INSTITUTO TECNOLÓGICO DE BUENOS AIRES – ITBA

ESCUELA DE INGENIERÍA Y TECNOLOGÍA

Relación entre la estructura y la dinámica en proteínas

AUTOR:

- ***Castillo, Maximiliano Sebastián*** **(Leg. Nº 55378)**

TUTORES:

- ***Dra. Fernández, María Laura***

UBA. Dep Física, FECyN. UBA-CONICET-Instituto de Física del Plasma (INFIP), ITBA

- ***Dra. Marino Buslje, Cristina***

Instituto de Investigaciones Bioquímicas de Buenos Aires (IIBBA), Fundación Instituto Leloir

TRABAJO FINAL PRESENTADO PARA LA OBTENCIÓN DEL TÍTULO DE BIOINGENIERO

BUENOS AIRES

PRIMER CUATRIMESTRE, 2019

Relación entre la estructura y la dinámica en proteínas

Agradecimientos

Al Instituto Tecnológico de Buenos Aires, por permitirme estudiar esta hermosa carrera.

A la Dra María Laura Fernández y la Dra Cristina Marino Buslje , que me brindaron sus conocimientos y toda su paciencia para poder llevar a cabo este Proyecto Final de Carrera.

A la Fundación Instituto Leloir y a Amazon, que me brindaron, junto con el ITBA, los servidores para que pueda realizar las simulaciones del presente escrito.

Por último, y no por eso menos importante, a mi familia y amigos que me acompañaron desde el instante cero, viviendo conmigo cada momento.

¡Gracias!

Relación entre la estructura y la dinámica en proteínas

Índice

Resumen

Este trabajo tiene como fin “aprender para predecir”: tal como el oráculo daba a los griegos de la antigüedad una idea del futuro, se busca utilizar la bioinformática como herramienta predictiva en biología.

El presente proyecto pretende poder encontrar relaciones entre la secuencia, función, estructura y flexibilidad proteica al combinar la información estructural de proteínas disponibles con la información dinámica obtenida mediante simulaciones de dinámica molecular (DM).

Se pudo establecer que los valores de RMSD (respecto a una estructura de referencia y entre todas las estructuras obtenidas de la simulación), los valores de radio de giro, la cantidad de grupos (*clusters*) identificados y la distribución de factor B entre los aminoácidos de la cadena proteica, nos dan cuenta, a partir de un estado nativo, la existencia de otros estados. Estos resultados nos permite identificar si las proteínas son rígidas o móviles, con posibilidad de discriminar con mayor precisión la separación de las últimas en parcialmente desordenadas y maleables. Este trabajo demuestra la relación entre conformación y dinámica: las simulaciones de DM despliegan las posibles conformaciones de una proteína, mostrando su flexibilidad, aún cuando se conozca un solo confórmero de la misma. Si bien no podemos asegurar que la información dinámica de una proteína está codificada en su secuencia, estos resultados sugieren que es una posibilidad.

Por último, lo enriquecedor de este trabajo es que suscita nuevas preguntas y permite la generación de nuevas hipótesis, cuya resolución nos permitirá ir un paso adelante en el camino infinito del conocimiento.

Introducción

Gran parte de las moléculas biológicas son macromoléculas, polímeros conformados a partir de precursores relativamente simples, conocidos como monómeros [1]. Las proteínas, los ácidos nucleicos y los polisacáridos son algunos ejemplos.

Las proteínas son polímeros de aminoácidos de distinta longitud; si no incluimos el agua, son el mayor componente de la células. Las proteínas cumplen diferentes funciones, por ejemplo transporte, catálisis, estructura, defensa, comunicación y almacenamiento [2].

Existen 20 aminoácidos diferentes que se distinguen entre sí por las propiedades fisicoquímicas que les confiere la cadena lateral (R) (Figura 1.a) [3].

Los aminoácidos se unen entre sí mediante enlaces covalentes que se generan de la unión de un carboxilo con el grupo amino del aminoácido siguiente, con pérdida de una molécula de agua. Este nuevo enlace generado, del tipo amida, se denomina enlace peptídico, y es por ello que se nombra a las proteínas como polipéptidos (Figura 1.b) [3].

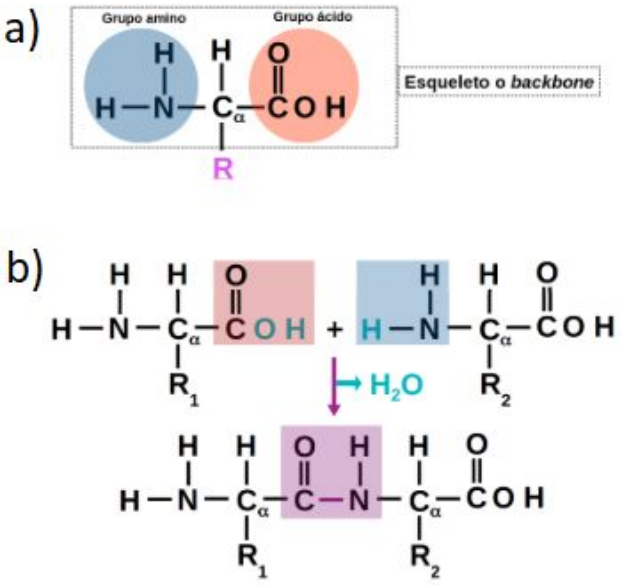

Figura 1. a) Estructura general de un aminoácido. Esta estructura es común a todos menos uno de los aminoácidos (la prolina, aminoácido cíclico, es la excepción). El grupo R o la cadena lateral (rosa) se une al carbono α, y es diferente para cada aminoácido. Imagen extraída y modificada de [3]. b) Reacción de formación del enlace peptídico. Imagen extraída y adaptada de [3].

