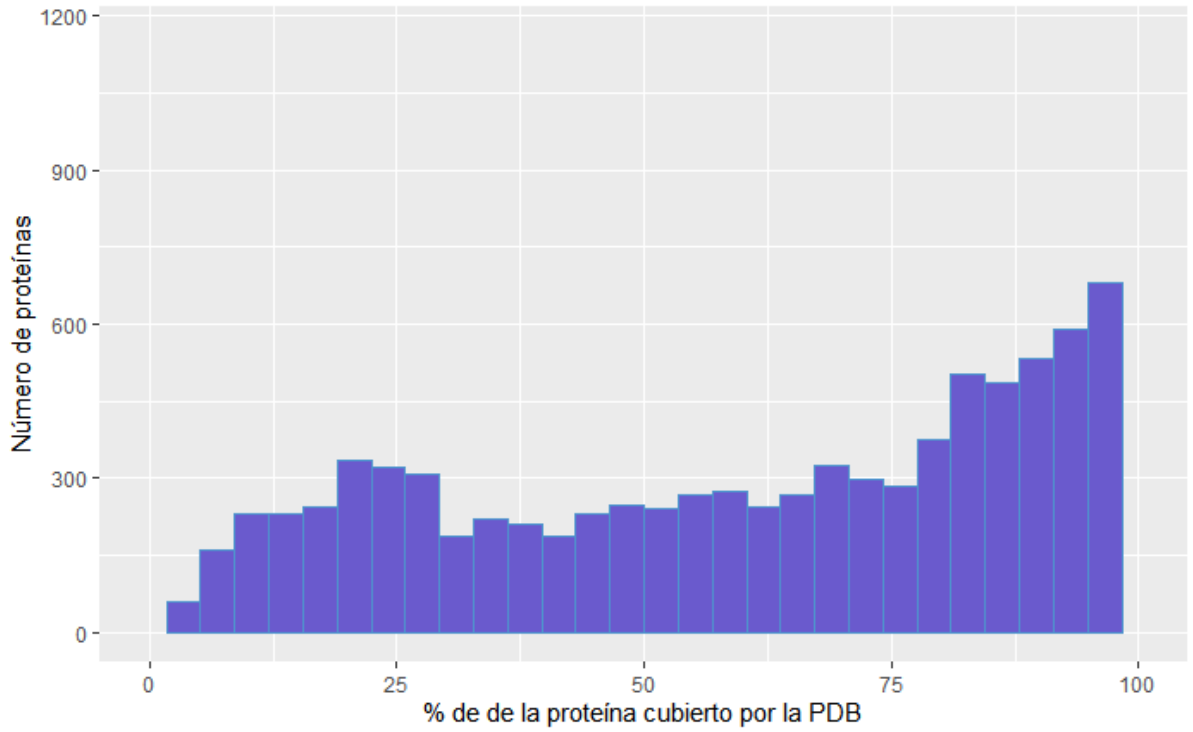
El siguiente parámetro a estudiar es qué tan parecidas son las regiones alineadas de las secuencias del proteoma humano con las estructuras: esto se realiza mediante un parámetro de identidad, que compara aminoácido a aminoácido si las posiciones alineadas son iguales. Es importante, sin embargo, tener en cuenta que este parámetro se calcula de forma local. Es posible que muchas estructuras tengan identidades muy altas, pero que representen una sección muy chica de la proteína en general.



Distribución de identidad del alineamiento para las proteínas. [59]

Al estudiar las identidades de los distintos matches obtenidos por el BLAST del dataset de estructuras obtenidas [58], podemos ver que la gran mayoría de la población obtiene identidades significativas (con un parámetro de corte por arriba del 30%), obteniendo un total de 8476 secuencia con scores superiores a este valor, un 87,04% del total.

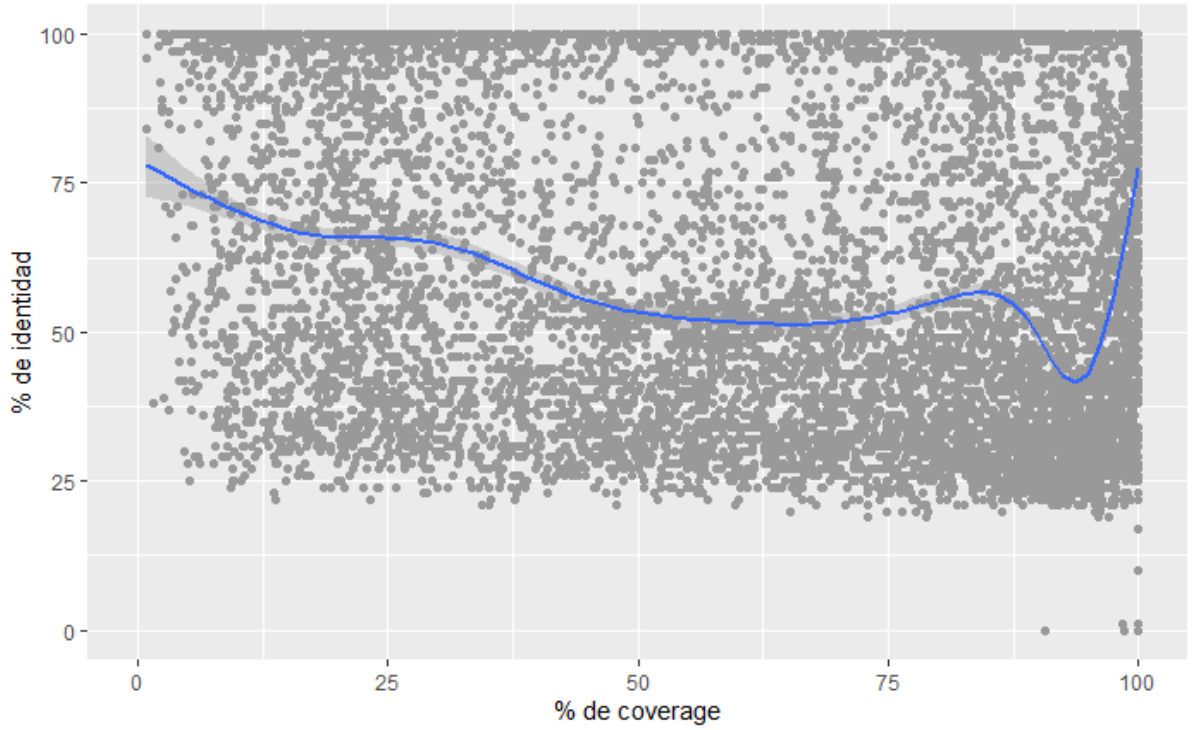
Finalmente, podemos estudiar cómo se comparan las estructuras reclutadas por el BLAST con las proteínas alineadas con respecto a su longitud: esto se hace alineando globalmente toda la estructura de la PDB con la secuencia humana, tomando como referencia la región localmente alineada. En este caso, podemos encontrar el siguiente gráfico:



Histograma del porcentaje cubierto de la secuencia por la PDB. [60]

Como se puede observar en el gráfico [60], hay una tendencia a valores por arriba del 50%. En total, 7799 proteínas tienen coverages mayores que 30%, representando un 80,08% de la población total.
Aplicando estos filtros (E-value menor a 1e-5, y coverages e identidades mayores al 30%) nos quedamos con un total de 6496 proteínas, un 58% del total.

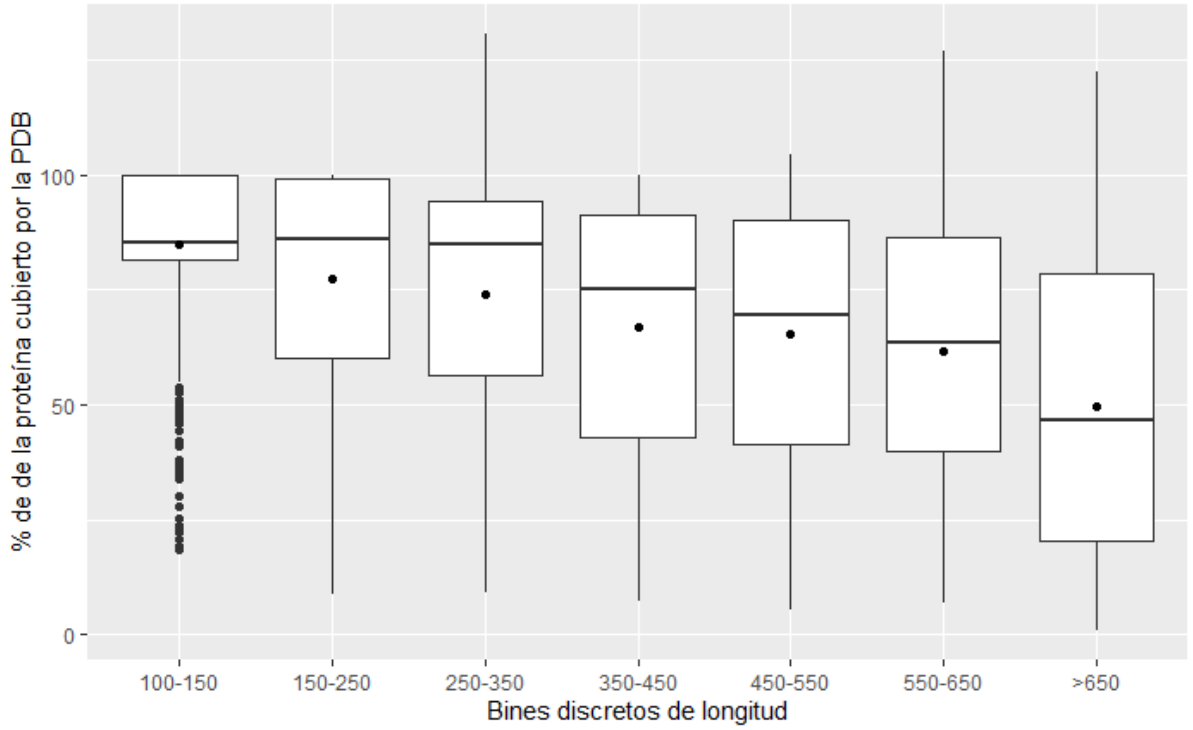
Si estudiamos cómo se relacionan los parámetros de identidad y coverage (de todo el set, no del subset filtrado), encontramos el siguiente resultado:



Relación entre la identidad y el coverage. [61]

Donde no parece existir una correlación constante entre el porcentaje cubierto de la proteína por la estructura y su identidad [60].

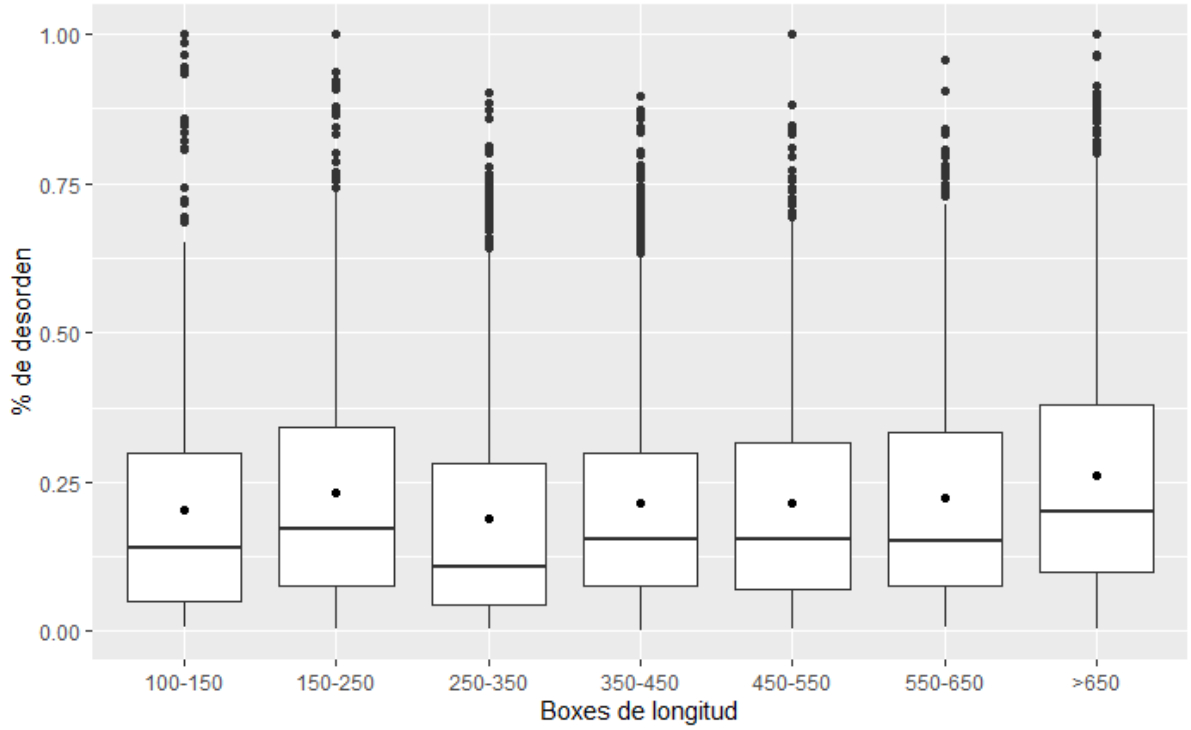
Realizando un análisis similar a los anteriores, podemos ver si existe una tendencia entre el coverage y la longitud de la proteína:



Relación entre boxes de longitud y el porcentaje mapeado. [62]

Donde podemos observar [61] que se sigue cumpliendo la tendencia de proteínas más cortas a tener coverages más altos, debido a que con una sola estructura ya se puede entender la totalidad de la proteína.

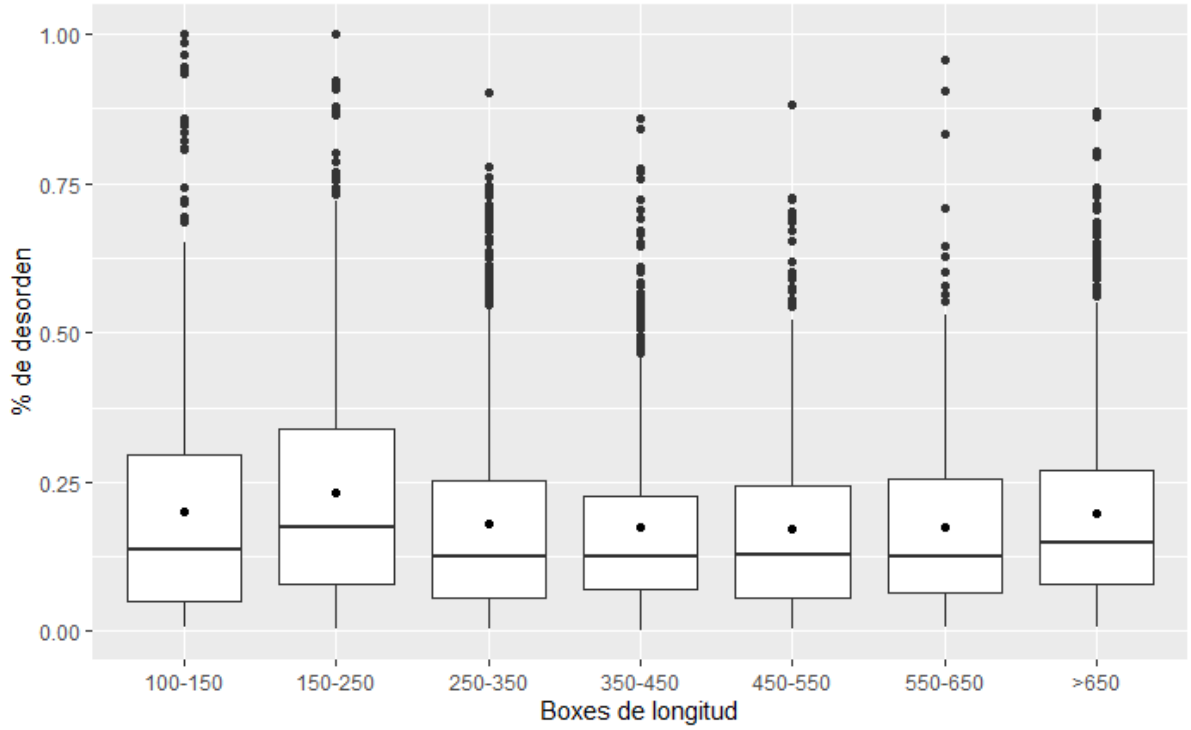
Análogamente, si vemos como es la relación de desorden de estas secuencias con su longitud podemos encontrar lo siguiente:



Relación desorden - longitud. [63]

Que también concuerda con lo esperado, pero con una tendencia menos marcada: estas secuencias de proteínas parecen tener niveles de desorden bastantes bajos.

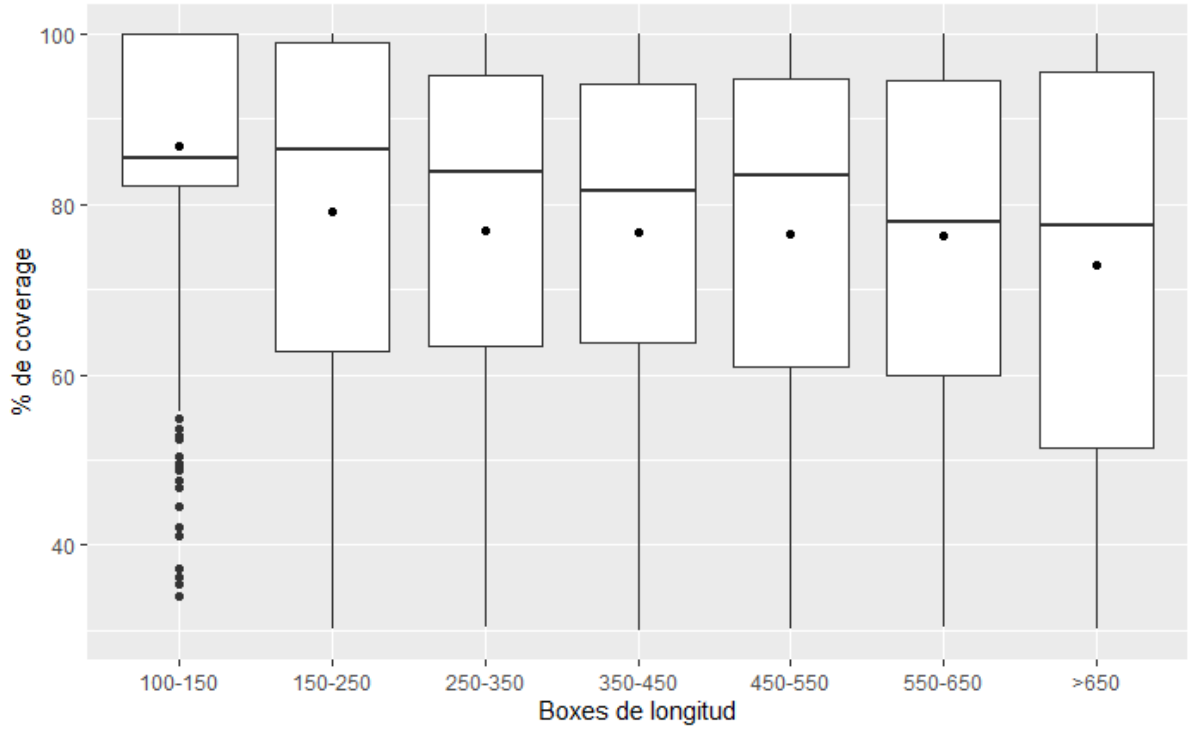
Finalmente, podemos estudiar cómo es la relación de desorden y estructuras pero con nuestro dataset ya filtrado por los parámetros de identidad, coverage e E-value de forma similar a la anterior:



Relación desorden - longitud (proteínas con estructura asignada con BLAST). [64]

El gráfico [64] indica un comportamiento totalmente similar al anterior, con una tendencia a tener valores levemente más bajos

Y, finalmente, podemos estudiar cómo se relaciona su longitud con el porcentaje de la proteína cubierto por la estructura:



Relación coverage - longitud (proteínas con estructura asignada con BLAST) [65]

Donde encontramos el resultado esperado [65]: se sigue cumpliendo la tendencia de que proteínas más cortas tengan un grado de cobertura mayor, pero a diferencia del gráfico anterior, podemos encontrar mucha menos dispersión en los datos, con una tendencia más marcada hacia porcentajes más altos.