Las proteínas, como toda macromolécula con actividad biológica, tienen una estructura tridimensional específica que depende de su composición aminoacídica y de cómo dichos aminoácidos se asocian en estructuras ordenadas. Esta organización se describe en distintos niveles estructurales denominados estructura primaria, secundaria, terciaria y cuaternaria. La estructura primaria (Figura 2.a) es la más básica y consiste en la descripción de todos los enlaces covalentes que unen a los aminoácidos de una cadena polipeptídica; es decir, la secuencia ordenada de aminoácidos. La estructura secundaria (Figura 2.b) se refiere a disposiciones particularmente estables de los aminoácidos que dan lugar a patrones estructurales repetitivos. Los dos tipos más comunes de estructuras secundarias son las hélices alfa y las hojas betas. Estas estructuras se conectan unas a otras a través de loops, y se estabilizan a través de puentes de hidrógeno. La estructura terciaria del polipéptido plegado es la disposición en el espacio de los elementos de estructura secundaria (Figura 2.c). Por último, la estructura cuaternaria (Figura 2.d) se forma mediante la unión de varias cadenas con estructura terciaria para formar un complejo proteico; en otras palabras, se refiere a la disposición en el espacio cuando una proteína posee dos o más cadenas polipeptídicas [1].

Es importante destacar que la secuencia de una proteína determina su estructura tridimensional, la cual, a su vez, determina su función. La estructura tridimensional en condiciones fisiológicas, donde la proteína se encuentra más estable y funcionalmente activa, es lo que se denomina “estructura nativa”. Una forma de describir a la estructura espacial es indicando los aminoácidos de la estructura primaria que se encuentran en contacto. Dichos contactos son los responsables de la estructura tridimensional. Cabe aclarar que contacto se considera cuando dos aminoácidos tienen átomos que se encuentran a distancia menor de *n* Ångström (Å). Si bien el número *n* puede variar según el estudio que se realiza, en general, suele encontrarse entre 4 y 8 [4].

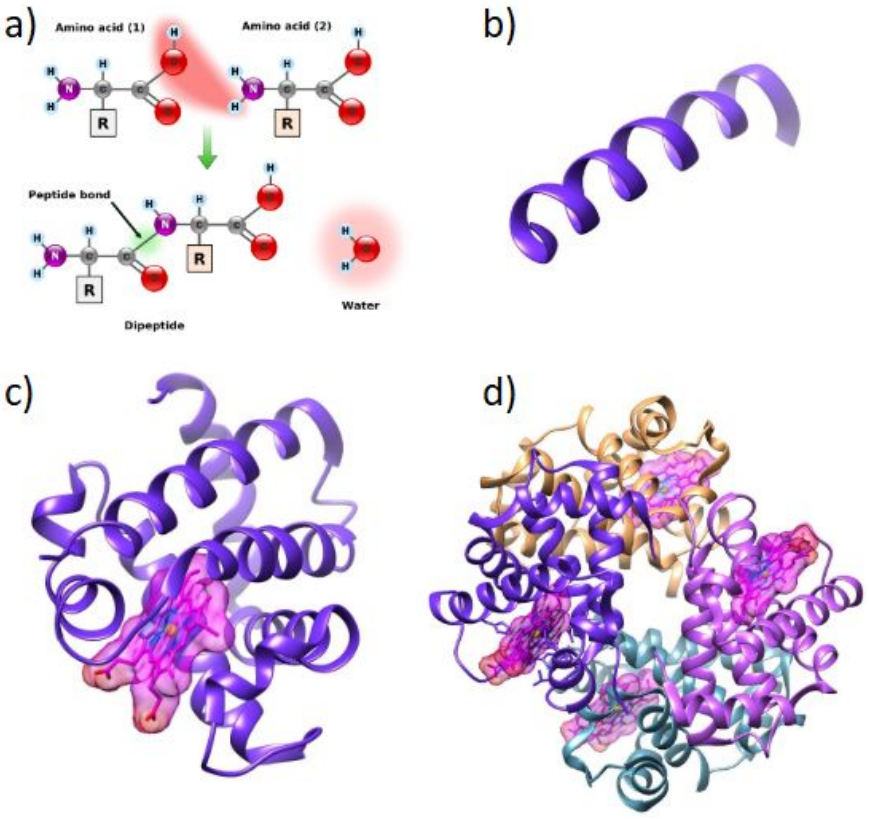

Figura 2. Niveles de estructura de proteínas. Ej: hemoglobina humana (pdb: 1N7N). a) La estructura primaria consiste en una secuencia de aminoácidos unidos entre sí por enlaces peptídicos. b) El polipéptido resultante puede adquirir unidades de estructura secundarias α hélice y hojas β (representación de cintas). c) La estructura terciaria de una de las cadenas de la hemoglobina con el grupo hemo en el sitio activo (representación en sticks). d) La disposición espacial de dos o más cadenas con estructura terciaria componen la estructura cuaternaria de la proteína, en este caso, la hemoglobina humana. Imágenes generadas con Chimera [5].

Una proteína puede adoptar disposiciones espaciales ligeramente diferentes, denominados confórmeros. Esto se debe a la rotación de los enlaces simples en el espacio. Un modelo de estructura es un confórmero determinado (Figura 3). La resolución de la estructura de la misma proteína, en diferentes condiciones, es lo que nos da información de los diferentes confórmeros.