Por último, se tomó las secuencias con estructura provenientes de MobiDB que, a pesar de tener estructura, tenían un mapeo muy bajo (menor al 30%) y se las volvió a someter al BLAST, con el propósito de encontrar otra posible estructura que otorgue un mejor mapeo. Del total de secuencias de MobiDB, 4792 secuencias caían dentro de este rango, y luego del BLAST, se encontraron 4555 resultados positivos.

3.4 Búsqueda de estructuras basada en homología: HHblits

Una vez que terminamos de reclutar estructuras vía BLAST, quedaba un conjunto de proteínas que no habían recuperado ninguna estructura por este método. Debido a que era un grupo más pequeño, recurrimos a un motor de búsqueda por homología más poderoso: el HHblits (Remmert et al. 2011). El HHblits es un software de asignación de estructura/función por homología que, en lugar de basarse en alineamientos secuencias-secuencia, utiliza alineamientos de HMM-HMM, por medio de hidden Markov models (HMMs, cadenas markovianas ocultas, por sus siglas en inglés) (Eddy 2004).

3.4.1 Fundamentos del uso de HMM en la asignación estructural por homología:

El funcionamiento básico del algoritmo se basa en generar alineamientos de secuencias a los cuales, por cada residuo, generan una pseudocuenta que contenga información en un contexto de 13 posiciones hacia ambos lados por cada posición en la cadena de aminoácidos. Luego, el HHblits filtra la base de datos de HMM para encontrar posibles candidatos con ciertos parámetros de E-value secuencias con perfiles similares. Una vez aplicado el filtro, el algoritmo aplica un alineamiento, pero en lugar de usar una matriz de 20x20 para los distintos aminoácidos, se genera un alfabeto de 219 caracteres que representan una columna típica del perfil [65].

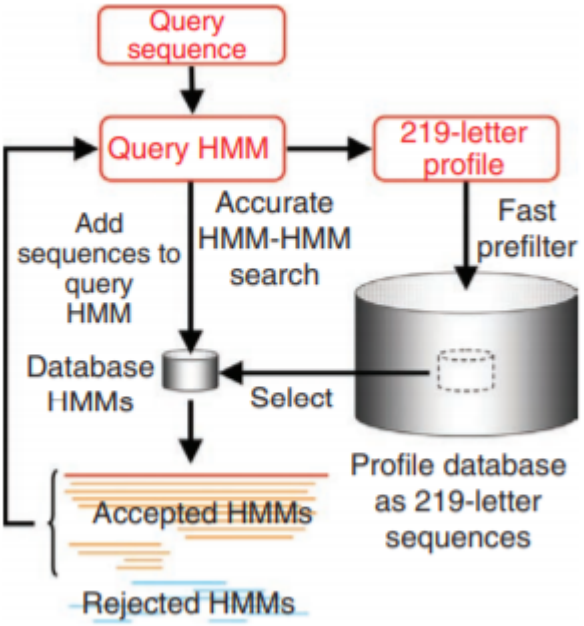


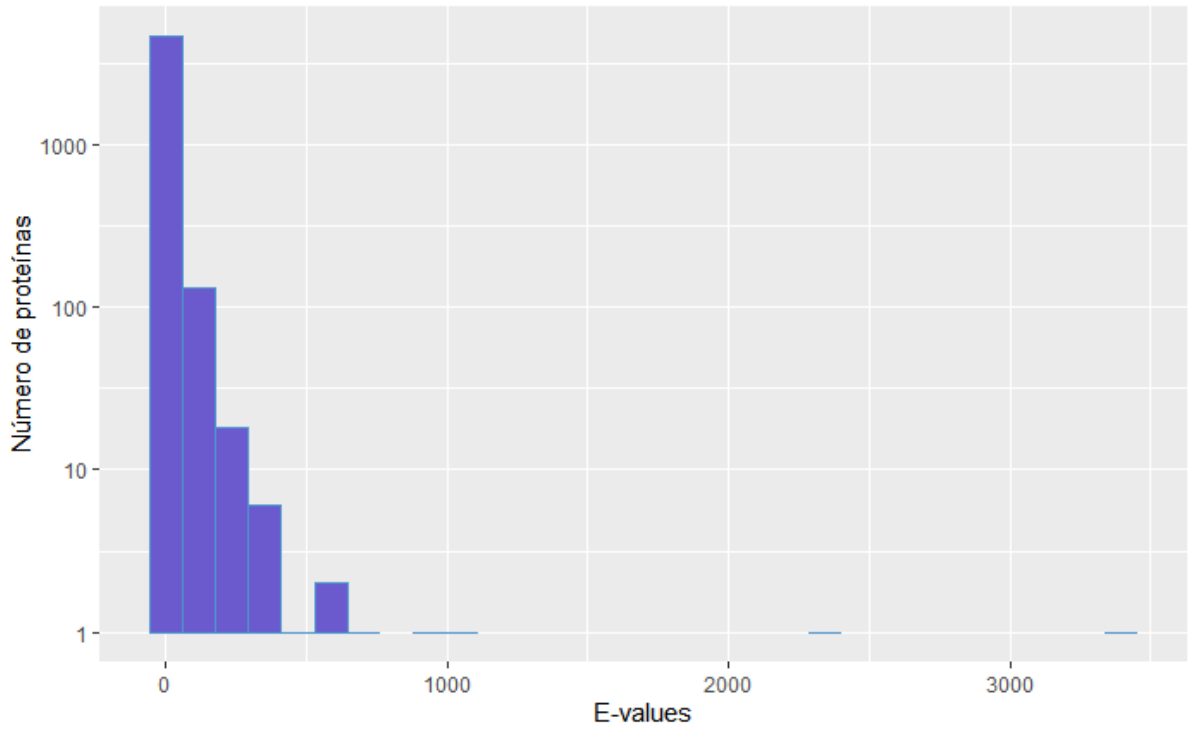
Diagrama del funcionamiento básico del HHbLits (extraído de Remmert et al. 2011). [66]

3.4.2 Análisis de los resultados del uso de HHblits:

En primera instancia, se analizaron las 4805 secuencias a las cuales todavía no se les había asignado ninguna estructura ni por evidencia preexistente (PDB), ni utilizando los métodos de similitud secuencial BLAST. El total de las 4805 retornaron, al menos, un hit demostrando alta capacidad de procesamiento de este algoritmo. Sin embargo, el problema es que al comprar los rangos de E-values e identidad y coverage de estas secuencias, encontramos que hay rangos mucho más amplios, indicando que, a pesar de tener una alta

capacidad de detectar secuencias homólogas, estos resultados tienen una probabilidad más alta de ser no significativos: el algoritmo es más sensible pero también tiene una mayor tendencia a encontrar falsos positivos.

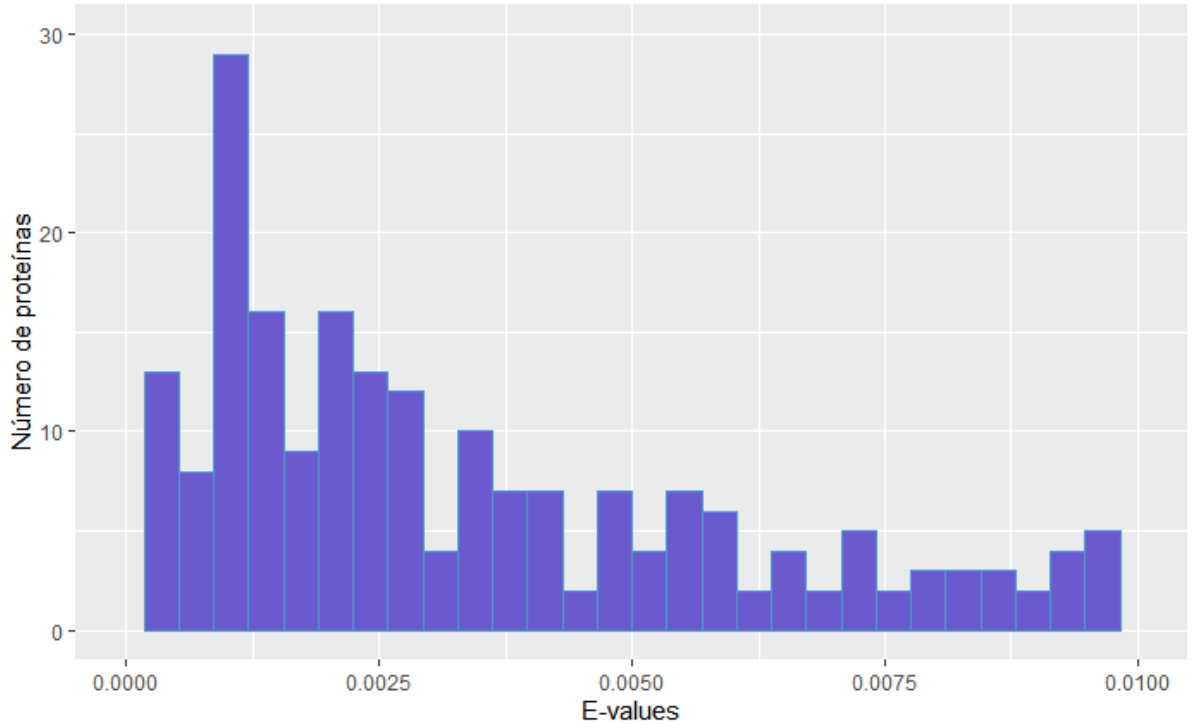
Haciendo el análisis en un orden similar al de la sección anterior, podemos comenzar estudiando los distintos E-values de cada uno de los pares secuencia-estructura:



Recuento de secuencias por E-value. [67]

A primera vista, no es posible observar en detalle la exacta distribución [67] en rangos aceptables de este score, pero podemos ver como el rango en el eje X es muchísimo más amplio que la del gráfico [47] de la sección anterior. Esto ya nos está indicando que, a pesar de haber obtenido un hit para todas las proteínas que fueron sometidas al HHblits, muchos de estos son falsos positivos.

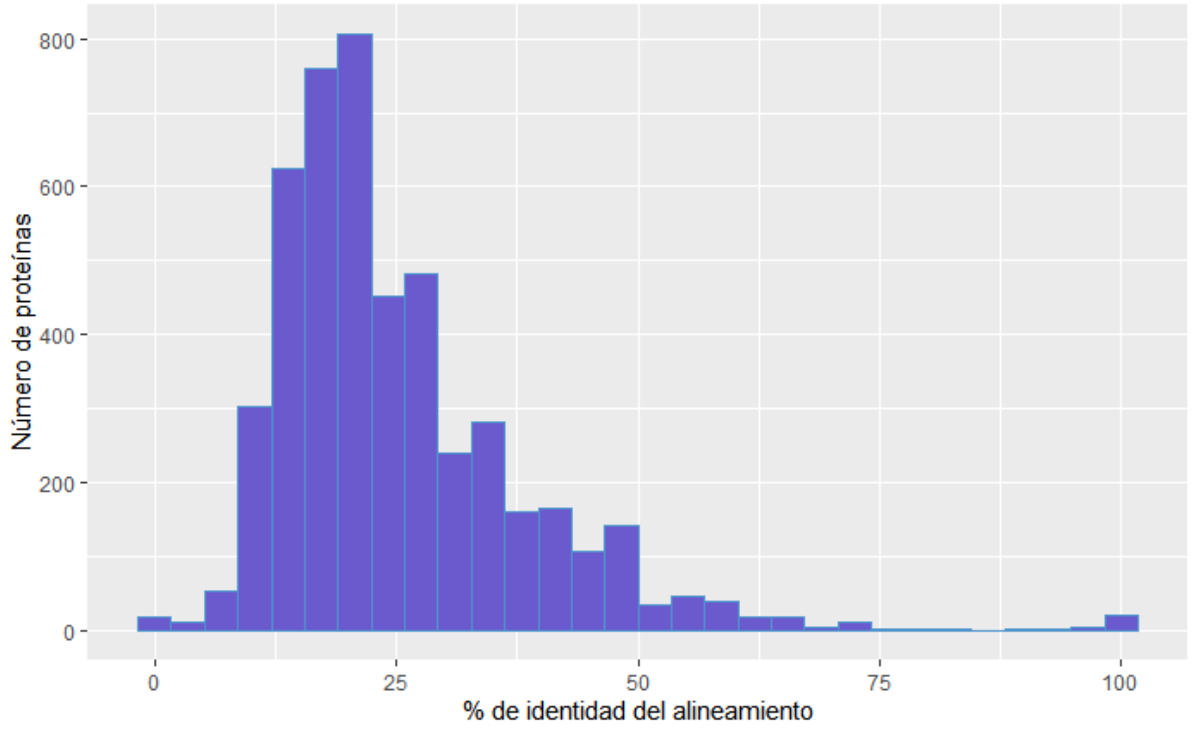
Si estudiamos los rangos entre 0 y un cutoff de 0.01 encontramos la siguiente distribución:



Distribución de E-values por proteínas. [68]

En este histograma [68], podemos ver como tenemos una distribución menos homogénea con una población mucho menor dentro del rango aceptable. En total, de las 4086 secuencias analizadas, solo 2098 están dentro de este gráfico: un 43,65% del total. Este es el punto de partida del que analizamos cuáles de estas estructuras podemos asignarles realmente a las secuencias y cuales debemos descartar.

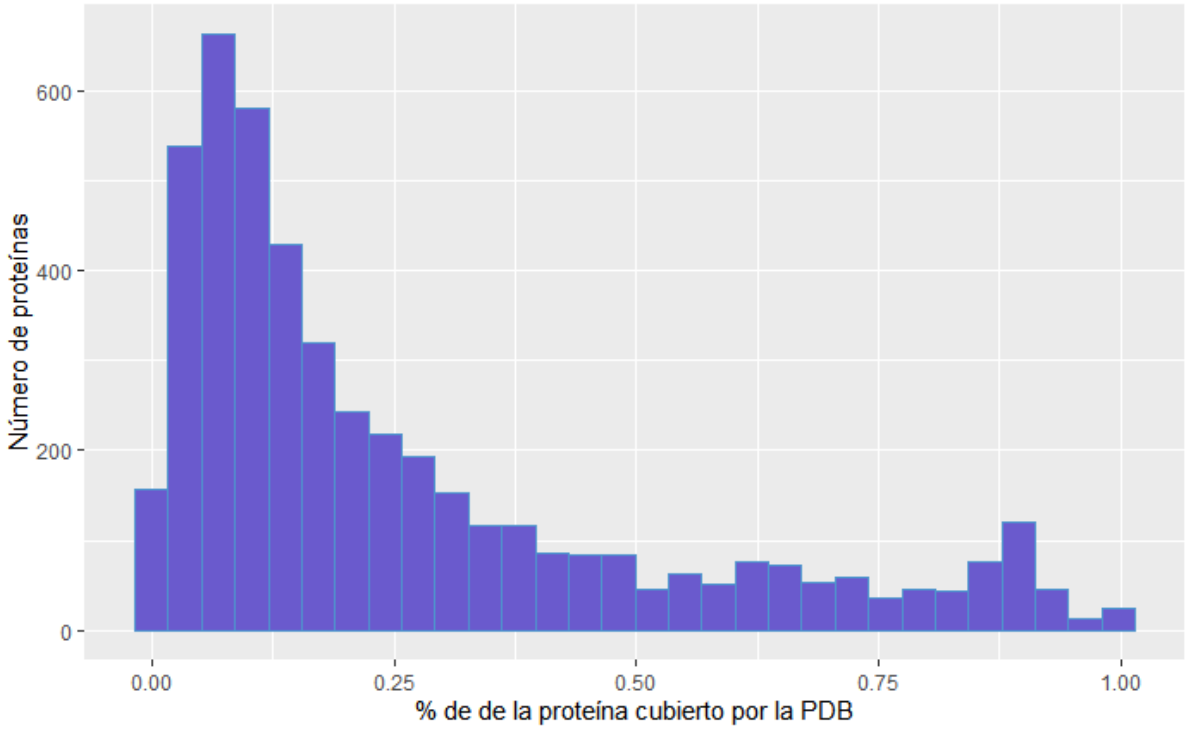
Continuando con el análisis, vamos a estudiar el nivel de identidad entre las secuencias y las estructuras:



Número de secuencias por % de identidad. [*69*]

En la figura [69], vemos un resultado que coincide con el razonamiento anterior: la población tiene un corrimiento hacia la izquierda, teniendo una densidad más alta en ella primera mitad del eje X (entre 0 y 50). Si seguimos usando un 30% mínimo de identidad como parámetro de corte, del total de secuencias iniciales solo nos quedamos con 1374 secuencias, que representan un 28,6%. Esto nos está indicando que a pesar de haber encontrado posibles estructuras a las distintas secuencias analizadas, los segmentos que matchean entre estos dos no son idénticos (son evolutivamente más distantes, Chothia and Lesk 1986).

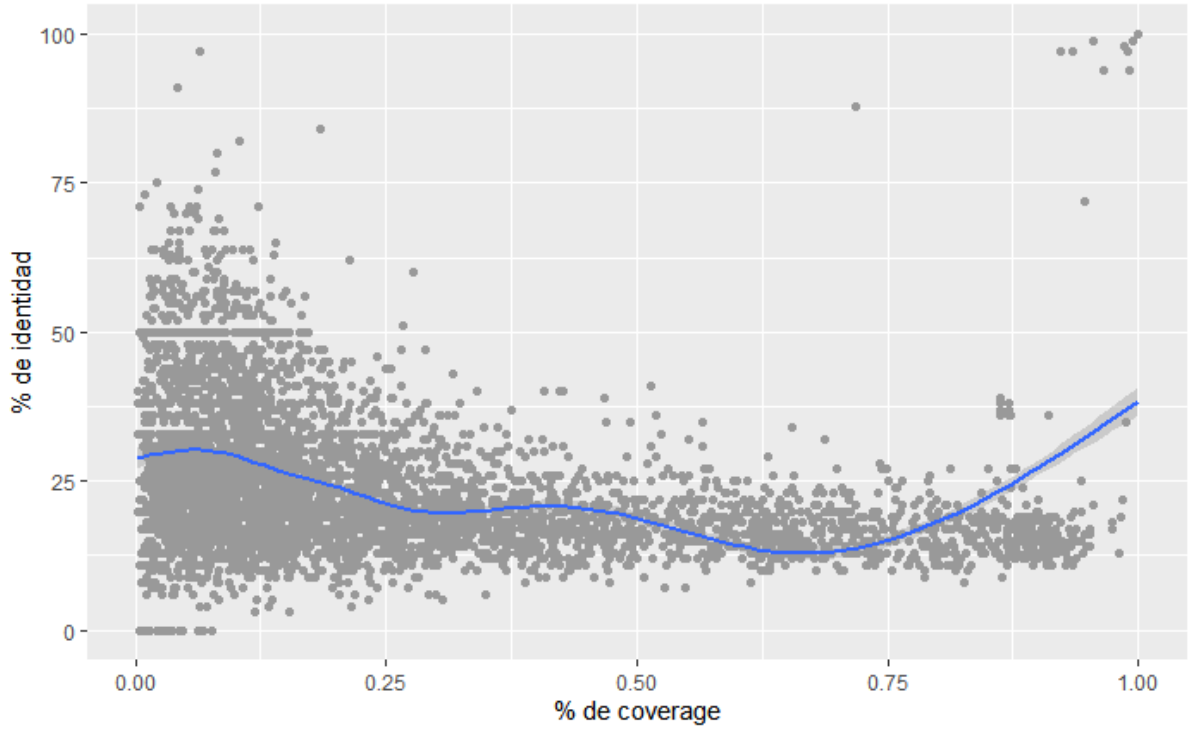
Por último, queda analizar la cobertura de las estructuras encontradas con el total de la longitud de la secuencia (el alineamiento que realiza HHblits el local, por lo que hay que hacer un paso extra, al igual que como hicimos con los resultados de BLAST, para encontrar estos resultados). Si graficamos un histograma, podemos encontrar lo siguiente:



Distribución del coverage por proteína. [70]

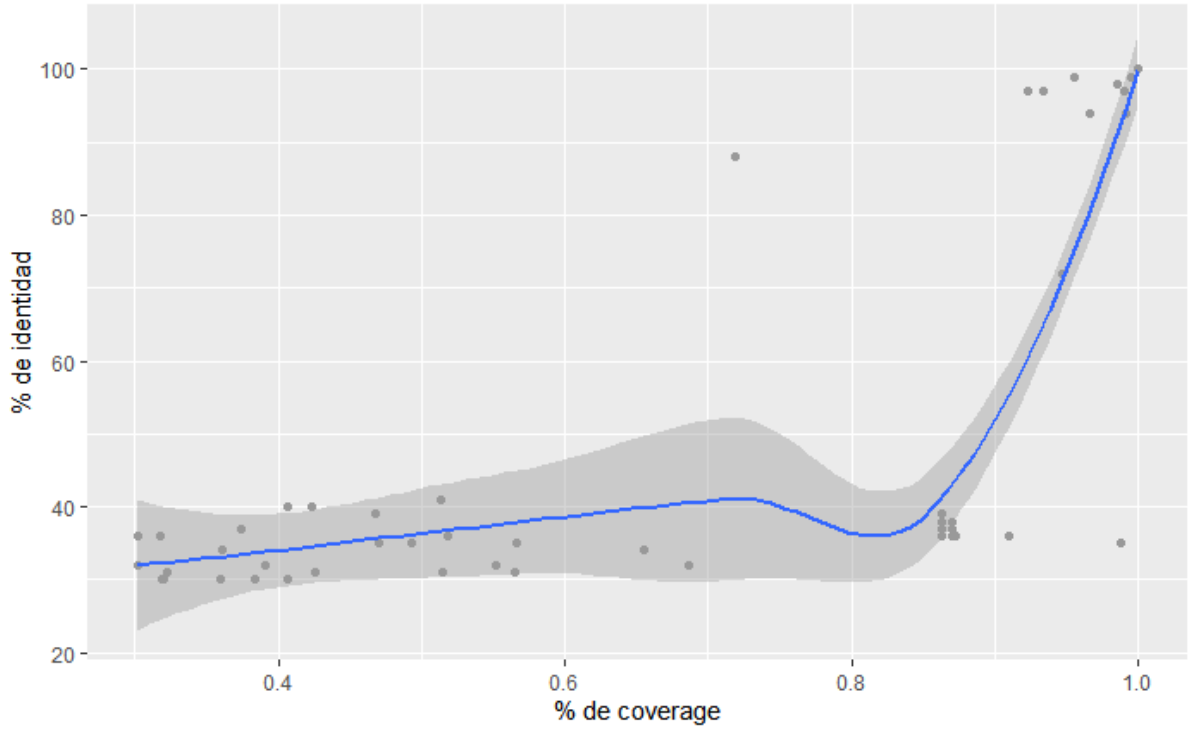
Observando la tendencia en este gráfico [70], podemos ver lo opuesto al resultado obtenido en la sección anterior, gráfico [60]. Hay muy pocas proteínas con coverages mayores al 50%, y la gran mayoría de estas secuencias se encuentran en el primer quintil (de 0 a 25%). Esto, de nuevo, nos demuestra que el método es lo suficientemente poderoso para detectar dominios evolutivamente lejanos y cortos, pero, para la aplicación que queremos darle nosotros, no necesariamente nos vemos beneficiados de esto. Si establecemos un cutoff del 30%, nos quedamos con 1428 secuencias (un 29,71%).
En total, si tenemos en cuenta los parámetros de corte para estos 3 valores, E-value menor a 0.01 e Identidad y Coverage mayores al 30%, nos quedamos solamente con 73 secuencias que cumplan con las 3 condiciones. Este número es increíblemente bajo, pero coincide con lo esperado.

Ahora podemos realizar unos últimos análisis, al igual que realizamos con los pares secuencias/estructuras de BLAST. Primero, podemos estudiar la distribución (para todo el dataset) de la relación de los porcentajes de Identidad y de Coverage:



Relación obtenida entre el porcentaje de identidad y el coverage. [71]

Como podemos observar en [71], hay una correlación casi constante entre estos dos parámetros. A pesar de que los coverages en general parecen estar más o menos bien distribuidos, la identidad de los distintos matches son casi constantes, alrededor del 30%. Existen secuencias que difieren con esta regla (hay un cluster de bajo coverage y alta identidad, y unas pocas secuencias que presentan alto coverage e identidad), pero la gran mayoría del dataset se comporta similarmente.

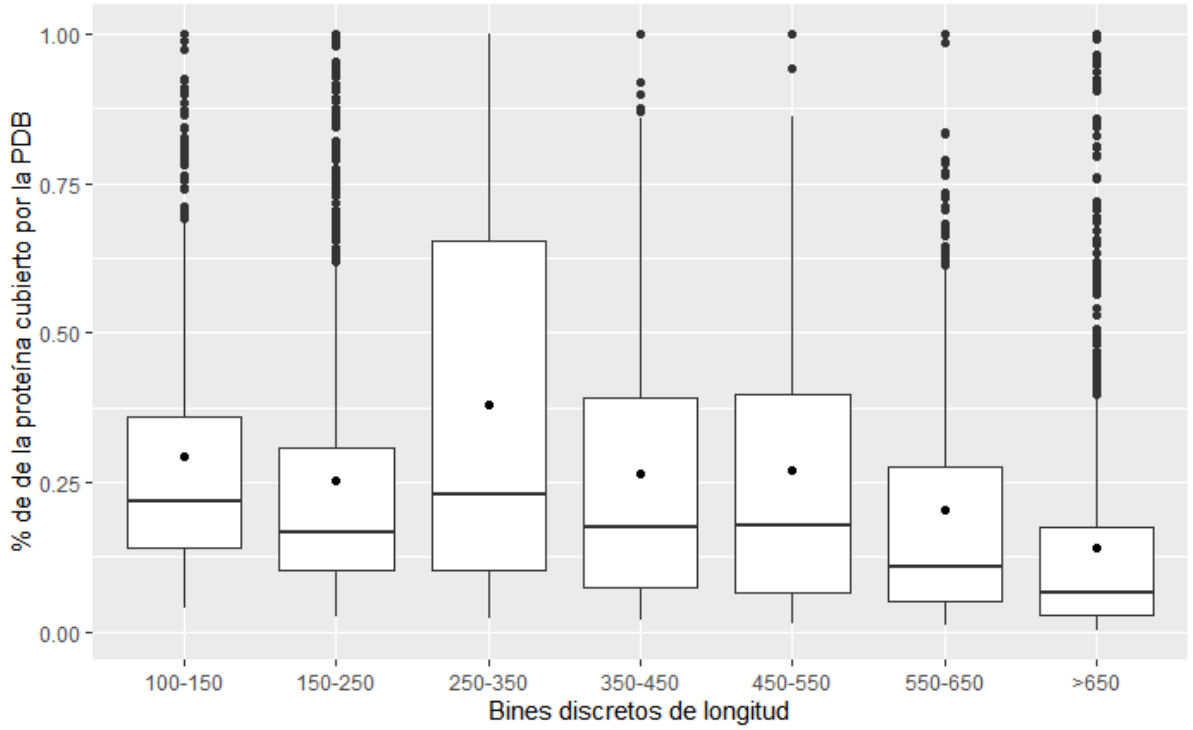


Relación Identidad/Coverage después del filtrado. [72]

Luego de usar los valores de cutoff para filtrar los matches no significativos, nos encontramos con un gráfico muchísimo menos poblado, como se observa en la figura [72], en el que observamos un panorama similar. A pesar de que existe un pequeño cluster de secuencias con alta identidad y coverage, la mayor tendencia se observa en secuencias que tienen niveles de identidad más bien pobres, sin importar el coverage que abarquen de la secuencias. Esto, una vez más, coincide con lo esperado: el método es, efectivamente, muy sensible, pero no es exacto. Estamos encontrando secuencias que si bien son homólogas, se encuentran muy distantes evolutivamente, por lo que su identidad es solo un remanente de la actual.

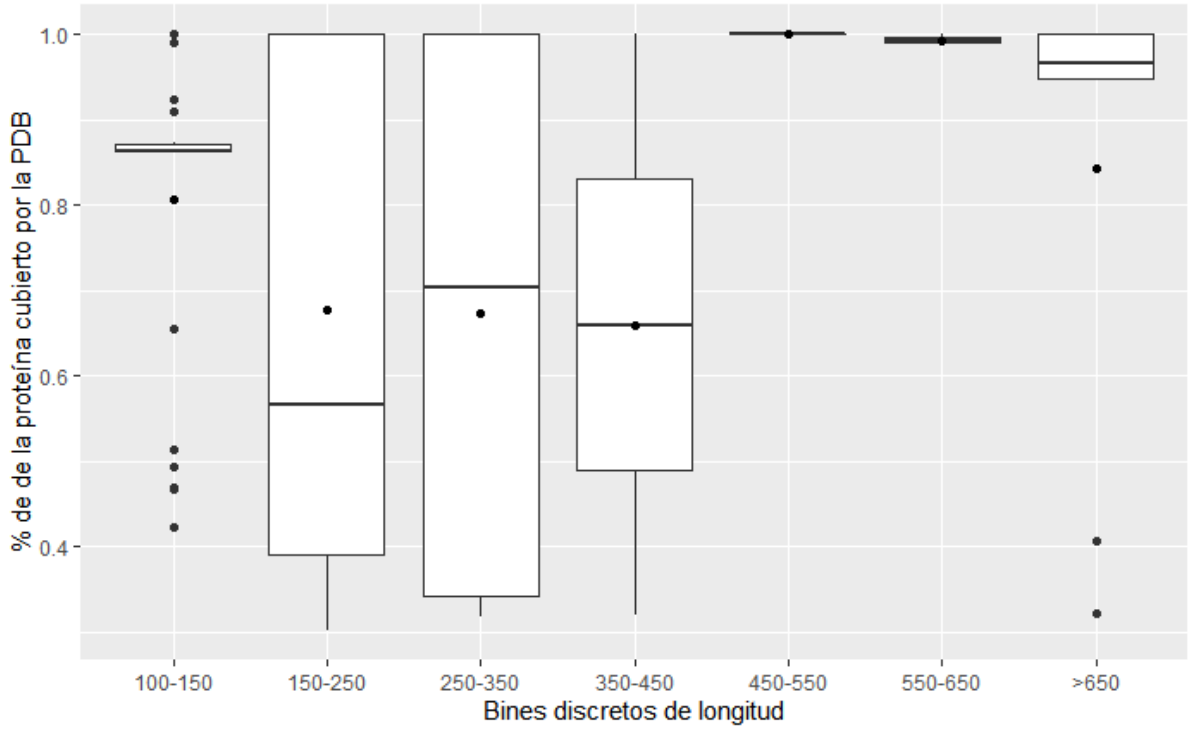
El siguiente paso es estudiar cómo se comparan las estructuras encontradas con respecto a las secuencias en términos de longitud: en los gráficos [61] y [64] de la sección anterior, pudimos ver una tendencia a que proteínas más cortas tengan porcentajes de coverages más altos, debido a que estas solo tenían un dominio, o presentaban bajos niveles de desorden, por lo que eran capaces de ser abarcadas en su enteridad por una estructura cristalográfica.

En el caso del dataset, obtenemos este gráfico para la relación % de Coverage/Longitud para todos los matches:



Relación Coverage / Longitud. [73]

Donde podemos encontrar, en la figura [73], una tendencia muy pobre entre los dos parámetros. Hay una caída a medida que tenemos longitudes más largas, pero desde el punto de partida los coverages medios están alrededor de un 25%, a pesar de haber outliers.

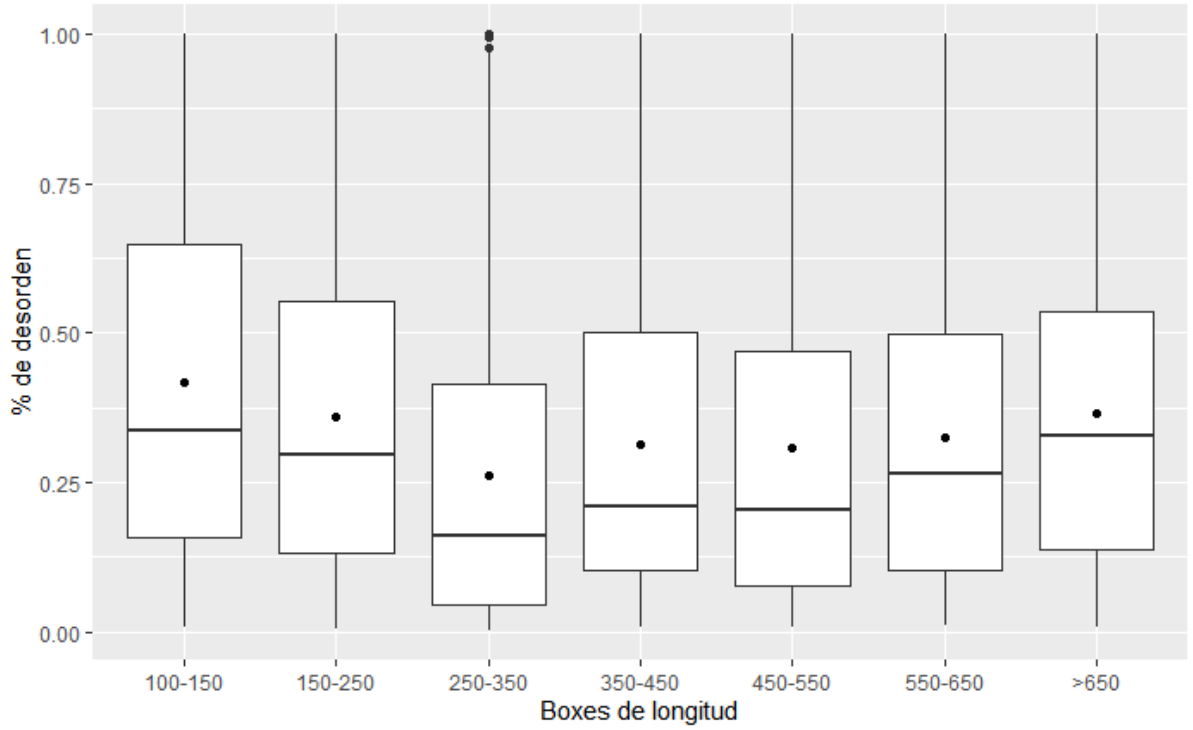


Relación coverage/ longitud post filtrado. [74]

Al analizar las secuencias después del filtrado, como se observa en la figura [74], es remarcable cómo aumenta la tendencia a tener coverages por arriba del 65% (el filtrado se hizo en base al 30% de coverage), demostrando que hay un “espacio” no ocupado por las estructuras: la gran mayoría tiene coverages bajos, del aproximadamente un 25%, pero existen unos pocas estructuras con coverages altos, con una gran tendencia a tener una cobertura por arriba del 70%.

Por último, debemos estudiar los perfiles de desorden de estas secuencias: la carencia de estructuras puede deberse a que, entre otras cosas, estas proteínas tienen altos niveles de desorden, por lo que sería imposible encontrarles una estructura homóloga.

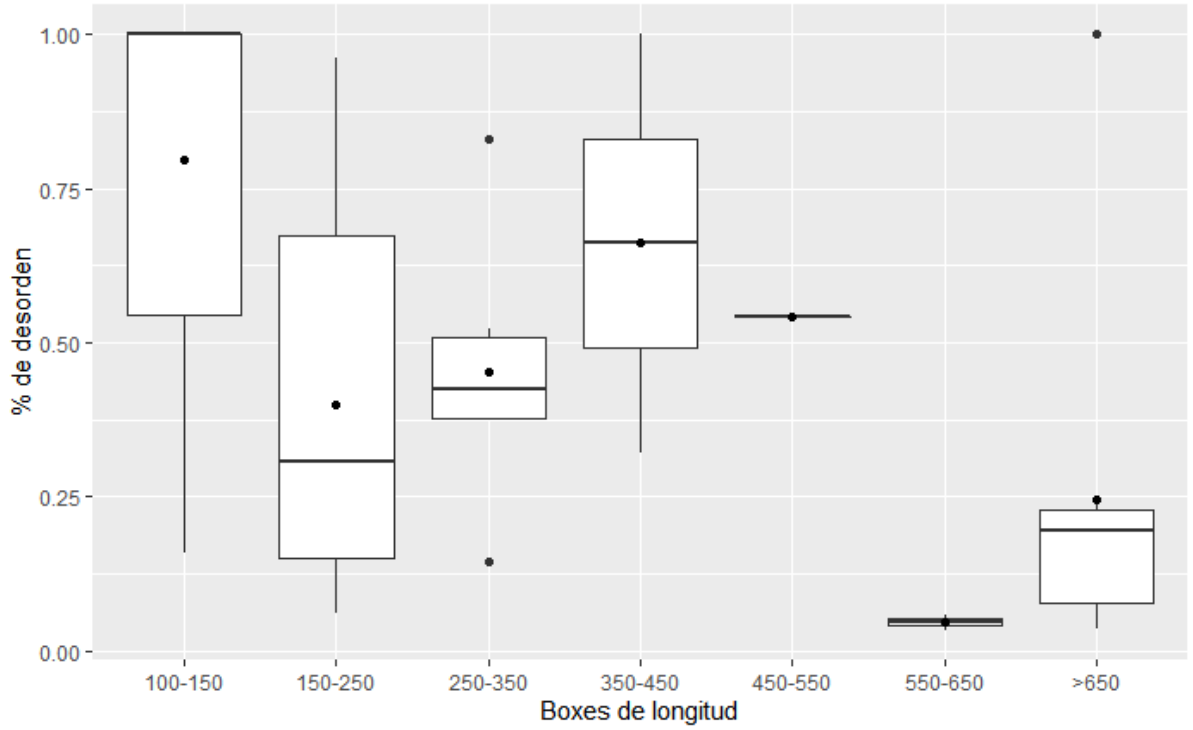
Si graficamos el % de desorden de las distintas secuencias con respecto a su longitud, obtenemos el siguiente resultado:



Porcentaje de desorden vs Longitud. [75]

En el gráfico [75] podemos ver otra tendencia que resulta muy interesante: las secuencias tienen niveles de desorden más alto que las que fueron reclutadas por medio del BLAST, gráfico [63] de la sección anterior (3.2). Esto resalta el siguiente razonamiento: puede que una de las razones que estas secuencias no tienen estructuras humanas en la PDB, ni que se les haya encontrado una estructura por medio del BLASTP es que tengan suficiente desorden en su estructura para impedir ambos procedimientos. El HHblits es capaz de reclutar estructuras homólogas debido a que su algoritmización más avanzada con respecto al BLASTP es capaz reclutar estructuras evolutivamente lejanas que hayan sido cristalizadas gracias que sus secuencias tengan niveles de desorden más bajo. Esto se alinea con los bajos niveles de identidad de las secuencias: pueden existir cambios en las mismas que faciliten el proceso, debido a que vuelven a las estructuras más o menos rígidas.