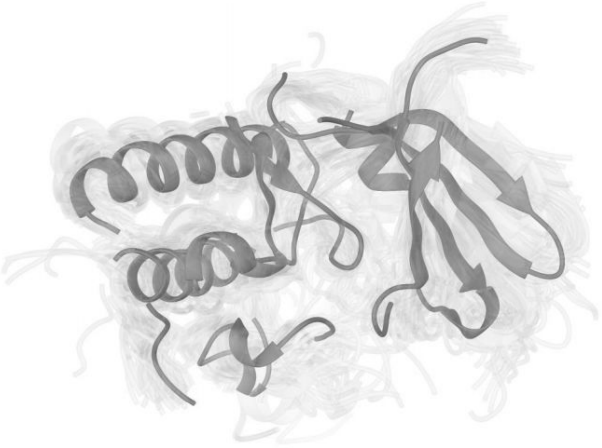

Figura 3. Superposición estructural de 77 confórmeros de una proteína quinasa. En gris oscuro se muestra la proteína PIM1 (pdb: 4N70). Imagen tomada de [6].

Se ha demostrado que la estructura y la función de las proteínas están estrechamente relacionadas. Cualquier cambio o alteración en la conformación nativa de una proteína puede dar como resultado que una función fisiológica normal derive en un proceso patológico.

Las proteínas en su entorno biológico natural no son objetos rígidos: se permite y es necesaria cierta flexibilidad estructural [7]. Por lo tanto, para cada proteína, el estado nativo no se limita a una conformación única, sino, que varios confórmeros coexisten en equilibrio. Una gran variedad de enfermedades humanas, como el cáncer, enfermedades cardiovasculares, la amiloidosis, enfermedades neurodegenerativas, diabetes, poliglutamina, la enfermedad de Parkinson, la deficiencia de antitripsina, el Alzheimer y las encefalopatías espongiformes, se deben, en parte, a errores en el plegamiento del conjunto de estructuras, o desequilibrios de los confórmeros del estado nativo, por lo tanto, pueden considerarse enfermedades conformacionales [8][9][10][11][12].

Es muy importante conocer la estructura tridimensional de las proteínas de modo de poder encontrar relaciones entre la estructura y la función. Actualmente, hay diferentes métodos utilizados para determinar la estructura de una proteína, incluyendo la cristalografía de rayos X (X-ray crystallography), la espectroscopía de resonancia magnética nuclear (*nuclear magnetic resonance spectroscopy - NMR spectroscopy*), y la criomicroscopía electrónica (*cryo electron microscopy*). Cada uno de los métodos tiene ventajas y desventaja y utilizan diferentes elementos de información para crear el modelo atómico final.

La cristalografía de rayos X (Figura 4.a) es una técnica para determinar la estructura tridimensional de las moléculas, incluidas macromoléculas biológicas complejas, como las proteínas y los ácidos nucleicos. Es una herramienta poderosa en la elucidación de la estructura tridimensional de una molécula en resolución atómica. Los datos se recolectan al difractar rayos X a un solo cristal, que tiene una disposición ordenada y regular de átomos. Sobre la base del patrón de difracción obtenido a partir de la dispersión de rayos X se puede reconstruir la densidad electrónica y a partir de ésta se puede inferir la posición atómica [13].

La espectroscopía de resonancia magnética nuclear (Figura 4.b) es una técnica que tiene como elementos principales un imán, un transmisor, una antena y una computadora. El imán produce un campo magnético estático estable (campo magnético principal) que se utiliza para generar magnetización macroscópica en una muestra de resonancia magnética. De esta manera, se logra orientar todos los núcleos en una misma dirección. Luego, el transmisor proporciona pulsos de radiofrecuencia para irradiar a la muestra con una frecuencia determinada, generando, al polarizar sus átomos, una desviación de los núcleos del campo magnético principal. Finalizada la emisión de dichos pulsos, los átomos de la muestra tienden a orientarse nuevamente según la dirección generada por el imán, emitiendo, durante el proceso de despolarización, la señal de resonancia magnética. Ésta, es detectada por una antena amplificadora y, mediante un convertidor analógico a digital (ADC), se digitaliza para su posterior procesamiento y visualización de datos en una computadora [15].

Como se mencionó, la difracción de rayos X puede dar estructuras de biomoléculas de muy alta resolución. Pero, para obtener una estructura mediante este método, es necesario poder cristalizar la molécula. Muchas proteínas no se cristalizan y, en algunos casos, el proceso de cristalización altera la estructura, por lo que no es representativo de cómo se ve la molécula en la vida real. Con la espectrometría de resonancia magnética también se puede obtener estructuras de buena resolución, aunque tiene un limitante: suele utilizarse para proteínas relativamente pequeñas, partes de proteínas, o proteínas intracelulares solubles (se suele evitar proteínas de la membrana celular). Si se desea estudiar proteínas grandes, receptores unidos a la membrana, complejos de varias biomoléculas juntas o evitar la cristalización, se puede recurrir a la microscopía electrónica criogénica (Figura 4.c) [17].