Finalmente, si observamos el mismo gráfico pero después de haber filtrado las estructuras, encontramos lo siguiente:

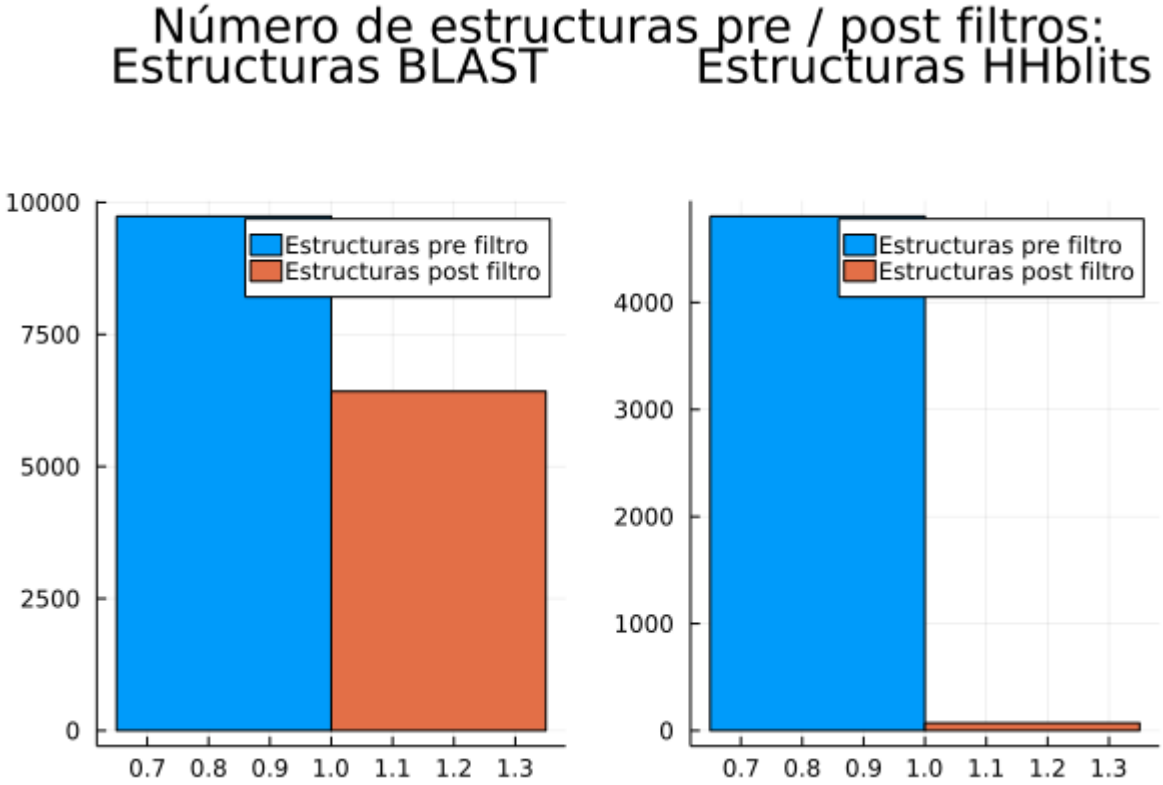


Porcentaje de desorden vs Longitud post filtrado. [76]

En este caso, figura [76], podemos ver un aumento en el % de desorden medio por estructura, pero con una tendencia muy incompleta. Esto se debe, en gran parte, al poco volumen de datos, solo hay 73 secuencias en este caso, y a pesar de que en varios rangos de longitudes parece haber un considerable aumento en el desorden (rangos 100-150, 350-450, 450-550), en otros se observa en comportamiento opuesto, con muchos datos dispersos en el eje Y.

3.5 Conclusión

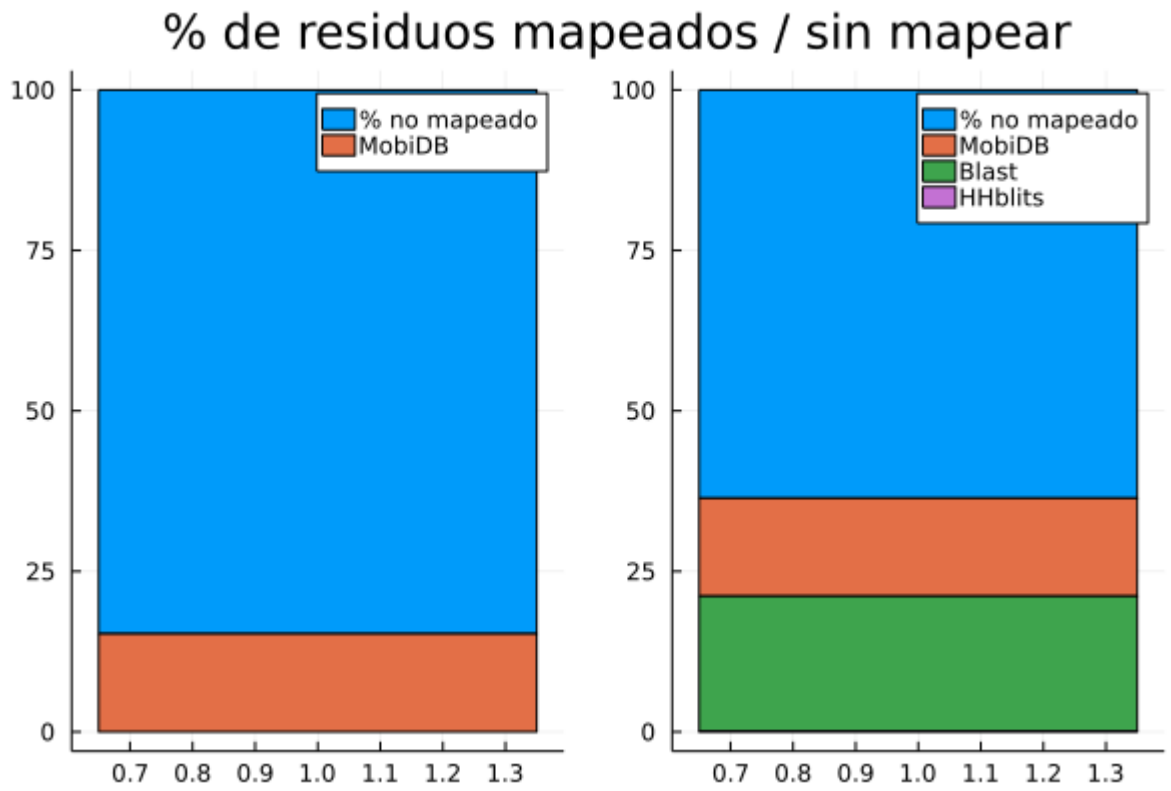
En este capítulo estudiamos dos métodos distintos para el reclutamiento y asignación de estructuras: BLAST y HHblits. Ambos métodos tienen la misma base (buscan proteínas homólogas en base a la información secuencial), pero difieren algorítmicamente. Los dos métodos cumplen su propósito para nuestro análisis, pero es importante señalar el volumen de estructuras que fueron reclutadas antes y después de aplicar los 3 filtros, identidad, coverage e E-value, para poder asignar (o no) la estructura a una cierta proteína:



Número de estructuras asignadas con BLAST / HHblits. [77]

Como se puede observar en el gráfico [77], el número de estructuras descartadas cuando usamos el BLAST es mucho menos que al usar el HHblits: de 9738 estructuras encontradas, descartamos 3315 para el set de BLAST, un 34%. En cambio, en el caso del HHblits, de 4806 secuencias encontradas descartamos un total de 4734, un 98,5%. Sin embargo, es importante tener en cuenta que las proteínas que fueron seleccionadas para correr en el HHblits eran las que no habían retornado resultados positivos en el BLAST (y que tampoco tenían ninguna parte mapeada en MobiDB), por lo que podemos explicar este alto volumen de resultados negativos en base a esto: son las proteínas del proteoma humano con menor información estructural, y sus homólogos disponibles tampoco cuentan con información confiable.

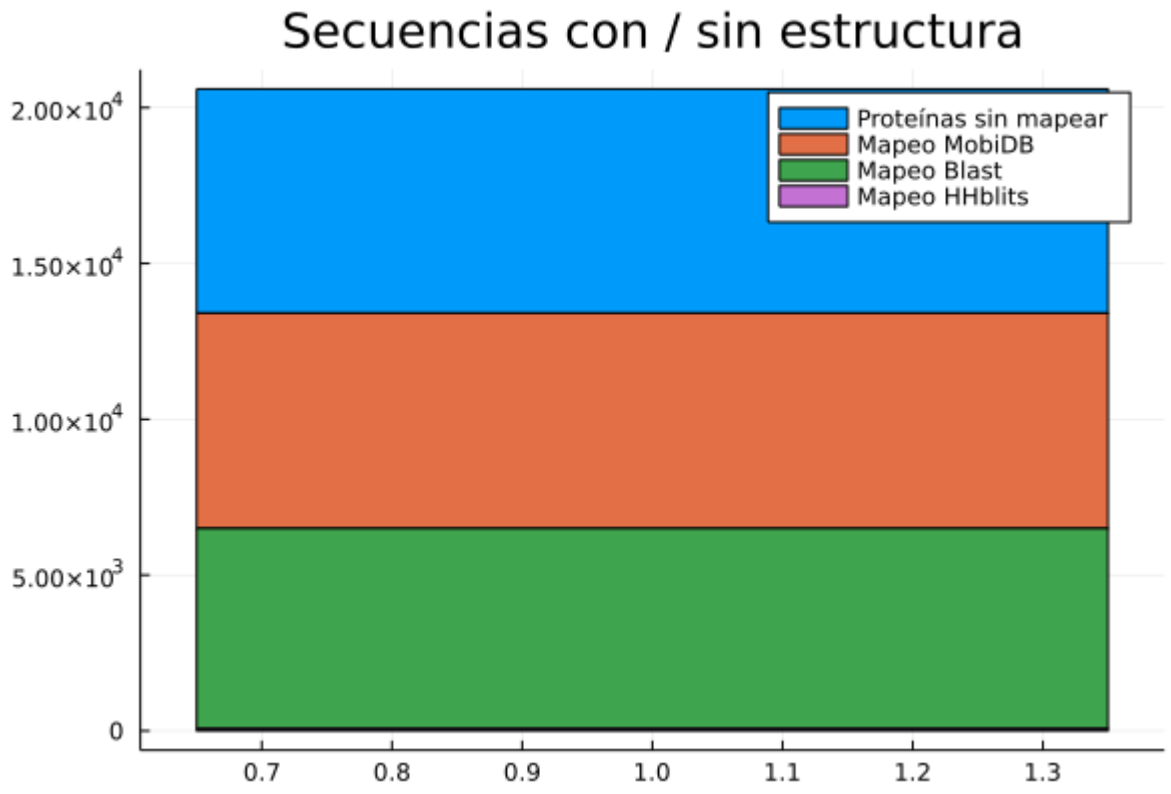
Ahora, pasando en limpio las estructuras encontradas, podemos hacer un análisis de en cuanto enriquecimos nuestro dataset. La métrica elegida para este análisis es el porcentaje de residuos mapeados:



Residuos del proteoma mapeados con estructuras. [78]

Como se observa en la figura [78], antes de la búsqueda por homología, había un 15,27% de los residuos totales del proteoma mapeados por la información estructural proveniente de MobiDB. Luego del uso de BLAST y HHblits, este porcentaje aumenta al 36,41% (un aumento del 21% y 0,14% respectivamente).

Esto parece indicarnos que tenemos un mapeo por residuo relativamente bajo (menor al 50%), pero cuando observamos las proteínas que ahora tienen al menos una estructura asignada, encontramos lo siguiente:



Mapeo estructuras del proteoma. [79]

Como vemos en el gráfico [79], el número de proteínas con al menos una estructura supera el número de proteínas sin estructura: quedan 7194 secuencias sin estructura (un 34,92%), y sumando la asignación por BLAST y MobiDB, tenemos un total de 13406 secuencias con al menos una estructura (65,08%): todas las estructuras pertenecientes a estas estas secuencias son los objetivos a caracterizar para poder finalizar este trabajo.

Como conclusión, en este capítulo pudimos verificar la potencia y utilidad de los métodos de búsqueda por homología: por medio de su uso, encontramos y asignamos estructuras a pesar de haber utilizado estructictos parámetros de corte a más de la mitad del proteoma humano. Sin embargo, no hay que perder de vista el objetivo: es importante analizar estas estructuras y caracterizarlas para llegar a las conclusiones pertinentes de esta tesis.

4. Diversidad estructural:

4.1 Introducción

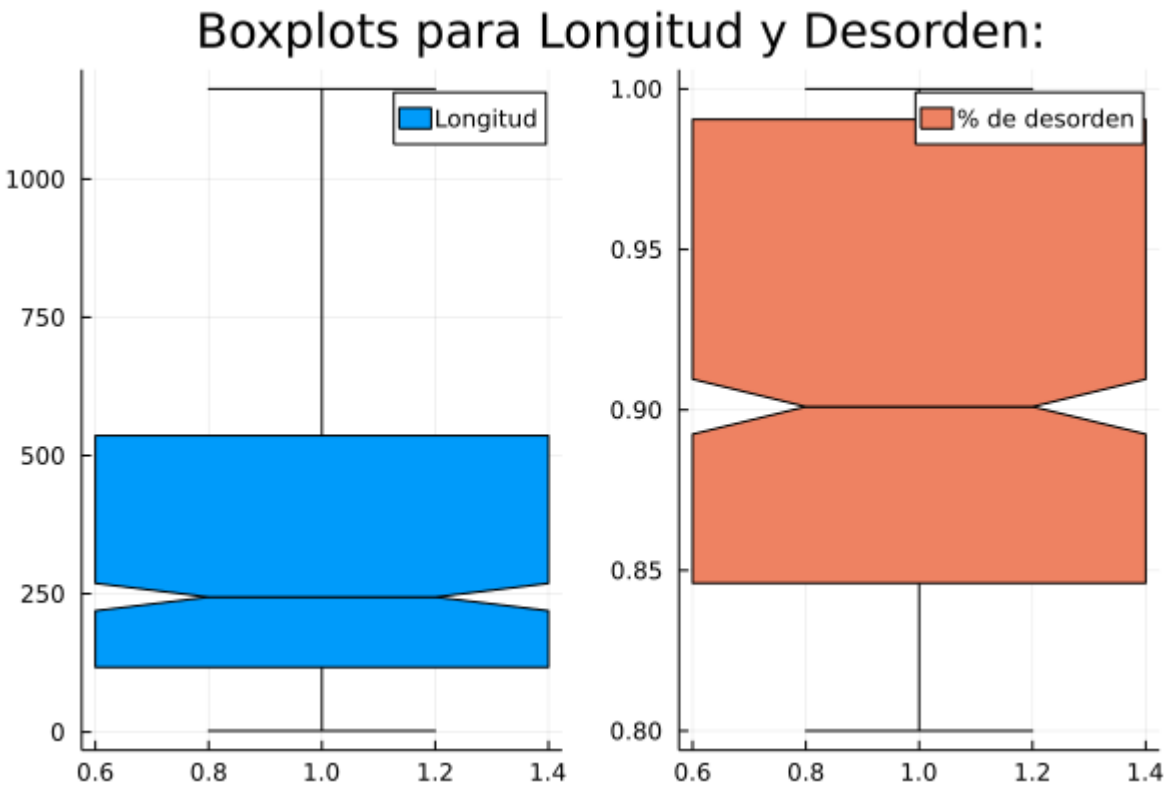
Las primeras oraciones de esta tesis proponían que las proteínas, análogas a criaturas de un cuento de Jorge Luis Borges, podían ser clasificadas casi de infinitas formas debido a que "...notoriamente no hay clasificación del universo que no sea arbitraria y conjetural. La razón es muy simple: no sabemos qué cosa es el universo". Esto nos lleva a tener que plantear una decisión muy importante: de todas las clasificaciones arbitrarias y conjeturales posibles de las proteínas, debemos definir cual vamos a usar nosotros. No importa cuál elijamos, es imposible que sea la correcta debido a que siempre van a existir modelos estructuras nuevos para las proteínas (AlQuraishi 2019), nuevos tipos de clasificación estructural (sin ir más lejos, el desorden se descubrió en la década de los '90, Dunker et al. 2001, Kriwacki et al. 1996, y recientemente CATH, una base de datos que estudiaremos más adelante, agregó una nueva clasificación para proteínas "especiales", Sillitoe et al. 2021), y también es posible que se agreguen nuevas secuencias, y por lo tanto, estructuras al pool de proteínas "consenso" humanas (UniProt Consortium 2008).

En los capítulos anteriores, partiendo de las 20600 secuencias en el proteoma humano, analizamos la estructura de 6911 proteínas con al menos una estructura por medio de la base de datos MobiDB, y con el uso de los programas BLAST y HHblits, pudimos asignar, de forma significativa, 6496 proteínas más, llegando a un total de 13406 secuencias con estructura asignada.
Por lo tanto, para finalizar este trabajo y a modo de conclusión, en este capítulo presentaremos la clasificación elegida y a modo de cierre un análisis que integre las distintas poblaciones de estructuras en el proteoma humano y perspectivas hacia el futuro.

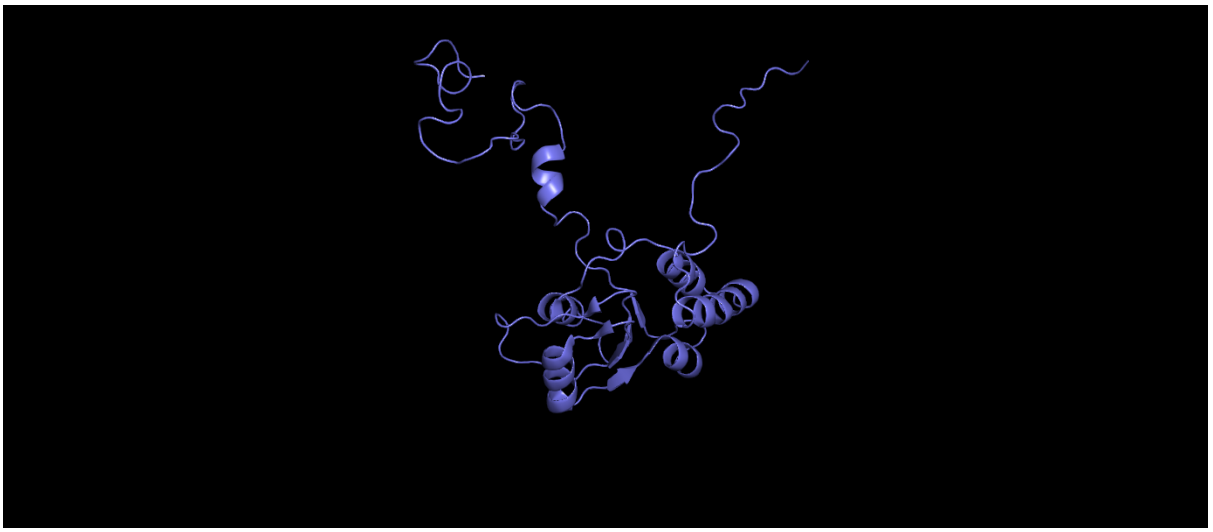
4.2 Proteínas desordenadas

El desorden fue, contradictoriamente, uno de los temas más hablados en esta tesis: realizamos un estudio preliminar en la sección 2.6, y posteriormente analizamos el contenido de desorden en relación a los mapeos estructurales que realizamos (secciones 3.2.2, 3.3.3 y 3.4.2). Sin embargo, la gran mayoría de estos análisis fueron muy generales, estudiando el total de proteínas con regiones desordenadas, sin importar mucho el tamaño de estas.
Es momento de hacernos la siguiente pregunta: ¿cuántas proteínas de nuestro proteoma son, en su mayoría, desordenadas? Para este análisis usamos el desorden consenso del 50% proveniente de las anotaciones de MobiDB (Piovesan et al. 2021, Di Domenico et al. 2012), que se basan en el consenso de distintos predictores: MobiDB-lite (Necci et al. 2020), ESpritz-DisProt, ESpritz-NMR y ESpritz-Xray (Walsh et al. 2012), IUPred-Long y IUPred-Sort (Dosztányi 2018), VSL2b (Katuwawala and Kurgan 2020), DisEMBL-465 y DisEMBL-HotLoops (Linding et al. 2003), GlobPlot (Linding et al. 2003) y por último JRONN (Yang et al. 2005).

Con este dataset podemos encontrar las proteínas que tengan más del 80% de su secuencia predicha como desordenada. Es importante no perder de vista que estas proteínas no necesariamente tienen estructura predicha: al ser un análisis secuencial (como fue explicado en la sección 2.6) estas predicciones dependen solamente de la composición aminoacídica de las mismas, por lo que puede que estas tengan o no una estructura asignada a las regiones desordenadas.
En total, encontramos **711** proteínas con un nivel igual o mayor al 80% de desorden predicho en su secuencia, que representan al **3,45% del proteoma total**: tienen una media de 464 aminoácidos de longitud, e interesantemente, una media del 90% de desorden.
De estas 711 proteínas, 85 ya tenían una estructura asignada en MobiDB, 23 están presentes en el dataset de estructuras asignadas por HHblits, y 37 provienen del dataset generado por medio del BLAST. Es remarcable el gran número de proteínas desordenadas que hay en el dataset de HHblits (un 32%) comparado con el resto.
En total, de las 711 proteínas desordenadas, 154 (un 21,66%) están mapeadas a, al menos, una estructura.



Dispersión de los datos para proteínas desordenadas, [80]



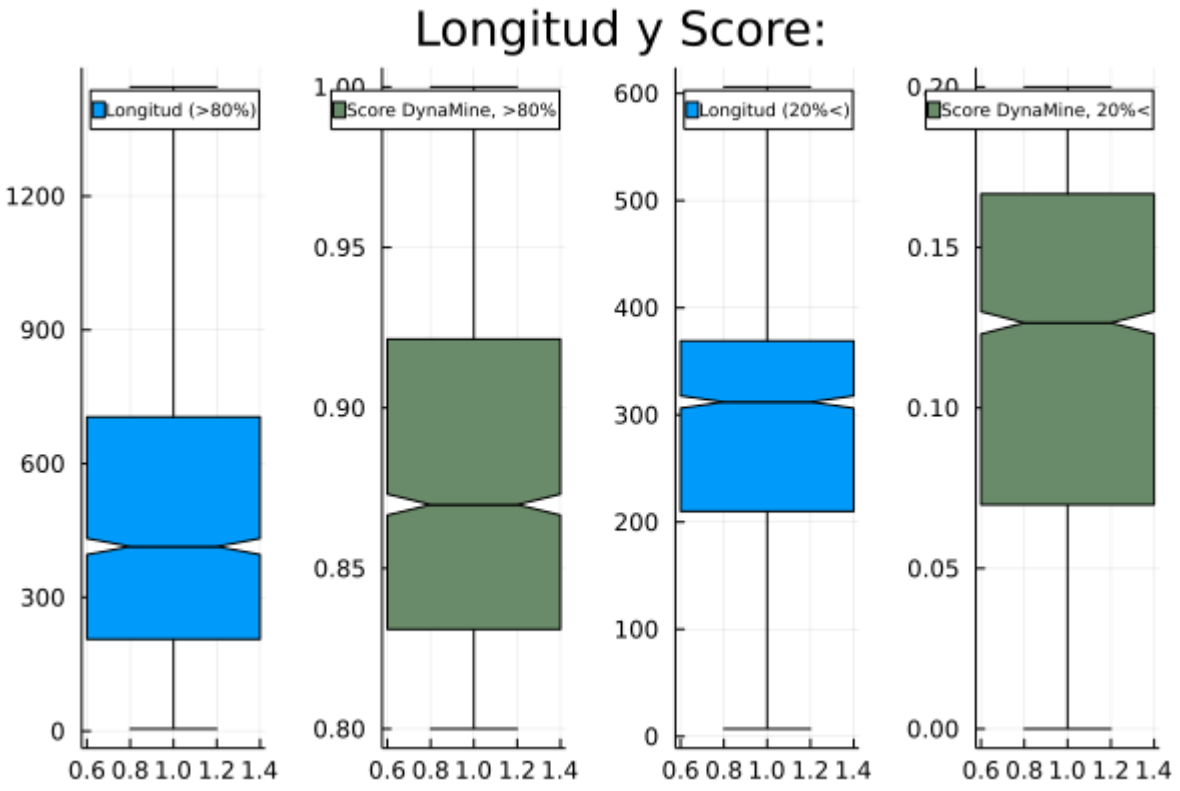
Estructura PDB de 6LU8, cadena Z, asignada a Q9BRT6 (83,7% de desorden predicho), [81].

4.3 Proteínas rígidas y flexibles

La clasificación de las proteínas entre flexibles o rígidas, en el contexto de scores globales de DynaMine para las proteínas (Cilia et al. 2013), es el otro tema abordado en capítulos anteriores además del desorden (secciones 1.5 y 2.7). Al igual que el desorden esta

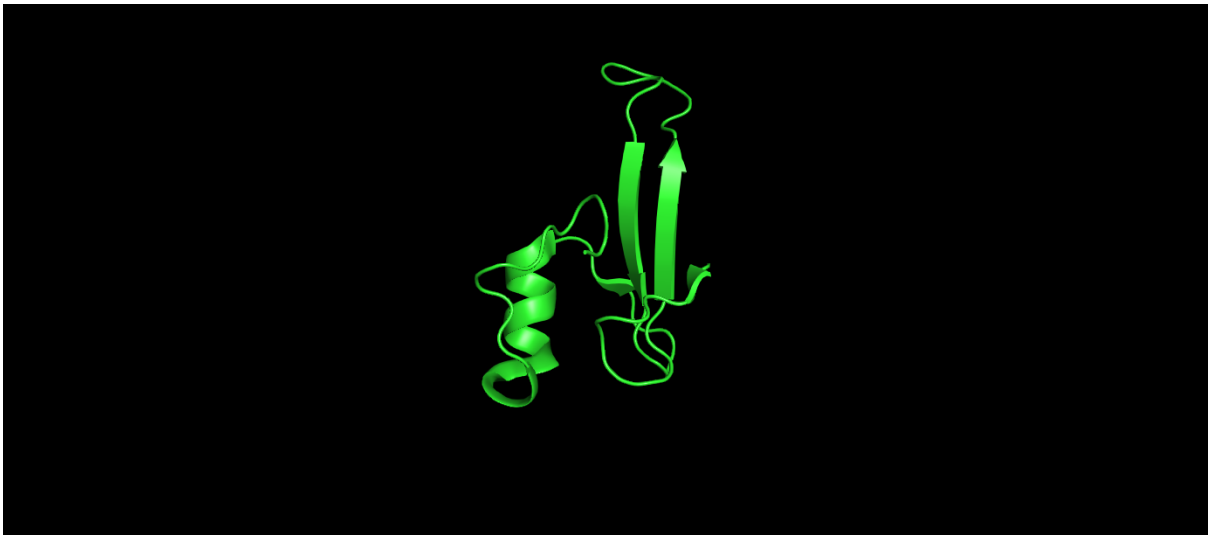
clasificación requiere solamente de un análisis secuencial que, en nuestro caso, fue obtenido de la base de datos MobiDB (citada en la sección anterior).
Análogamente, en lugar de analizar el panorama general para todo el proteoma, en esta sección nos vamos a centrar en dos grupos opuestos: las proteínas que clasificaremos como **rígidas**, que se caracterizan por tener un score global menor o igual al 20%, y las proteínas **flexibles**, que se caracterizan por scores mayor o iguales al 80%. De esta forma, podemos realizar un análisis similar al estudiado en la sección 1.5 (Monzon et al. 2017).

En nuestro caso, al analizar el dataset completo encontramos que **1894** secuencias entran dentro de la clasificación de flexibles (más del 80% de su secuencia tiene scores significativos de flexibilidad), representando al 9,2% total del proteoma, con una media de 88% en su score y una longitud media de 606 Aa, y **1905** secuencias son clasificadas como rígidas (mismo análisis pero aplicado a menos del 20% de su secuencia), el 9,24% del proteoma, con una media del 11% y una longitud media de 331 Aa. Es un resultado remarcable como las proteínas más largas parecen estar favorecidas por la flexibilidad, como puede observarse en la imagen [82]:

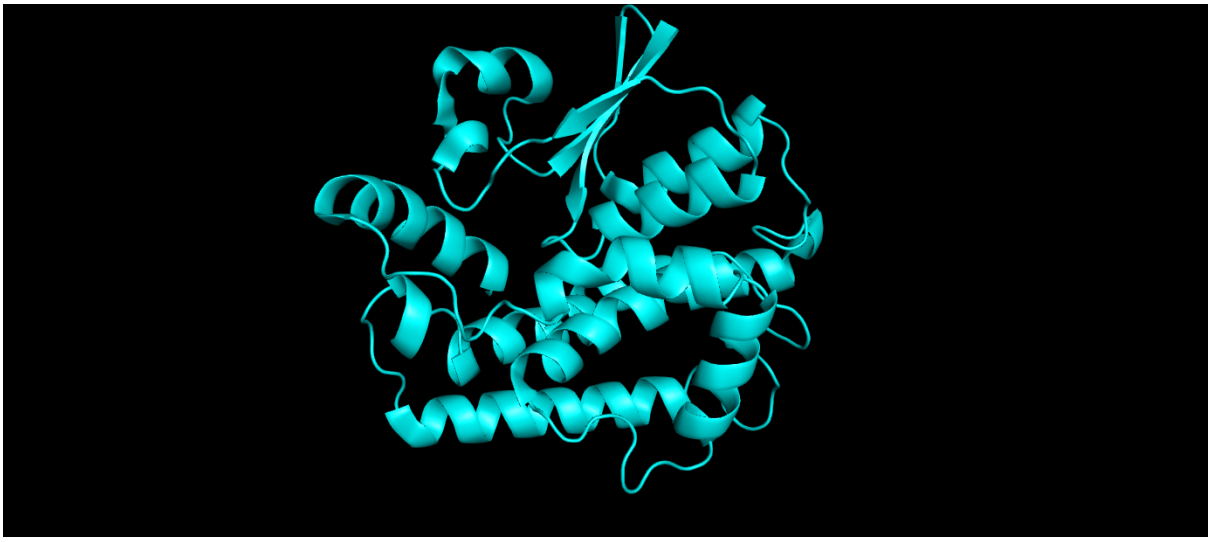


Diferencias de longitud entre proteínas flexibles y rígidas, [82].

En cuanto al mapeo estructural, hay un total de 528 proteínas flexibles con estructuras asignadas (277 por MobiDB, 222 mediante el BLAST y 29 mediante el HHblits), generando un mapeo efectivo del 27,88%, y un total de 726 secuencias rígidas con estructura asignada (257 provenientes de MobiDB y 469 asignadas mediante el BLAST), un 38,11% del total de las proteínas rígidas. Al tener en cuenta la relación del desorden y la flexibilidad, el resultado del menor mapeo estructural para con las proteínas flexibles se vuelve lógico.



Estructura PDB de 7AOA, cadena C, asignada a A0A1B0GVZ6 (Score en DynaMine del 99,01%), [83].



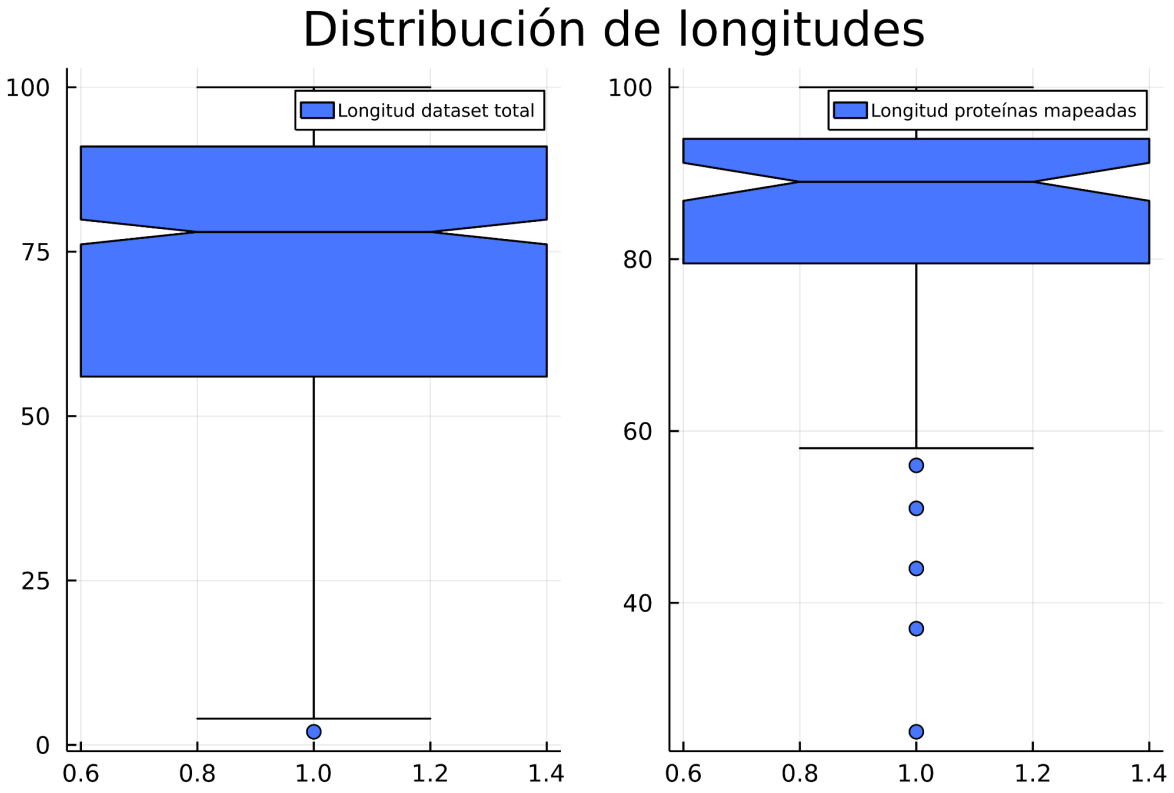
Estructura PDB de 2C3T, cadena D, asignada a A0A1W2PRG0 (Score del 15,3%), [84].

4.4 Proteínas pequeñas

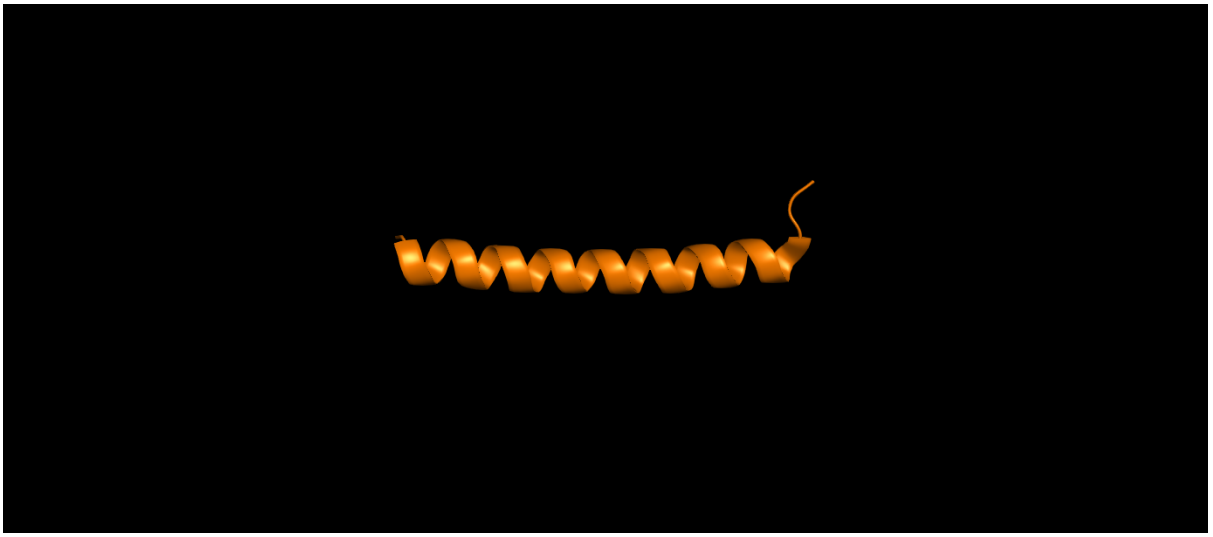
Las proteínas pequeñas (y sus estructuras) van a ser las últimas proteínas que vamos a caracterizar de las cuales ya hemos hablado anteriormente. Como vimos en la sección 2.4, hay un pequeño subset de proteínas que integra este dataset, y que presenta un interesante grupo de estudio debido a sus distintos roles en las células (Muranova et al. 2019; Collier and Benesch 2020; Camby et al. 2006).
En nuestro caso, del total de las 20600 secuencias en el proteoma de referencia, **855** (4,15%) son proteínas cortas, con una longitud menor o igual a 100 aminoácidos, con una longitud media de 69,32 aminoácidos.

De este total de secuencias sólo 107 (un 12,51%) tienen alguna estructura asignada, y todas provienen de las estructuras humanas anotadas en MobiDB. Ninguna proteína corta

tuvo un match válido tanto para el análisis y búsqueda por BLAST como por HHblits. Un resultado relevante es que la longitud promedio de las proteínas con estructura en MobiDB es mayor a la del todo el dataset, con un valor de 84,65 Aa.



Longitudes de proteínas cortas sin y con estructura, figura [85].

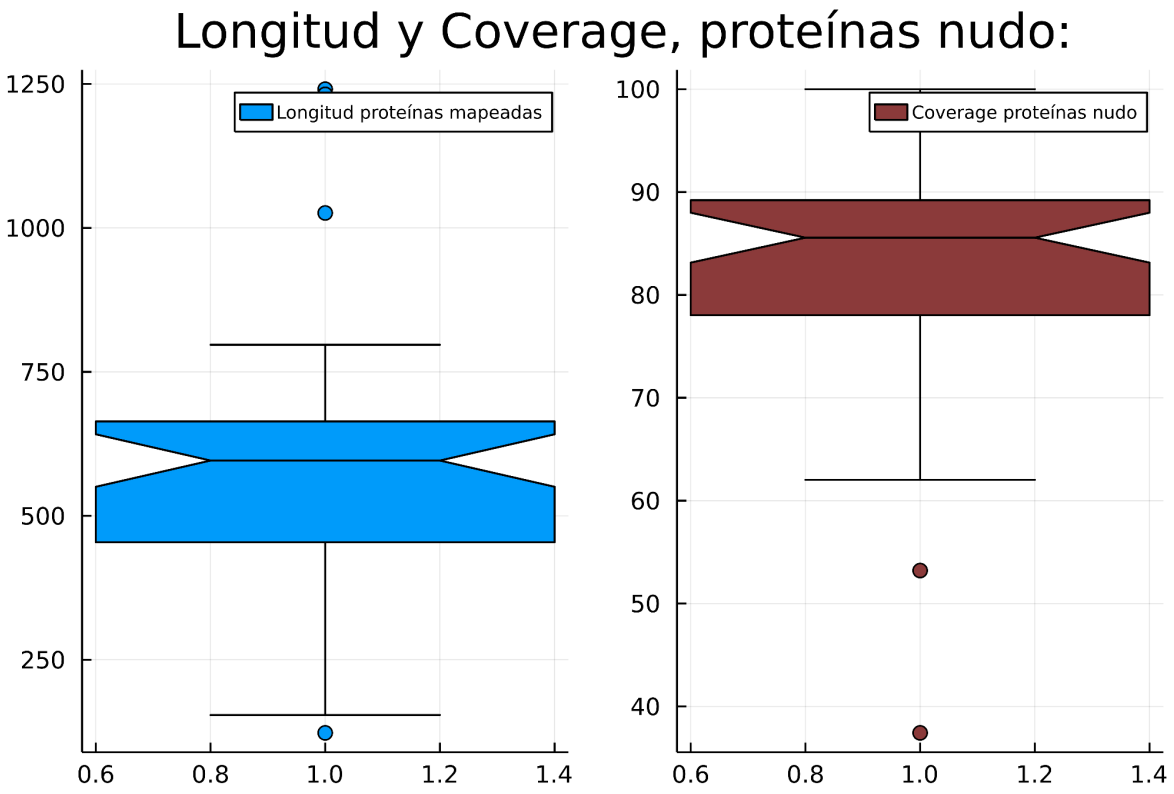


Estructura PDB de 6S7T, cadena B, asignada a P0C6T2, (Longitud de 37 aminoácidos), [86].