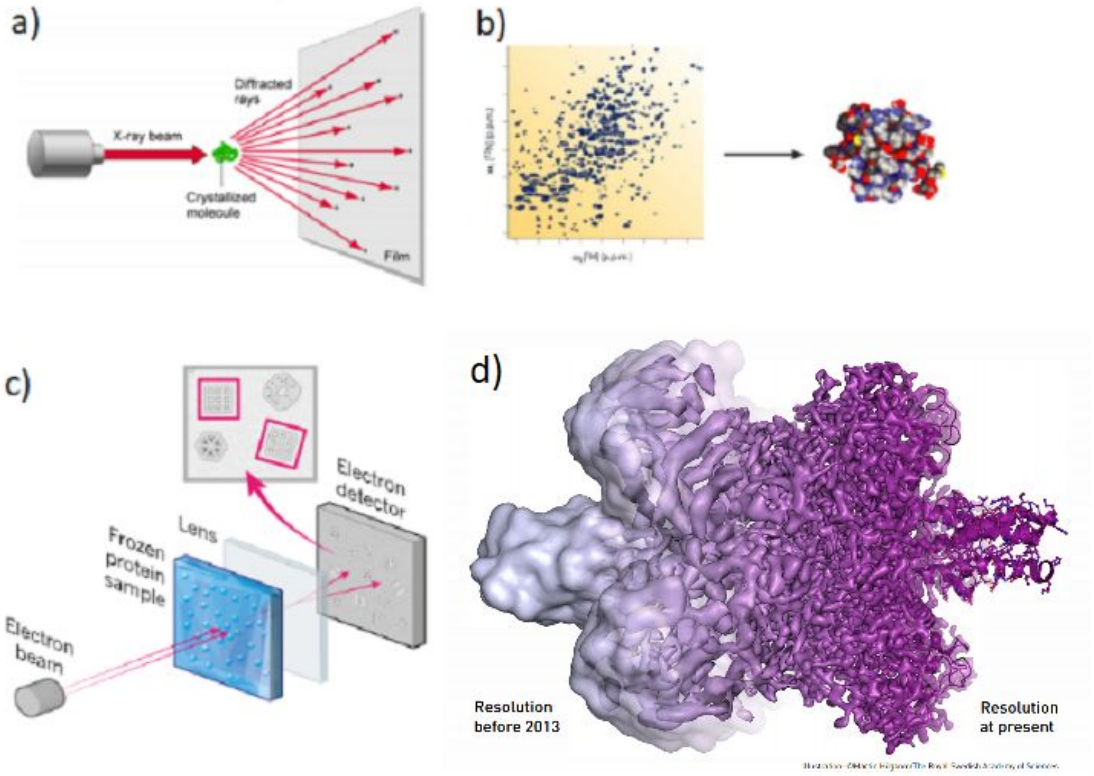

Figura 4. a) Un patrón de difracción de rayos X de una enzima cristalizada. El patrón de puntos (reflexiones) y la fuerza relativa de cada punto (intensidades) se pueden utilizar para determinar la estructura de la enzima. Imagen tomada de [14]. b) Espectroscopía de Resonancia Magnética Nuclear. Un ejemplo del patrón complejo obtenido de la espectroscopia de NMR (izquierda), que correlaciona los átomos de nitrógeno (eje vertical) de los grupos amida con los átomos de hidrógeno (eje horizontal) en una proteína. Una serie compleja de mediciones que correlacionan otros núcleos (carbono-hidrógeno, hidrógeno-hidrógeno), cálculos y deducciones, permiten derivar la estructura de la proteína (derecha). Imagen tomada de [16]. c) Microscopía electrónica criogénica. Un haz de electrones se dispara a una solución de proteína congelada. Los electrones dispersos emergentes pasan a través de una lente para crear una imagen ampliada en el detector, a partir de la cual se puede determinar su estructura. Imagen tomada de [19]. d) La progresión de la resolución de criomicroscopía electrónica, ilustrada por una representación de glutamato deshidrogenasa con un nivel de detalle creciente de izquierda a derecha. Para una proteína de este tamaño, 334 kDa, una resolución de 1.8 Å, la cual se pudo lograr después del 2012 [18].

La criomicroscopía electrónica es una técnica derivada de la microscopía electrónica. Los microscopios electrónicos de transmisión (TEM) utilizan un haz de electrones para examinar las estructuras de las moléculas y los materiales a escala atómica. A medida que el haz pasa a través de una muestra muy delgada, interactúa con las moléculas, que proyectan una imagen de la

muestra en el detector. Sin embargo, algunos materiales, especialmente las biomoléculas, no son compatibles con las condiciones de alto vacío y los haces de electrones intensos utilizados en los TEM tradicionales. El agua que rodea las moléculas se evapora, y los electrones de alta energía queman y destruyen las moléculas. En cambio, la microscopía electrónica criogénica utiliza muestras congeladas, haces de electrones más suaves y un procesamiento de imágenes sofisticado para superar estos problemas. La microscopía electrónica criogénica no requiere cristales, y también permite ver cómo las biomoléculas se mueven e interactúan a medida que realizan sus funciones, lo que es difícil con la cristalografía. Esta técnica es única, ya que no requiere cristalización, utiliza poco material y resuelve partículas del tamaño de la hemoglobina (64 kDa) hasta grandes partículas en el orden de los mega Dalton. La evolución en la mejora de la resolución de esta técnica se ve reflejada en la Figura 4.d. La importancia de esta técnica y de la investigación en este campo disciplinar se vio reflejada en el premio Nobel de Química otorgado a los investigadores Joachim Frank, Richard Henderson y Jacques Dubochet en el año 2017 [17][18].

Cualquiera de los tres métodos mencionados tienen como resultado un archivo pdb (*Protein Data Bank*) [20]. Este, es un archivo de texto (Figura 5) que presenta la información espacial de la proteína, es decir, las coordenadas espaciales X, Y, y Z, correspondientes a la posición de cada uno de los átomos de los aminoácidos, además de incluir cierta información complementaria [20]:

- nombre de la proteína,
- fecha de generación del archivo,
- código del pdb (4 caracteres alfanuméricos),
- la cantidad de cadenas y de residuos presentes,
- la técnica experimental utilizada para la determinación de la estructura,
- los autores del archivo,
- la fuente biológica de la macromolécula de entrada,
- la resolución obtenida,
- las asignaciones de estructuras secundarias,
- la conectividad atómica,