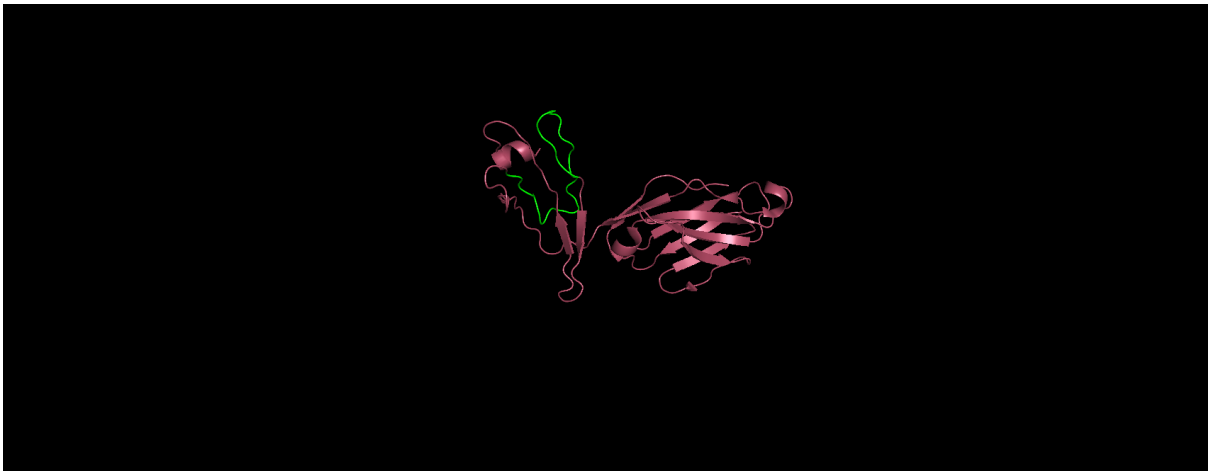
4.5 Proteínas nudo

Las proteínas nudo son un tipo de estructura recientemente descubierta (Mansfield 1994, King et al. 2007,Sułkowska et al. 2012) que se caracteriza por tener un nudo lineal y abierto en su estructura nativa. Este tipo de estructura, si bien parece simple de identificar, requiere complejos cálculos matemáticos aplicados a nudos polinomiales para poder identificar si es posible formar esta estructura.
Desde un punto de vista evolutivo, se cree que las proteínas nudo fueron en parte eliminadas durante la evolución debido a que las proteínas que se pliegan lentamentamente y/o no-reproduciblemente deberían ser evolutivamente desventajosas para el organismo que las expresa (Virnau et al. 2006, Prentiss et al. 2010). Sin embargo, análisis más actuales parecen indicar que tienen un muy alto grado de conservación aún en proteínas provenientes de especies muy alejadas evolutivamente (Sułkowska et al. 2009).

Para nuestro análisis, usamos la base de datos KnotProt (Jamroz et al. 2015), que utiliza complejos cálculos matemáticos derivados del campo de teoría de nudos para identificar estructuras que los contengan. A la hora de identificar las distintas estructuras de proteínas humanas (o asignadas a proteínas humanas) que contienen nudos, encontramos un total de **241** proteínas humanas que contienen nudos. **188** provienen de MobiDB, y **53** fueron asignadas mediante el BLAST.
Podemos evaluar la distribución de longitudes y de coverages de estas proteínas en la figura [87], donde encontramos que tienden a tener buenos scores de coverages (las estructuras cubren la gran mayoría de la proteína) y que, la gran mayoría, son proteínas más grandes que el promedio.



Boxplots para la distribución de las estructuras nudo, imagen [87].



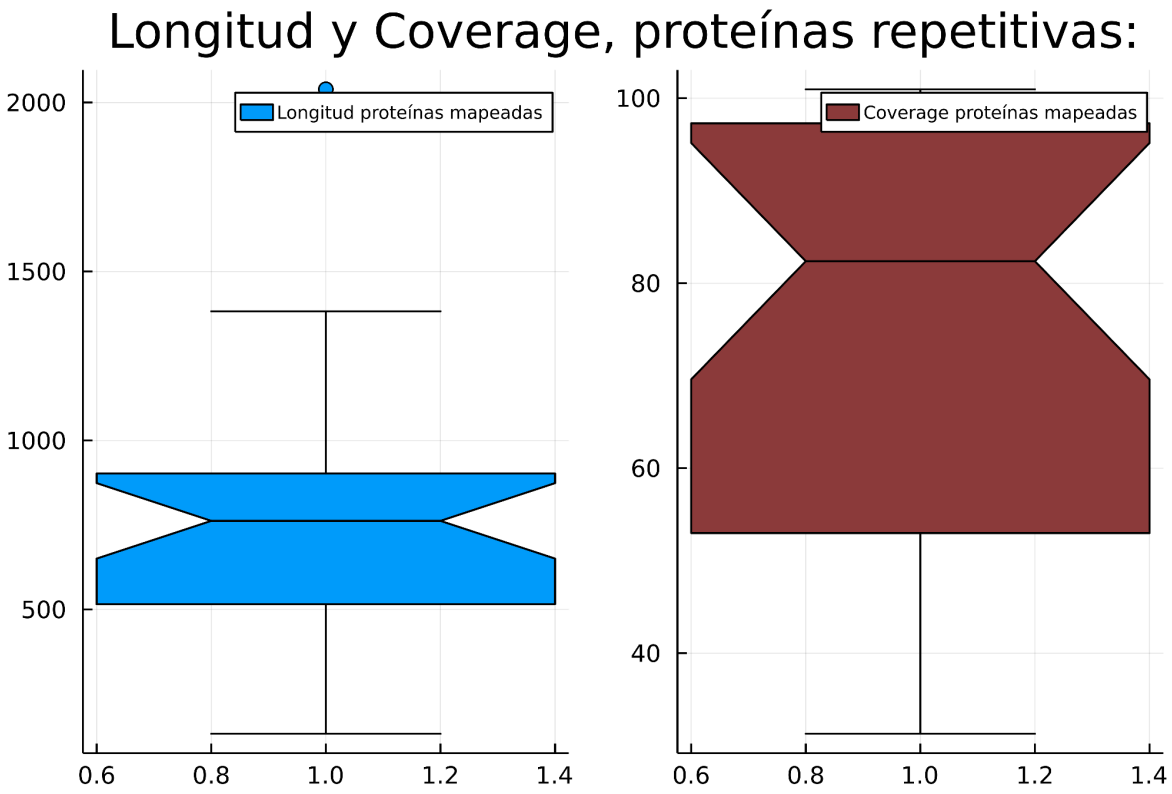
Estructura nudo (4WWK, cadena A, nudo interno señalado en verde) de la proteína humana A0A0B4J271. Imagen [88].

4.6 Proteínas repetitivas

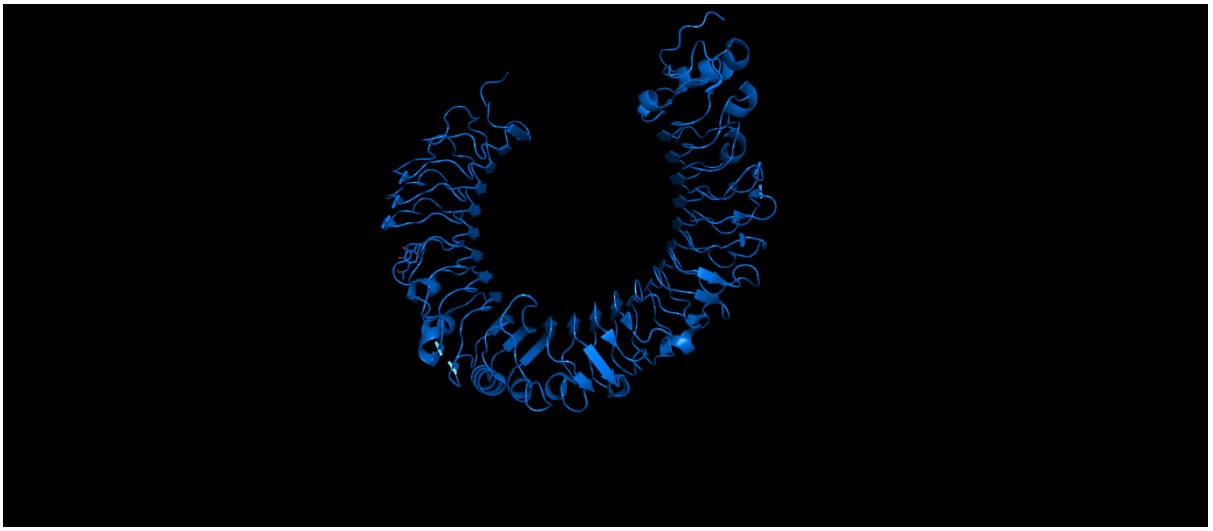
Las proteínas repetitivas se caracterizan por tener duplicaciones internas en su secuencia, generando motivos repetitivos, y son de interés biológico por su relevancia en procesos como el desarrollo neuronal y la salud (Jorda and Kajava 2010, de Wit et al. 2011, Kajava and Steven 2006). Estas repeticiones internas, que pueden ser detectadas secuencialmente, pueden variar en rango desde los 5 hasta los 50 aminoácidos en longitud,

formando estructuras del estilo “solenoide”, son conocidos como los dominios de plegamiento autónomo (sin la necesidad de interactuar con proteínas como las chaperonas) más grandes (Bateman et al. 1998).

Para la búsqueda de proteínas repetitivas en secuencias humanas utilizamos la base de datos RepeatsDB (Di Domenico et al. 2014), que implementa un método de búsqueda combinado, primero utilizando el software RAPHAEL (Walsh et al. 2012), que utiliza análisis geométricos sobre las estructuras para analizar si contienen o no repeticiones, y posteriormente un paso de curación humana para verificar que exista la estructura.
En total, de todas las proteínas humanas, un total de **300** están caracterizadas como repetitivas: **270** provienen de MobiDB, y **30** de la búsqueda por BLAST.
En la figura [89] podemos estudiar su longitud y el coverage de las estructuras asignadas. Se puede observar que tienden a ser proteínas largas, con longitudes que superan al promedio, y que las estructuras asignadas tienden, en general, a cubrir un gran porcentaje de la secuencia:



Distribución de longitudes y coverages, proteínas nudo. [89].

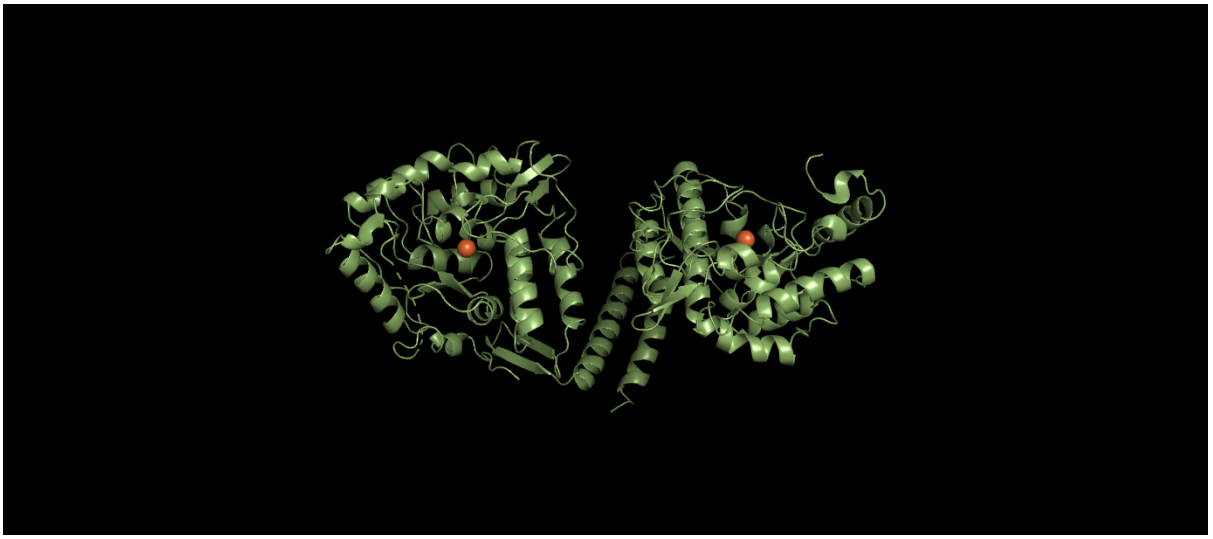


Estructura de 4G8A, cadena B, asignada a la proteína O00206 (839 Aa de longitud), figura [90].

4.7 Proteínas con intercambio de dominios

El intercambio de dominios (3D Domain Swapping) es una capacidad de ciertas proteínas, que forman un dímero u oligómero por medio del intercambio de un elemento estructural único, de modo que varías de sus interacciones monoméricas son reemplazadas por interacciones entre cadenas de un oligómero (Liu and Eisenberg 2002). Este fenómeno fue descubierto inicialmente en la toxina de la Diphtheria (DT, Bennett et al. 1994), y fue propuesto como el responsable de la evolución de proteínas desde monómeros hasta conformaciones oligoméricas (Bennett et al. 1994).

Para la identificación de estas estructuras utilizamos la base de datos 3DSwap (Shameer et al. 2011). Esta base de datos usa un método de predicción que combina curación humana (búsqueda manual de la evidencia en otras publicaciones de que cierta proteína tenga intercambio de dominios) y análisis estructural de dominios.
Para el dataset humano, encontramos un total de **43** estructuras con intercambio de dominios, **33** proviniendo de MobiDB y **10** del dataset asignado vía BLAST. Los números son relativamente bajos, pero esto puede deberse al bajo número de proteínas disponibles en la base de datos (293).



Estructura de P00439 (2PAH, cadena A), [91].

4.8 Proteínas amieloidogénicas

Los amiloides son estructuras caracterizadas por un plegamiento fibrilar y extremadamente estable formado por hojas beta plegadas enfrentadas entre sí, a partir de un proceso de nucleación y polimerización conocido, y, teóricamente, son el plegamiento más estable que cualquier proteína puede adoptar (Nelson et al. 2005, Sawaya et al. 2007). Estos plegamientos son de gran interés para muchos campos de la biología y la salud debido a que están involucrados (de manera protagónica) en distintos desórdenes degenerativos humanos, como la diabetes tipo II, la enfermedad de Huntington y el Alzheimer (Mukherjee et al. 2015, Lotz and Legleiter 2013, Knowles et al. 2014). Sumado a esto, muchas proteínas pueden adoptar este plegamiento de forma funcional, a modo de empaquetarse y almacenarse efectivamente dentro de la célula (Fowler et al. 2007, Maji et al. 2009).

Para poder encontrar plegamientos amiloides en el proteoma recurrimos a la base de datos AmyPro (Varadi et al. 2018), debido a que contiene información de amiloides con base experimental. En total encontramos **73** proteínas con evidencia de formar amiloides, de las cuales **53** provenientes de Mobidb y **6** del dataset de BLAST tienen al menos una estructura asignada, y las restantes **14** a pesar de tener evidencia experimental de formarlos, no tienen ninguna estructura resuelta u homóloga asignada.

Estructura amiloide formado por 2OCT (cadenas A y B), pertenecientes a la estructura de P04080 (Cystain-B). Su mutante es responsable de la Epilepsia Mioclónica, [92].

Al unísono de este trabajo de tesis, las proteínas amiloides humanas fueron un tema de estudio muy trabajado por nuestro grupo. Partiendo desde su asignación estructural, y estudio de su desorden y flexibilidad, pudimos encontrar que son proteínas muy interesantes desde un punto de vista evolutivo debido a que son proteínas con un muy alto grado de expresión, y contra intuitivamente, son de las proteínas que evolucionan más rápido de todo el proteoma (la velocidad de evolución fue medida por medio del análisis de sustituciones no-sinónimas a nivel genómico, a partir de un solo árbol filogenético conformado por 7 especies). Esto es un comportamiento opuesto al establecido por distintos autores: las proteínas que se expresan mucho evolucionan marcadamente más lento que las proteínas que se expresan poco (Drummond et al. 2005).
A pesar de que es un tema en estudio, nuestra hipótesis al momento está dirigida por la idea de que los amiloides están enriquecidos en interacciones con chaperonas (Chiti and Dobson 2017, Killian et al. 2019, Wentink et al. 2020), y las proteínas clientes de estas son proteínas con altos rates evolutivos (Bogumil and Dagan 2012), debido a que estas, mediante su interacción, “empujan” a las proteínas a sus plegamientos funcionales. Esto tiene un efecto contraproducente: les da la libertad de sufrir mutaciones, ya que la interacción permite (hasta un punto) a las proteínas a adoptar aún así los plegamientos necesarios. Este círculo vicioso eventualmente se rompe cuando las proteínas capaces de formar amiloides sufren las suficientes mutaciones para convertirse en plegamientos patológicos.

4.9 Clasificación acorde a CATH

CATH es una base de datos de dominios estructurales que divide a las proteínas según una jerarquía dividida en: Clase, Arquitectura, Topología y Homologas (por sus siglas en inglés, Class, Architecture, Topology and Homologous superfamily, CA Orengo et al, 1997). Fue desarrollada en el año 1997, cuando el número de estructuras de proteínas conocidas era de alrededor de 5000. Actualmente, CATH integra un total de 151 millones de dominios proteicos, y los divide entre 5 grandes clases (la quinta clase fue agregada en 2020):

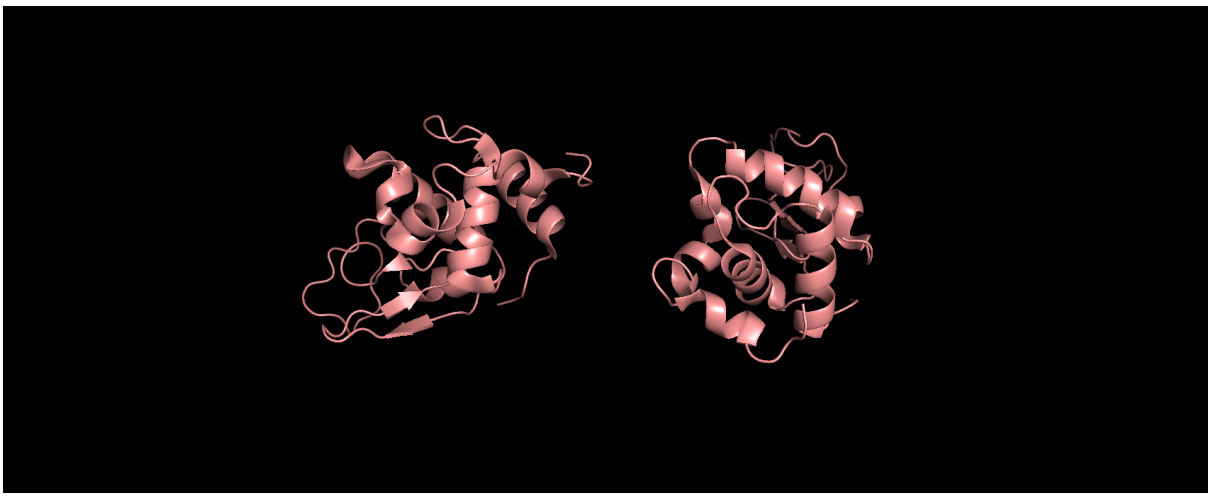
Mayoritariamente Alfa o 1 (que integra 5 arquitecturas, 404 plegamientos, 2003 superfamilias y 103788 dominios), Mayoritariamente Beta o 2 (contiene 21 arquitecturas, 244 plegamientos, 1290 superfamilias y 124032 dominios), Alfa Beta o 3 (contiene 14 arquitecturas, 634 plegamientos, 2337 superfamilias y 262275 dominios), con pocas estructuras secundarias o 4 (1 arquitectura, 108 plegamientos, 181 superfamilias y 5716 dominios) y por último especiales o 6 (que integra 2 arquitecturas, 82 plegamientos, 790 superfamilias y 4427 dominios) (Dawson et al. 2017, Sillitoe et al. 2019,Sillitoe et al. 2021).

Estas divisiones de Clase dividen a las proteínas de forma general, acorde a su composición de estructura secundaria y contactos, y se basan en alineamientos secuenciales y estructurales. Las tres primeras incluyen plegamientos globulares, y los dos últimos incluyen proteínas con plegamientos más desordenados. A la hora de buscar proteínas con estos plegamientos en nuestro dataset, es importante tener en cuenta que existe solapamiento entre varias categorías (muchas proteínas pueden estar anotadas como Mayoritariamente Alfa y Alfa Beta), y también puede que algunas proteínas que estén clasificadas bajo alguna estructura ya estudiada (desordenada, rígida, etc) sea también clasificada dentro de CATH.

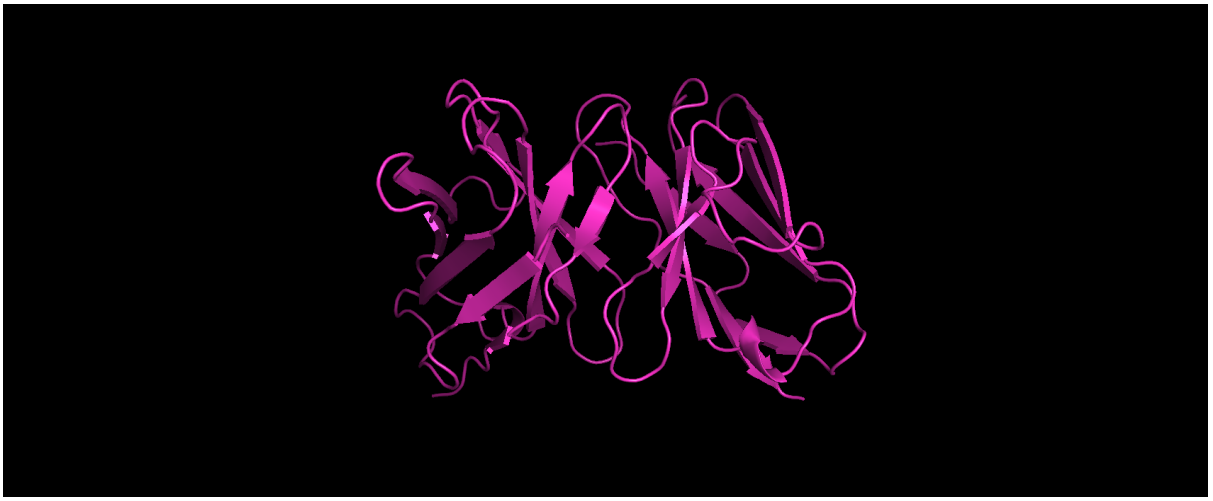
Para el proteoma humano, un total de **6773** proteínas tienen al menos una estructura asignada a alguna de las 5 clases de CATH (un 32,88% total del proteoma). Del total, **4283** pertenecen al grupo de MobiDB, **2484** al de BLAST, y **6** al de HHblits.

De las proteínas que provienen de MobiDB, hay **1939** estructuras anotadas bajo la clase 1, **1734** en la clase 2, **2701** en la clase 3, **154** en la clase 4 y **262** en la clase 6 (la suma excede al número total de proteínas debido a que varias proteínas están asignadas a más de una clase).
Con respecto a las estructuras asignadas por homología vía BLAST, **836** pertenecen al grupo 1, **889** al grupo 2, **1500** al grupo 3, **74** al grupo 4 y **52** al grupo 6.
Por último, de las estructuras asignadas mediante el uso del HHblits, **5** proteínas pertenecen al grupo 2 y **4** pertenecen al grupo 4 (de las cuales 2 también pertenecían al grupo 2).

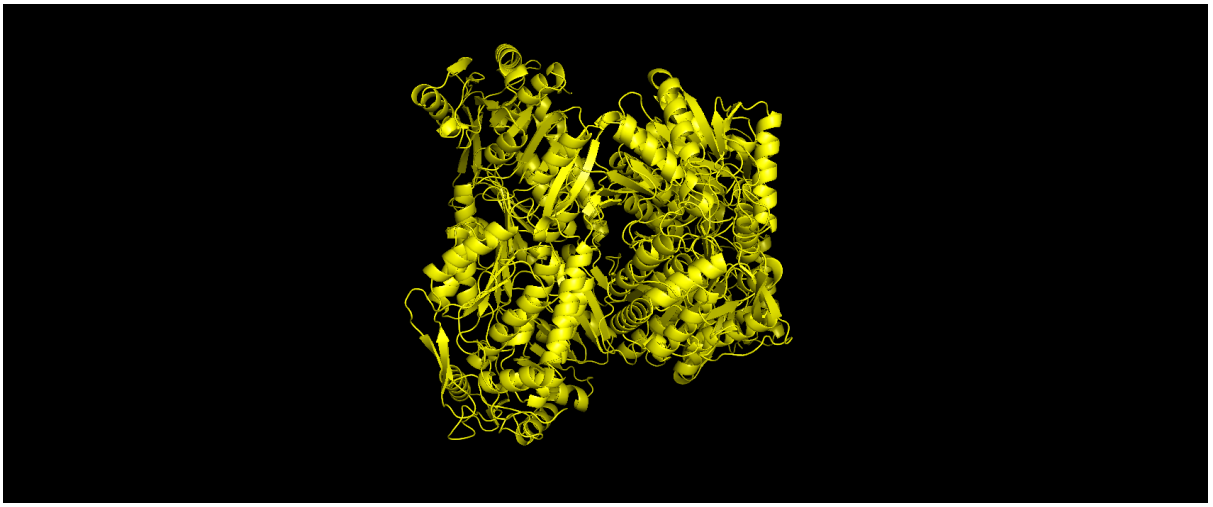
No debería ser llamativo que la asignación de estructuras por CATH haya cubierto un porcentaje mucho más alto que el resto de estructuras. Parte de la idea al crear la base de datos fue generar una forma fácil y práctica de anotar proteínas, y en el momento en el que se creó, la gran mayoría de los plegamientos conocidos eran globulares (sin ir más lejos, varios de las estructuras estudiadas anteriormente en este capítulo se descubrieron después de que se haya creado CATH).



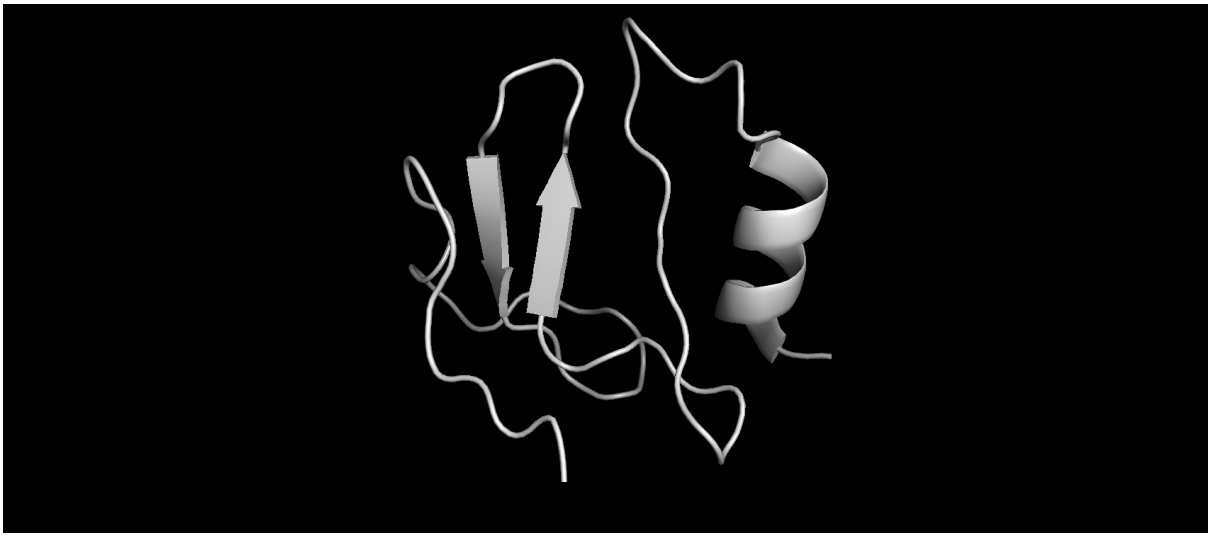
Estructura (mayoritariamente Alfa) de 1GHL, ambas cadenas, asignada a O75951. [93]



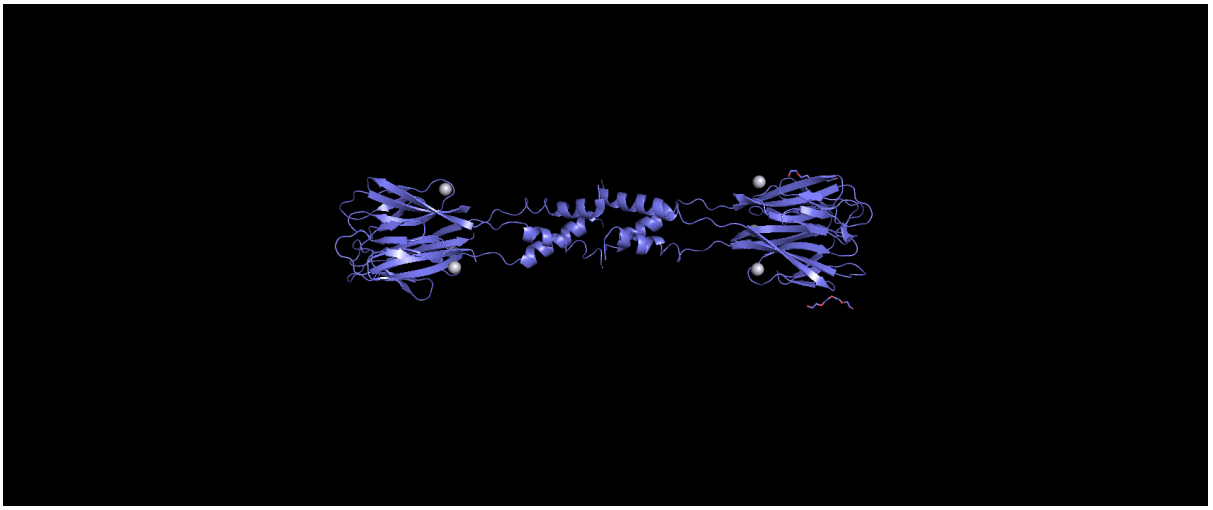
Estructura (mayoritariamente Beta) de 1A70, ambas cadenas, asignada a P09564. [94]



Estructura (Alfa Beta) de 1AG8, ambas cadenas, asignada a P30837. [95]



Estructura (poca estructura secundarias) de 1L3H, ambas cadenas, perteneciente a P04233. [96]

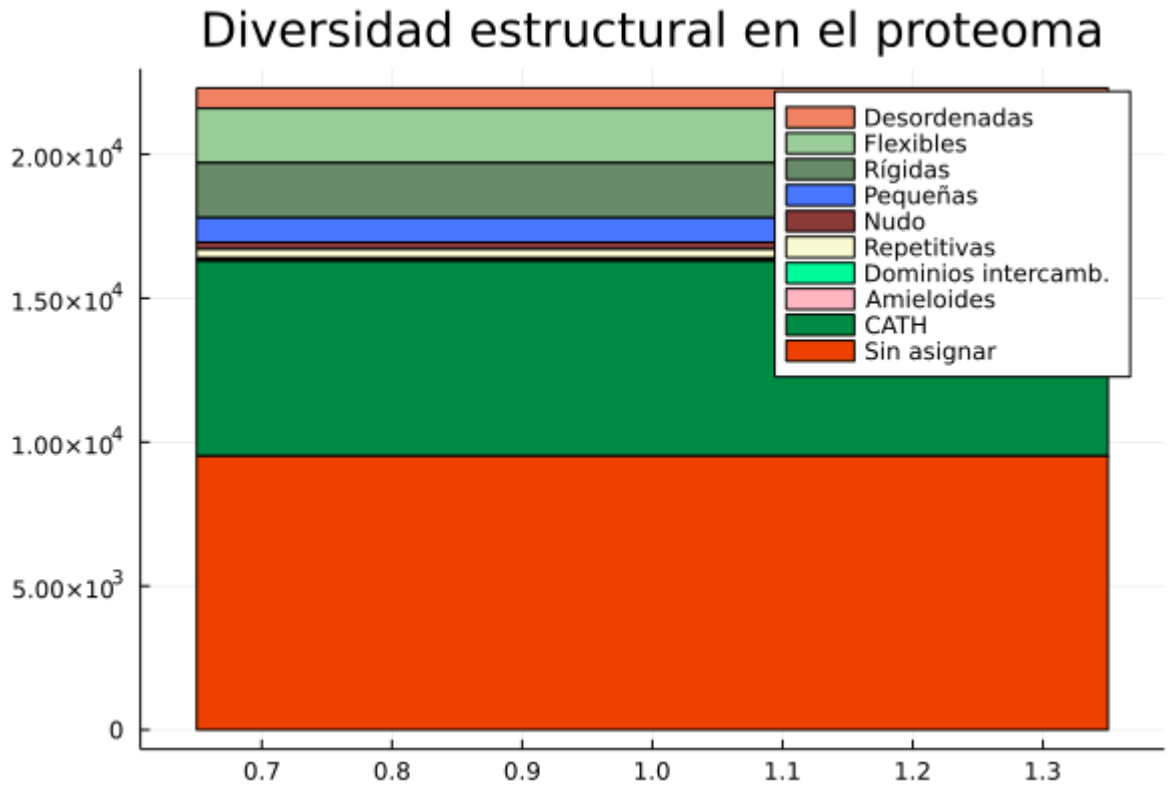


Estructura (especial) de 1RH7, ambas cadenas, asignada a Q9BQ08. [97]

Conclusiones generales del trabajo:

Para finalizar este trabajo podemos evaluar nuestros resultados globales, y en base a esto, generar perspectivas para futuros análisis e investigaciones:

Partiendo de las 20600 proteínas por gen con las que comenzamos, primero las caracterizamos efectivamente al proteoma en términos generales (longitud, expresión y abundancia, desorden y flexibilidad). Luego buscamos estructuras para estas distintas proteínas (antes de analizar las estructuras que ya estaban asignadas a algunas de estas) y las caracterizamos. Realizamos una asignación a más de la mitad del proteoma, llegando al 65,08% total de proteínas con estructuras, por medio del uso de dos métodos de búsqueda por homología, BLAST y HHblits.
Por último, realizamos una búsqueda en distintas bases de datos para categorizar estas estructuras (o proteínas usando sus características secuenciales). En total pudimos categorizar estructuralmente a 13406 proteínas, llegando a un total del 65,08% del proteoma, (11085 proteínas fueron categorizadas en el capítulo 4, un 53,81%. El número es distinto debido a que algunas no requieren de una estructura asignada para ser categorizada).



Distribución de la diversidad estructural en el proteoma humano, figura [98.1]

Proteínas con estructuras clasificadas / sin estructura o sin asignar [98.2].
Misma distribución, distinto gráfico.

De esta manera, hasta cierto grado, hemos demostrado la gran diversidad presente en el proteoma humano. Sin embargo muchas proteínas continúan estando parcialmente mapeadas o sin ninguna estructura asignada. Una posible solución para este problema es el uso de AlphaFold (AlQuraishi 2019, Jumper et al. 2021). Este algoritmo desarrollado por DeepMind permite la predicción de estructuras (5 o 10 modelos por proteína) a partir sólo del uso de secuencias. AlphaFold liberó una versión del estructuroma humano el 22 de Julio de 2021 (cuando este trabajo ya estaba empezando a finalizarse) que podría usarse de forma suplementaria a la nuestra asignación por homología, pudiendo hacer un análisis entre los modelos predichos y los modelos obtenidos (Tunyasuvunakool et al. 2021). Sumado a esto, es de nuestro interés caracterizar las estructuras predichas por AlphaFold con respecto a la librería de estructuras que generamos en este trabajo. Algunos resultados preliminares nos permitieron observar que el algoritmo no necesariamente favorece la predicción de estructuras Apo/Holo, si no que tiende a tener un bias hacia alguno de los dos modelos. También encontramos que al aumentar la diversidad conformacional de las proteínas, este también parece generar modelos menos confiables. Debido a esto, a pesar de que AlphaFold es una herramienta nueva e invaluable para el campo, es necesario realizar este tipo de estudios para calibrar y evaluar su verdadero valor a la hora de predecir estructuras de forma masiva.

Por otro lado, también hay que remarcar otro aspecto que hay que tener en cuenta hacia el futuro: muchas de las bases de datos estructurales (no solo las que usamos en este trabajo) se encuentran desactualizadas. Es decir, es muy probable que estemos subrepresentado a muchas estructuras dentro de nuestra clasificación, y también es posible que haya tipos estructuras que no hayamos estudiado debido a la falta de información

actual acerca de ellos. Esto nos permite proponer un análisis para el futuro: la solución para este problema es, en lugar de buscar en distintas bases de datos, usar predictores para identificar las distintas estructuras dentro de nuestro dataset. Esto es un objetivo de gran tamaño, no solo por el gran volumen de secuencias a analizar, si no porque también deberíamos generar consensos entre los distintos predictores usados para una misma categoría, análogamente a como MobiDB generó un consenso para la predicción del desorden.

Por último, es necesario no perder de vista uno de los puntos más importantes de esta tesis: citando de nuevo a Ernest Rutherford, “toda la ciencia se puede dividir entre física o coleccionar estampitas.” A pesar de que esta frase no es necesariamente aplicable a nuestro contexto actual (Rutherford vivió desde 1871 hasta 1937), tiene una importancia muy grande en el contexto de este trabajo. No tiene ningúna utilidad práctica generar una caracterización de la diversidad de estructura - función del proteoma humano sin seguir investigando y caracterizando los datos obtenidos. Con este trabajo hemos, en gran parte, formado el primer escalón para una nueva línea de investigación. Nos hemos familiarizado y realizado estudios preliminares sobre nuestro objeto de estudio, el proteoma humano.

Pero ahora es momento de dar el siguiente paso, y no solo mejorar y curar aún más los datos estructurales obtenidos, si no también ampliar estos, para poder generar un análisis más contundente. Como se habló brevemente en la sección 4.8, al unísono que se trabajaba en esta tesis, con el grupo de trabajo realizamos investigaciónes exhaustivas a nivel evolutivo sobre un tipo de estructuras particulares presentes en el proteoma.
Es de alto interés generar los mismos análisis para el resto de las proteínas, aumentando el conocimiento en el campo y comprobando a niveles más altos la idea de que el concepto de “las proteínas” incluye a uno de los grupos más diversos y complejos de la biología molecular.

Anexo I

1. Principales bases de datos

- **Uniprot:** (The Universal Protein Resource): es un recurso integral para el manejo de secuencias proteicas y anotación de datos accesorios. Las bases de datos que integran Uniprot son UniProt Knowledgebase (UniProtKB), UniProt Reference Clusters (UniRef) y UniProt Archive (UniParc). UniProt es un esfuerzo de colaboración entre European Bioinformatics Institute (EMBL-EBI), SIB Swiss Institute of Bioinformatics y Protein Information Resource (PIR).
  URL: http://www.uniprot.org/

- **PDB:** (Protein Data Bank): es una base de datos que contiene información de la estructura tridimensional de proteínas y ácidos nucleicos. El RCSB (Research Collaboratory for Structural Bioinformatics) es el ente encargado del curado y la anotación de las estructuras.
  URL: http://www.rcsb.org/pdb/home/home.do

- **MOBIDB:** Es una base de datos generada por investigadores de Italia (Universidad de Padua) y Argentina (Universidad Nacional de Quilmes) en el 2012 (cuentan con actualizaciones recientes) en la cual se recluta información pertinente a la movilidad y desorden anotado de proteínas.
  URL: https://mobidb.bio.unipd.it/

2. Herramientas de programación

- **Python:** Es un lenguaje de programación creado por Guido van Rossum en el Centro para las Matemáticas y la Informática (CWI) en los Países Bajos. Fue generado a finales de los ochenta, pero su versión 0.9.0 fue publicada en 1991 en *alt.sources* como sucesor del lenguaje de programación ABC.
  Se caracteriza por hacer hincapié en la legibilidad de su código, siendo un lenguaje dinámico y multiplataforma.

- **BASH:** (Bourne again shell): Es un programa informático cuya función consiste en interpretar órdenes. Está basado en la Shell de Unix y es compatible con POSIX.

- **R:** Es un lenguaje y un entorno para computación estadística y gráficos. R ofrece una amplia variedad de estadísticas (modelado lineal y no lineal, pruebas estadísticas clásicas, análisis de series temporales, clasificación, agrupación) y las técnicas gráficas, y es altamente extensible. Uno de los puntos fuertes de R es la facilidad con la que pueden generarse gráficos bien diseñados y de alta calidad para

publicaciones, incluidos símbolos matemáticos y formulaciones donde sean necesarios.