- la conectividad entre residuos,
- los ligandos (en caso de que presenten),
- las moléculas de aguas que forman parte de la solución (puede detallarse si algunas se encuentran coordinadas con la proteína), y/o,
- la presencia de átomos o residuos faltantes (átomos o residuos que no se pudieron determinar por el método utilizado).

```
HEADER    CELL ADHESION                           03-DEC-15   5F4V
TITLE     CRYSTAL STRUCTURE OF THE HUMAN SPERM IZUMO1 RESIDUES 22-268
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: IZUMO SPERM-EGG FUSION PROTEIN 1;
COMPND   3 CHAIN: A;
COMPND   4 FRAGMENT: UNP RESIDUES 22-268;
COMPND   5 SYNONYM: OOCYTE BINDING/FUSION FACTOR,OBF,SPERM-SPECIFIC PROTEIN
COMPND   6 IZUMO;
COMPND   7 ENGINEERED: YES
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: HOMO SAPIENS;
SOURCE   3 ORGANISM_COMMON: HUMAN;
SOURCE   4 ORGANISM_TAXID: 9606;
SOURCE   5 GENE: IZUMO1;
SOURCE   6 EXPRESSION_SYSTEM: DROSOPHILA MELANOGASTER;
SOURCE   7 EXPRESSION_SYSTEM_COMMON: FRUIT FLY;
SOURCE   8 EXPRESSION_SYSTEM_TAXID: 7227
KEYWDS    GLYCOPROTEIN, MEMBRANE-BOUND, CYSTEINE-RICH, ADHESION, FUSION, CELL
KEYWDS   2 ADHESION
EXPDTA    X-RAY DIFFRACTION
AUTHOR    H.AYDIN,A.SULTANA,J.E.LEE
REVDAT   4   28-DEC-16 5F4V    1       TITLE
REVDAT   3   06-JUL-16 5F4V    1       JRNL
REVDAT   2   29-JUN-16 5F4V    1       JRNL
REVDAT   1   15-JUN-16 5F4V    0
JRNL        AUTH   H.AYDIN,A.SULTANA,S.LI,A.THAVALINGAM,J.E.LEE
JRNL        TITL   MOLECULAR ARCHITECTURE OF THE HUMAN SPERM IZUMO1 AND EGG
JRNL        TITL 2 JUNO FERTILIZATION COMPLEX.
JRNL        REF    NATURE                         V. 534   562 2016
JRNL        REFN                   ESSN 1476-4687
JRNL        PMID   27309818
JRNL        DOI    10.1038/NATURE18595
REMARK   2
```

Figura 5. Captura de pantalla del archivo *5F4V.pdb* para ejemplificar la información de un archivo "*.pdb*".

El Banco de Datos de Proteínas (*Protein Data Bank*, PDB) es el único archivo mundial de datos estructurales de macromoléculas biológicas. Es un recurso global líder de datos digitales (proveniente de datos experimentales) de acceso abierto en el área de la biología y la medicina, y es un recurso central para el libre uso científico [21].

Muchas fuentes secundarias de información se derivan de datos de PDB. Es el punto de partida para estudios en bioinformática estructural. PDB es un repositorio de coordenadas atómicas y otra

información anexa para describir macromoléculas de importancia biológica. Los datos depositados fueron obtenidos mediante las técnicas antes descriptas (cristalografía de rayos X, RMN y criomicroscopía electrónica), son anotados y forman parte de un archivo que está disponible de manera pública [21].

El *Protein Data Bank* se estableció en el Laboratorio Nacional Brookhaven (*BNL*) [22] en 1971, como un repositorio de archivos de estructuras cristalinas macromoleculares biológicas. Representa una de las primeras colecciones de datos de biología molecular impulsadas por la comunidad científica [23]. Al día de hoy, en el PDB se encuentran determinadas y depositadas, para ser utilizadas públicamente, 154.214 estructuras. En la Figura 6 se puede apreciar la estructura de la porción extracelular de la proteína humana Izumo1 de espermatozoide, depositada con el código pdb 5F4V. Esta proteína de membrana del espermatozoide, junto con su contraparte JUNO, proteína de membrana del huevo, participan en la fusión de las gametas.

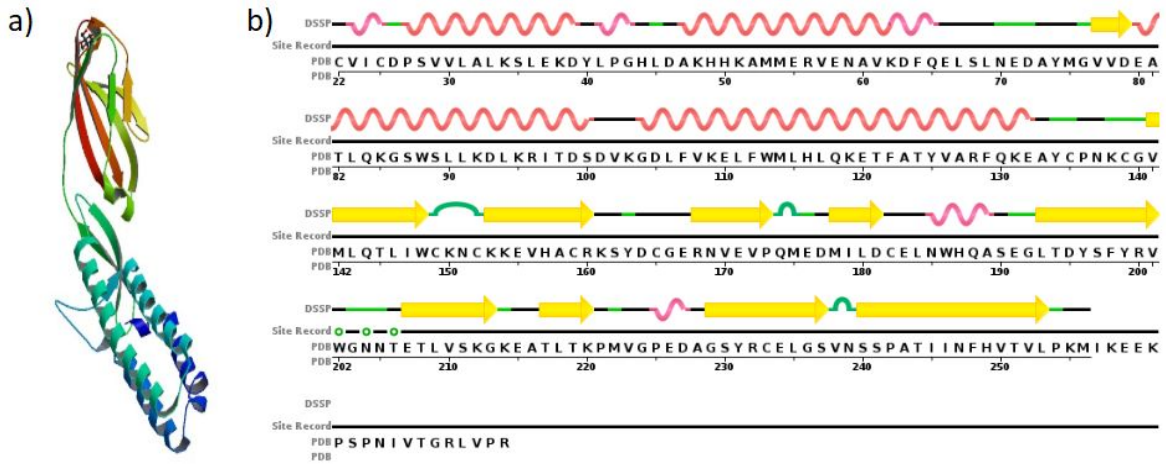

Figura 6. Estructura espacial de la proteína Izumo1, entrada de pdb 5F4V. a) Podemos apreciar las alfa hélices, hojas beta y loops que las conectan. b) Se presenta la secuencia primaria de aminoácidos de la proteína y, sobre ella, la secuencia secundaria: las alfa hélices esquematizadas como resortes y las hojas beta esquematizadas como flechas. Imágenes obtenida de https://www.rcsb.org/structure/5F4V.

Al explorar la base de datos PDB, se puede encontrar archivos de coordenadas para la “unidad biológica" y para la "unidad asimétrica". Muchas entradas de PDB son iguales [24]. Sin embargo, para algunas proteínas (principalmente aquellas resueltas por cristalografía de rayos X), es posible que ambas unidades no coincidan [24]. El archivo de coordenadas primario de una estructura

cristalina normalmente contiene solo una unidad asimétrica de cristal y puede o no ser el mismo que el del la unidad biológica [24].

La unidad asimétrica es la porción más pequeña de una estructura de cristal a la que se pueden aplicar operaciones de simetría para generar la celda de la unidad completa (la unidad de repetición de cristal) [24]. Las operaciones de simetría más comunes para los cristales de macromoléculas biológicas son rotaciones, traslaciones y sus combinaciones . La aplicación de operaciones de simetría cristalográfica a una unidad asimétrica produce una celda única que permite reconstruir el cristal completo (Figura 7) [24].

Una unidad asimétrica de cristal puede contener [24]:

- una unidad biológica, unidad asimétrica y biológica coinciden;
- una porción de un conjunto biológico, es el caso donde puede reconstruirse la unidad biológica a partir de la unidad asimétrica;
- múltiples unidades biológicas.

Vamos a hacer un resumen de los ejemplos que mencionan los autores del artículo:

> Unidad asimétrica = Unidad biológica: Hemoglobina con sus cuatro cadenas (Figura 7.a).
>
> Unidad asimétrica = Parte de la unidad biológica: Hemoglobina con solo dos cadenas, a partir de las cuales aplicando funciones de rotación y traslación y sus combinaciones se puede reconstruir el tetrámero biológicamente funcional de la proteína (Figura 7.b).
>
> Unidad asimétrica = Muchas unidades biológicas: Hemoglobina en cuyo cristal se encuentran dos tetrámeros no exactamente idénticos, donde cada uno es una unidad biológicamente funcional (Figura 7.c).

Finalmente, la unidad biológica es la estructura macromolecular que se ha demostrado que es o que se presume que es la forma biológicamente funcional de la molécula [24].

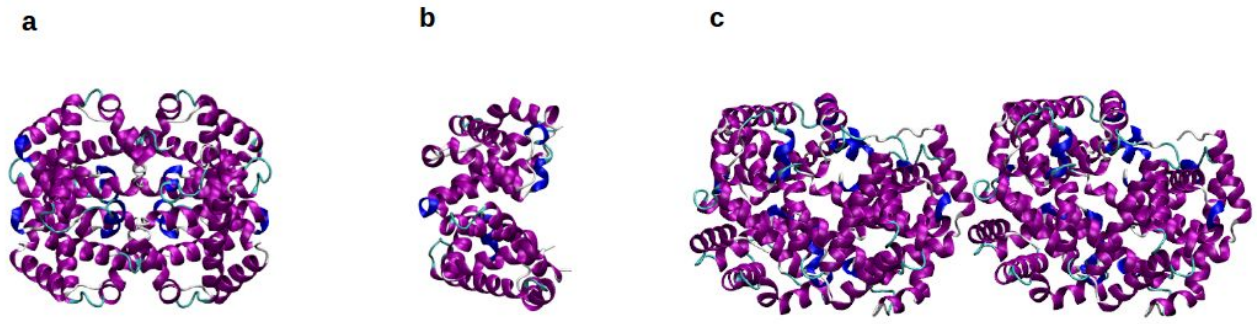

Figura 7. Diferentes casos de de unidad asimétrica (ejemplo: hemoglobina). a) Unidad asimétrica = Unidad biológica (PDB: 2HHB), b) Unidad asimétrica = Porción de la unidad biológica (PDB: 1OUT), c) Unidad asimétrica = Dos veces la unidad biológica (PDB:) Imagen adaptada de [24] generada con VMD.

Por otro lado, si bien las proteínas en su estado nativo son un conjunto de estructuras ligeramente diferentes en equilibrio, es decir, confórmeros, ¿cuál es la dimensión de estas diferencias? Las diferencias entre estructuras de los confórmeros de cada proteína es medida por la desviación media cuadrática *(Root Mean Square Deviation*, RMSD) de las posiciones atómicas al superponer los distintos confórmeros. Este parámetro refiere a la similitud o diferencias entre dos estructuras a través de una superposición de los conjuntos de coordenadas atómicas. El RMSD de estructuras alineadas da una idea de la divergencia: valores bajos de RMSD indican que las estructuras se superponen mejor, por lo tanto, las estructuras son relativamente similares. Mientras que valores altos indican diferencias mayores entre estructuras. En general, los valores de RMSD se encuentran alrededor 1-10 Å.

Como describen Marchetti y col. [25], a las proteínas con diferencias entre los confórmeros muy chica, o nula, se las denominan “rígidas”. Por otro lado, cuando las diferencias entre confórmeros son mayores, las proteínas se las denomina “móviles”. Por último, en las proteínas “intrínsecamente desestructuradas” (IDP), la superposición en el espacio de los confórmeros es imposible. La Figura 8 muestra un ejemplo de cada uno de estos conjuntos. A la izquierda se muestra un ejemplo de una proteína globular ordenada. Estas proteínas muestran grandes proporciones de estructura secundaria. La mayor cantidad de interacciones entre residuos de del núcleo hidrofóbico de este tipo de proteínas previene de sustitución de los aminoácidos a lo largo de la evolución que evolucionan más lento en comparación con los residuos expuestos. En el centro se muestra una proteína con regiones ordenadas o globulares, pero también con regiones

muy flexibles que le otorga un comportamiento dinámico. A las derecha se muestra una proteína IDP, con varios confórmeros que presentan grandes diferencias estructurales. Son confórmeros que muestran cadenas altamente flexibles y segmentos eventualmente pequeños y transitorios de estructura secundaria o terciaria [25].