- **Julia:** Julia es un lenguaje de programación dinámico de alto nivel y velocidad, desarrollado por Jeff Bezanson, Stefan Karpinski, Viral B. Shah y Alan Edelman en 2012. Cumple el rol de ser un lenguaje de programación accesible y rápido de aprender (como R o Python), pero también tiene un alto desempeño y velocidad de ejecución (donde R y Python fallan). Es un lenguaje homoicónico, es decir, el código de Julia está escrito en Julia, y es muy sencillo interiorizarse en su ambiente. Actualmente representa una opción muy viable a la hora de necesitar manejar grandes volúmenes de datos, debido a su velocidad, accesibilidad y rendimiento.

Anexo II

Anexo II: Listado de proteínas humanas de re…

Bibliografía

1. , Rohdein der Carminsaure angenommenen Molekeln Wasser noch eine dritte enthalten ist. Die Substanz wurde fiir die Analyse im Kohlensaurestrom bei.

2. A. E. Mirsky, L. Pauling, On the structure of native, denatured, and coagulated proteins. *Proc Natl Acad Sci USA* **22**, 439–447 (1936).

3. L. Pauling, A theory of the structure and process of formation of antibodies*. *J. Am. Chem. Soc.* **62**, 2643–2657 (1940).

4. TABLE, June, 1950 HETEROGENEITY OF BINDING SITES OF BOVINE.

5. J. Monod, J. P. Changeux, F. Jacob, Allosteric proteins and cellular control systems. *J. Mol. Biol.* **6**, 306–329 (1963).

6. J.-P. Changeux, Allostery and the Monod-Wyman-Changeux model after 50 years. *Annu. Rev. Biophys.* **41**, 103–133 (2012).

7. H. N. Motlagh, J. O. Wrabl, J. Li, V. J. Hilser, The ensemble nature of allostery. *Nature* **508**, 331–339 (2014).

8. R. Nussinov, Introduction to protein ensembles and allostery. *Chem. Rev.* **116**, 6263–6266 (2016).

9. G.-W. Wei, Protein structure prediction beyond AlphaFold. *Nat. Mach. Intell.* **1**, 336–337 (2019).

10. M. AlQuraishi, AlphaFold at CASP13. *Bioinformatics* **35**, 4862–4865 (2019).

11. J. N. Onuchic, N. D. Socci, Z. Luthey-Schulten, P. G. Wolynes, Protein folding funnels: the nature of the transition state ensemble. *Fold. Des.* **1**, 441–450 (1996).

12. C. A. Orengo, A. E. Todd, J. M. Thornton, From protein structure to function. *Curr. Opin. Struct. Biol.* **9**, 374–382 (1999).

13. A. M. Monzon, *et al.*, Conformational diversity analysis reveals three functional mechanisms in proteins. *PLoS Comput. Biol.* **13**, e1005398 (2017).

14. G. Parisi, N. Palopoli, S. C. E. Tosatto, M. S. Fornasari, P. Tompa, “Protein” no longer means what it used to. *Current Research in Structural Biology* **3**, 146–152 (2021).

15. M. Gerstein, A. M. Lesk, C. Chothia, Structural mechanisms for domain movements in proteins. *Biochemistry* **33**, 6739–6749 (1994).

16. L. C. James, P. Roversi, D. S. Tawfik, Antibody multispecificity mediated by conformational diversity. *Science* **299**, 1362–1367 (2003).

17. I. Kufareva, R. Abagyan, Methods of protein structure comparison. *Methods Mol. Biol.* **857**, 231–257 (2012).

18. P. V. Burra, Y. Zhang, A. Godzik, B. Stec, Global distribution of conformational states derived from redundant models in the PDB points to non-uniqueness of the protein structure. *Proc Natl Acad Sci USA* **106**, 10505–10510 (2009).

19. D. Piovesan, *et al.*, DisProt 7.0: a major update of the database of disordered

proteins. *Nucleic Acids Res.* **45**, D219–D227 (2017).

20. , On the richness of the red globules in hemoglobine. *Edinb. Med. J.* **23**, 757–758 (1878).

21. R. Origa, β-Thalassemia. *Genet. Med.* **19**, 609–619 (2017).

22. L. R. Manning, *et al.*, Energetic differences at the subunit interfaces of normal human hemoglobins correlate with their developmental profile. *Biochemistry* **48**, 7568–7574 (2009).

23. L. R. Manning, A. M. Popowicz, J. C. Padovan, B. T. Chait, J. M. Manning, Gel filtration of dilute human embryonic hemoglobins reveals basis for their increased oxygen binding. *Anal. Biochem.* **519**, 38–41 (2017).

24. W. Chen, *et al.*, Transposing sequences between fetal and adult hemoglobins indicates which subunits and regulatory molecule interfaces are functionally related. *Biochemistry* **39**, 3774–3781 (2000).

25. U. Hoeger, J. R. Harris, Eds., *Vertebrate and Invertebrate Respiratory Proteins, Lipoproteins and other Body Fluid Proteins* (Springer International Publishing, 2020).

26. J. Shendure, *et al.*, DNA sequencing at 40: past, present and future. *Nature* **550**, 345–353 (2017).

27. E. Papalexi, R. Satija, Single-cell RNA sequencing to explore immune cell heterogeneity. *Nat. Rev. Immunol.* **18**, 35–45 (2018).

28. B. Hwang, J. H. Lee, D. Bang, Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp. Mol. Med.* **50**, 1–14 (2018).

29. C. Scheckel, A. Aguzzi, Prions, prionoids and protein misfolding disorders. *Nat. Rev. Genet.* **19**, 405–418 (2018).

30. N. Palopoli, *et al.*, Intrinsically disordered protein ensembles shape evolutionary rates revealing conformational patterns. *J. Mol. Biol.* **433**, 166751 (2021).

31. , Ernest Rutherford - Wikiquote (September 20, 2021).

32. UniProt Consortium, The universal protein resource (UniProt). *Nucleic Acids Res.* **36**, D190-5 (2008).

33. B. Modrek, C. Lee, A genomic view of alternative splicing. *Nat. Genet.* **30**, 13–19 (2002).

34. L. Brocchieri, S. Karlin, Protein length in eukaryotic and prokaryotic proteomes. *Nucleic Acids Res.* **33**, 3390–3400 (2005).

35. D. J. Lipman, A. Souvorov, E. V. Koonin, A. R. Panchenko, T. A. Tatusova, The relationship of protein conservation and sequence length. *BMC Evol. Biol.* **2**, 20 (2002).

36. G. Storz, Y. I. Wolf, K. S. Ramamurthi, Small proteins can no longer be ignored. *Annu. Rev. Biochem.* **83**, 753–777 (2014).

37. J. Zhang, Protein-length distributions for the three domains of life. *Trends Genet.* **16**, 107–109 (2000).

38. J. Surkont, Y. Diekmann, P. V. Ryder, J. B. Pereira-Leal, Coiled-coil length: Size does matter. *Proteins* **83**, 2162–2169 (2015).

39. D. Greenbaum, C. Colangelo, K. Williams, M. Gerstein, Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biol.* **4**, 117 (2003).

40. A. Panda, D. Acharya, T. Chandra Ghosh, Insights into human intrinsically disordered proteins from their gene expression profile. *Mol. Biosyst.* **13**, 2521–2530 (2017).

41. M. Wang, C. J. Herrmann, M. Simonovic, D. Szklarczyk, C. von Mering, Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines. *Proteomics* **15**, 3163–3168 (2015).

42. A. O. Urrutia, L. D. Hurst, The signature of selection mediated by expression on human genes. *Genome Res.* **13**, 2260–2264 (2003).

43. A. K. Dunker, *et al.*, Intrinsically disordered protein. *J. Mol. Graph. Model.* **19**, 26–59 (2001).

44. R. W. Kriwacki, L. Hengst, L. Tennant, S. I. Reed, P. E. Wright, Structural studies of p21Waf1/Cip1/Sdi1 in the free and Cdk2-bound state: conformational disorder mediates binding diversity. *Proc Natl Acad Sci USA* **93**, 11504–11509 (1996).

45. W. Bode, P. Schwager, R. Huber, The transition of bovine trypsinogen to a trypsin-like state upon strong ligand binding. *J. Mol. Biol.* **118**, 99–112 (1978).

46. A. C. Bloomer, J. N. Champness, G. Bricogne, R. Staden, A. Klug, Protein disk of tobacco mosaic virus at 2.8 A resolution showing the interactions within and between subunits. *Nature* **276**, 362–368 (1978).

47. A. K. Dunker, *et al.*, The unfoldomics decade: an update on intrinsically disordered proteins. *BMC Genomics* **9 Suppl 2**, S1 (2008).

48. E. Schad, P. Tompa, H. Hegyi, The relationship between proteome size, structural disorder and organism complexity. *Genome Biol.* **12**, R120 (2011).

49. V. N. Uversky, *et al.*, Pathological unfoldomics of uncontrolled chaos: intrinsically disordered proteins and human diseases. *Chem. Rev.* **114**, 6844–6879 (2014).

50. M. Necci, D. Piovesan, Z. Dosztányi, S. C. E. Tosatto, MobiDB-lite: fast and highly specific consensus prediction of intrinsic disorder in proteins. *Bioinformatics* **33**, 1402–1404 (2017).

51. E. Cilia, R. Pancsa, P. Tompa, T. Lenaerts, W. F. Vranken, From protein sequence to dynamics and disorder with DynaMine. *Nat. Commun.* **4**, 2741 (2013).

52. M. V. Berjanskii, D. S. Wishart, Application of the random coil index to studying protein flexibility. *J. Biomol. NMR* **40**, 31–48 (2008).

53. H. Hebert, CryoEM: a crystals to single particles round-trip. *Curr. Opin. Struct. Biol.* **58**, 59–67 (2019).

54. J. M. Dana, *et al.*, SIFTS: updated Structure Integration with Function, Taxonomy and Sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins. *Nucleic Acids Res.* **47**, D482–D489 (2019).

55. S. Velankar, *et al.*, SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. *Nucleic Acids Res.* **41**, D483-9 (2013).

56. D. Piovesan, *et al.*, MobiDB: intrinsically disordered proteins in 2021. *Nucleic Acids Res.* **49**, D361–D367 (2021).

57. T. Di Domenico, I. Walsh, A. J. M. Martin, S. C. E. Tosatto, MobiDB: a comprehensive database of intrinsic protein disorder annotations. *Bioinformatics* **28**, 2080–2081 (2012).

58. A. M. Monzon, E. Juritz, M. S. Fornasari, G. Parisi, CoDNaS: a database of conformational diversity in the native state of proteins. *Bioinformatics* **29**, 2512–2514 (2013).

59. A. M. Monzon, M. S. Fornasari, D. J. Zea, G. Parisi, Exploring protein conformational diversity. *Methods Mol. Biol.* **1851**, 353–365 (2019).

60. A. M. Monzon, C. O. Rohr, M. S. Fornasari, G. Parisi, CoDNaS 2.0: a comprehensive database of protein conformational diversity in the native state. *Database (Oxford)* **2016** (2016).

61. C. Marino-Buslje, A. M. Monzon, D. J. Zea, M. S. Fornasari, G. Parisi, On the dynamical incompleteness of the Protein Data Bank. *Brief. Bioinformatics* **20**, 356–359 (2019).

62. D. A. Keen, A. L. Goodwin, The crystallography of correlated disorder. *Nature* **521**, 303–309 (2015).

63. K. Djinovic-Carugo, O. Carugo, Missing strings of residues in protein crystal structures. *Intrinsically Disord. Proteins* **3**, e1095697 (2015).

64. C. Chothia, A. M. Lesk, The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5**, 823–826 (1986).

65. S. Henikoff, J. G. Henikoff, Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA* **89**, 10915–10919 (1992).

66. S. tschuP, W. sh, W. Myers, D. Lipman, Basic Local Al ignment Search Tool.

67. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

68. S. F. Altschul, *et al.*, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).

69. M. Remmert, A. Biegert, A. Hauser, J. Söding, HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* **9**, 173–175 (2011).

70. S. R. Eddy, What is a hidden Markov model? *Nat. Biotechnol.* **22**, 1315–1316 (2004).

71. I. Sillitoe, *et al.*, CATH: increased structural coverage of functional space. *Nucleic Acids Res.* **49**, D266–D273 (2021).

72. M. Necci, D. Piovesan, D. Clementel, Z. Dosztányi, S. C. E. Tosatto, MobiDB-lite 3.0: fast consensus annotation of intrinsic disorder flavours in proteins. *Bioinformatics* (2020) https:/doi.org/10.1093/bioinformatics/btaa1045.

73. I. Walsh, A. J. M. Martin, T. Di Domenico, S. C. E. Tosatto, ESpritz: accurate and fast prediction of protein disorder. *Bioinformatics* **28**, 503–509 (2012).

74. Z. Dosztányi, Prediction of protein disorder based on IUPred. *Protein Sci.* **27**, 331–340 (2018).

75. A. Katuwawala, L. Kurgan, Comparative Assessment of Intrinsic Disorder Predictions with a Focus on Protein and Nucleic Acid-Binding Proteins. *Biomolecules* **10** (2020).

76. R. Linding, *et al.*, Protein disorder prediction: implications for structural proteomics. *Structure* **11**, 1453–1459 (2003).

77. R. Linding, R. B. Russell, V. Neduva, T. J. Gibson, GlobPlot: Exploring protein sequences for globularity and disorder. *Nucleic Acids Res.* **31**, 3701–3708 (2003).

78. Z. R. Yang, R. Thomson, P. McNeil, R. M. Esnouf, RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics* **21**, 3369–3376 (2005).

79. L. K. Muranova, A. S. Ryzhavskaya, M. V. Sudnitsyna, V. M. Shatov, N. B. Gusev, Small heat shock proteins and human neurodegenerative diseases. *Biochemistry Mosc* **84**, 1256–1267 (2019).

80. M. P. Collier, J. L. P. Benesch, Small heat-shock proteins and their role in mechanical stress. *Cell Stress Chaperones* **25**, 601–613 (2020).

81. I. Camby, M. Le Mercier, F. Lefranc, R. Kiss, Galectin-1: a small protein with major functions. *Glycobiology* **16**, 137R-157R (2006).

82. M. L. Mansfield, Are there knots in proteins? *Nat. Struct. Mol. Biol.* **1**, 213–214 (1994).

83. N. P. King, E. O. Yeates, T. O. Yeates, Identification of rare slipknots in proteins and their implications for stability and folding. *J. Mol. Biol.* **373**, 153–166 (2007).

84. J. I. Sułkowska, E. J. Rawdon, K. C. Millett, J. N. Onuchic, A. Stasiak, Conservation of complex knotting and slipknotting patterns in proteins. *Proc Natl Acad Sci USA* **109**, E1715-23 (2012).

85. P. Virnau, L. A. Mirny, M. Kardar, Intricate knots in proteins: Function and evolution. *PLoS Comput. Biol.* **2**, e122 (2006).

86. M. C. Prentiss, D. J. Wales, P. G. Wolynes, The energy landscape, folding pathways and the kinetics of a knotted protein. *PLoS Comput. Biol.* **6**, e1000835 (2010).

87. J. I. Sułkowska, P. Sułkowski, J. Onuchic, Dodging the crisis of folding proteins with knots. *Proc Natl Acad Sci USA* **106**, 3119–3124 (2009).

88. M. Jamroz, *et al.*, KnotProt: a database of proteins with knots and slipknots. *Nucleic Acids Res.* **43**, D306-14 (2015).

89. J. Jorda, A. V. Kajava, “Protein Homorepeats” in Advances in protein chemistry and structural biology., (Elsevier, 2010), pp. 59–88.

90. J. de Wit, W. Hong, L. Luo, A. Ghosh, Role of leucine-rich repeat proteins in the development and function of neural circuits. *Annu. Rev. Cell Dev. Biol.* **27**, 697–729

(2011).

91. A. V. Kajava, A. C. Steven, Beta-rolls, beta-helices, and other beta-solenoid proteins. *Adv. Protein Chem.* **73**, 55–96 (2006).

92. A. Bateman, A. G. Murzin, S. A. Teichmann, Structure and distribution of pentapeptide repeats in bacteria. *Protein Sci.* **7**, 1477–1480 (1998).

93. T. Di Domenico, *et al.*, RepeatsDB: a database of tandem repeat protein structures. *Nucleic Acids Res.* **42**, D352-7 (2014).

94. I. Walsh, *et al.*, RAPHAEL: recognition, periodicity and insertion assignment of solenoid protein structures. *Bioinformatics* **28**, 3257–3264 (2012).

95. Y. Liu, D. Eisenberg, 3D domain swapping: as domains continue to swap. *Protein Sci.* **11**, 1285–1299 (2002).

96. M. J. Bennett, S. Choe, D. Eisenberg, Refined structure of dimeric diphtheria toxin at 2.0 A resolution. *Protein Sci.* **3**, 1444–1463 (1994).

97. M. J. Bennett, S. Choe, D. Eisenberg, Domain swapping: entangling alliances between proteins. *Proc Natl Acad Sci USA* **91**, 3127–3131 (1994).

98. K. Shameer, *et al.*, 3DSwap: curated knowledgebase of proteins involved in 3D domain swapping. *Database (Oxford)* **2011**, bar042 (2011).

99. R. Nelson, *et al.*, Structure of the cross-beta spine of amyloid-like fibrils. *Nature* **435**, 773–778 (2005).

100. M. R. Sawaya, *et al.*, Atomic structures of amyloid cross-beta spines reveal varied steric zippers. *Nature* **447**, 453–457 (2007).

101. A. Mukherjee, D. Morales-Scheihing, P. C. Butler, C. Soto, Type 2 diabetes as a protein misfolding disease. *Trends Mol. Med.* **21**, 439–449 (2015).

102. G. P. Lotz, J. Legleiter, The role of amyloidogenic protein oligomerization in neurodegenerative disease. *J. Mol. Med.* **91**, 653–664 (2013).

103. T. P. J. Knowles, M. Vendruscolo, C. M. Dobson, The amyloid state and its association with protein misfolding diseases. *Nat. Rev. Mol. Cell Biol.* **15**, 384–396 (2014).

104. D. M. Fowler, A. V. Koulov, W. E. Balch, J. W. Kelly, Functional amyloid--from bacteria to humans. *Trends Biochem. Sci.* **32**, 217–224 (2007).

105. S. K. Maji, *et al.*, Functional amyloids as natural storage of peptide hormones in pituitary secretory granules. *Science* **325**, 328–332 (2009).

106. M. Varadi, G. De Baets, W. F. Vranken, P. Tompa, R. Pancsa, AmyPro: a database of proteins with validated amyloidogenic regions. *Nucleic Acids Res.* **46**, D387–D392 (2018).

107. D. A. Drummond, J. D. Bloom, C. Adami, C. O. Wilke, F. H. Arnold, Why highly expressed proteins evolve slowly. *Proc Natl Acad Sci USA* **102**, 14338–14343 (2005).

108. F. Chiti, C. M. Dobson, Protein misfolding, amyloid formation, and human disease: A summary of progress over the last decade. *Annu. Rev. Biochem.* **86**, 27–68 (2017).

109. A. N. Killian, S. C. Miller, J. K. Hines, Impact of Amyloid Polymorphism on Prion-Chaperone Interactions in Yeast. *Viruses* **11** (2019).

110. A. S. Wentink, *et al.*, Molecular dissection of amyloid disaggregation by human HSP70. *Nature* **587**, 483–488 (2020).

111. D. Bogumil, T. Dagan, Cumulative impact of chaperone-mediated folding on genome evolution. *Biochemistry* **51**, 9941–9953 (2012).

112. , Sci-Hub | CATH – a hierarchic classification of protein domain structures. Structure, 5(8), 1093–1109 | 10.1016/S0969-2126(97)00260-8 (September 23, 2021).

113. N. L. Dawson, *et al.*, CATH: an expanded resource to predict protein function through structure and sequence. *Nucleic Acids Res.* **45**, D289–D295 (2017).

114. I. Sillitoe, *et al.*, CATH: expanding the horizons of structure-based functional annotations for genome sequences. *Nucleic Acids Res.* **47**, D280–D284 (2019).

115. J. Jumper, *et al.*, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

116. K. Tunyasuvunakool, *et al.*, Highly accurate protein structure prediction for the human proteome. *Nature* **5**