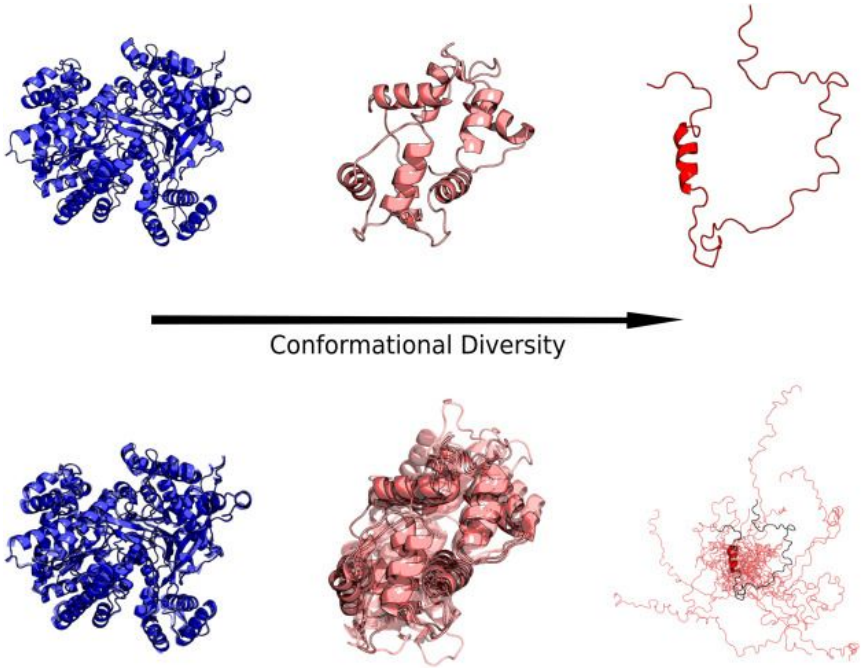

Figura 8. Diferentes conjuntos de proteínas en función del incremento de la flexibilidad. El panel superior muestra un confórmero determinado de la proteína, mientras que el panel inferior muestra todos los confórmeros disponibles superpuestos. A la izquierda, la maltodextrina fosforilasa, (códigos PDB: 1AHP_A, 1AHP_B, 1L5V_B). Se muestra como una proteína rígida con un 6,53% de desorden y se tomó como representante de proteínas ordenadas. La calmodulina (códigos PDB: 2FOT_A, 1LIN_A, 1NIW_E, 3G43_A, 2BE6_A, 1CDL_A, 3GP2_A, 4L79_B, 1CLL_A) muestra un 10.64% de desorden. La fosfoproteína soluble de tilacoides de espinaca, (ID PDB: 2FFT_A) es un conjunto típico de IDP con el 100% de desorden estimado. Los porcentajes de desorden se estimaron con ESpritz. Imagen tomada de [25].

Como la función biológica de una proteína está intrínsecamente relacionada con su dinámica, se puede analizar mediante el estudio de diferentes estructuras de la misma proteína determinadas experimentalmente y depositadas en el PDB. Una misma proteína pudo haber sido depositada varias veces en el PDB debido a que las distintas estructuras fueron obtenidas en diferentes condiciones experimentales (pH, temperatura, presencia de un ligando, etc.). Todas ellas son una muestra de las posibles estructuras que esa proteína puede tomar, llamadas confórmeros, y da

cuenta de la diversidad estructural/conformacional: diferencias estructurales entre los confórmeros. Por lo tanto, la redundancia de estructuras en el PDB (varias conformaciones para la misma proteína), es un dato esencial que contribuye a los estudios basados en experimentos de dinámica de proteínas y proporciona información sobre su función [26].

La diversidad conformacional es un concepto clave en la comprensión de diferentes problemas relacionados con la función de las proteínas, como el estudio de procesos catalíticos en enzimas, el reconocimiento de proteínas, la evolución de las mismas y los orígenes de las nuevas funciones biológicas. Para dar cuenta de estas diferencias se desarrolló la base de datos CoDNaS (Diversidad Conformacional del Estado Nativo - *Conformational Diversity of Native State*), que es una base de datos de proteínas con diferentes grados de diversidad conformacional (Figura 9) [27]. Es una “colección redundante”, y es aquí donde reside la importancia de esta base de estructuras tridimensionales para la misma proteína tomada del PDB. Así, las estructuras para la misma proteína obtenidas en diferentes condiciones cristalográficas se asocian para dar cuenta de la dinámica de las proteínas. Es muy importante destacar que CoDNaS utiliza cadenas individuales para hacer los cálculos, aunque las cadenas pertenezcan a una misma proteína multimérica. Para poder relacionar nuestros datos con los de esta base, analizamos cada cadena de forma individual.

CoDNaS permite explorar las diferencias estructurales globales y locales entre los confórmeros en función de diferentes parámetros, como la presencia de ligandos, las modificaciones postraduccionales, los cambios en los estados oligoméricos y las diferencias en el pH y la temperatura. CoDNaS presenta también información sobre la taxonomía y la función de las proteínas. La clasificación estructural ofrece información útil para explorar el mecanismo subyacente de la diversidad conformacional y su estrecha relación con la función proteica [28]. Actualmente, CoDNaS tiene un total de 320.144 conformaciones provenientes de 21.152 cadenas proteicas.

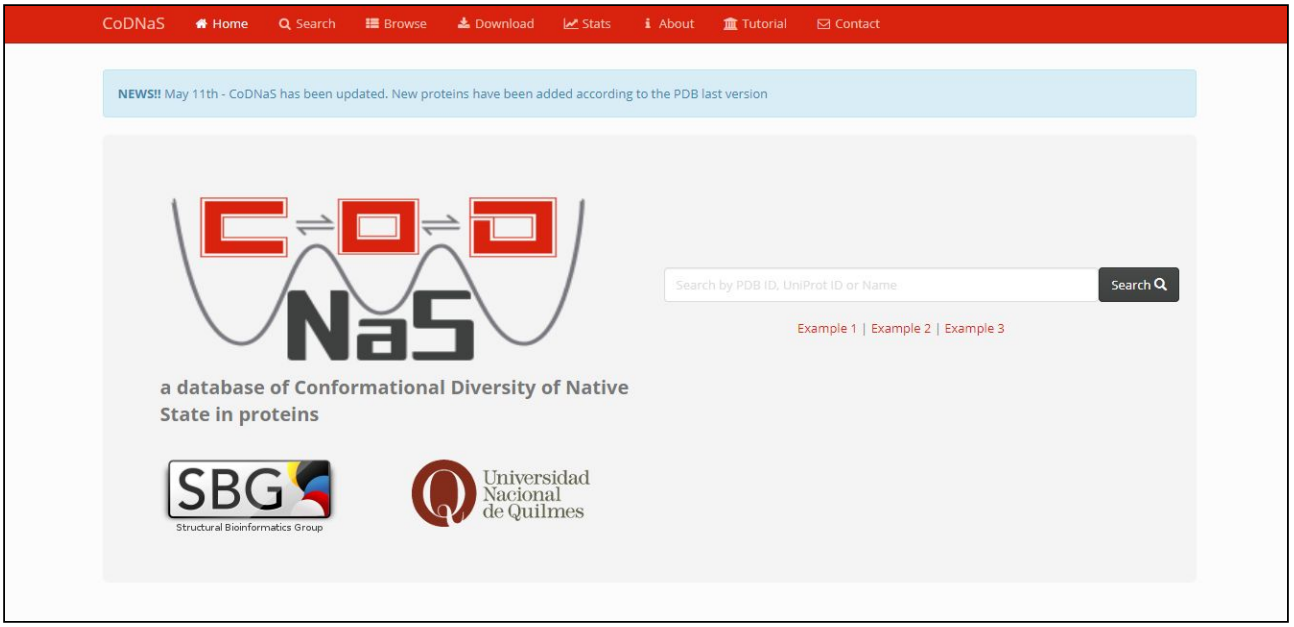

Figura 9. Captura de pantalla de la web de CoDNaS [28].

Mediante el análisis de los confórmeros, Monzón y col. han definido, en función de su diversidad conformacional, tres grupos de proteínas (Figura 10) [29]:

- Proteínas rígidas: el mayor de estos subconjuntos de proteínas (~60%). Presentan un RMSD promedio de 0.83 Å, no tienen regiones desordenadas, muestran una baja diversidad conformacional, tienen túneles grandes y cavidades pequeñas y enterradas.
- Proteínas parcialmente desordenadas: tienen en promedio el 67% de sus confórmeros con regiones desordenadas, un RMSD promedio de 1.1 Å, el mayor número de bisagras o *hinges*, y regiones desordenadas más largas.
- Proteínas maleables: tienen en promedio solo el 25% de los confórmeros desordenados, un RMSD promedio de 1.3 Å, cavidades flexibles afectadas en tamaño por la presencia de regiones desordenadas y muestran la mayor diversidad de ligandos afines.

Los dos últimos subgrupos se los puede unir y considerar como proteínas móviles.

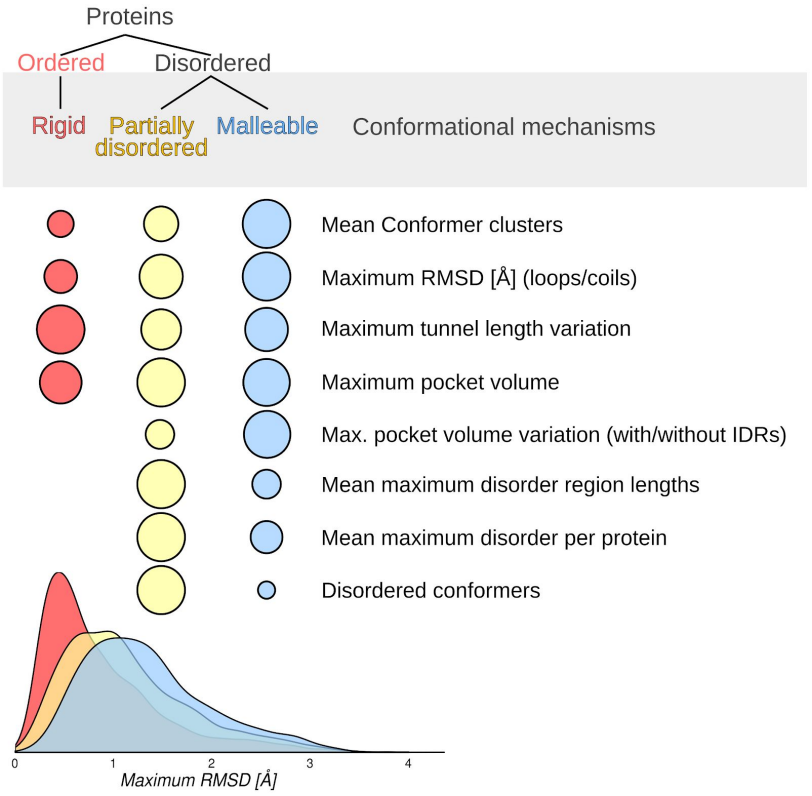

Figura 10. Comparación visual de las principales características estructurales en cada uno de los tres conjuntos de proteínas descritos. El área de cada círculo es proporcional a la medida cuantitativa correspondiente promedio. Imagen tomada de [29].

Para aportar a la claridad de estos conceptos, en la Figura 11 se muestra la superposición de los confórmeros de una proteína del subconjunto rígido y una del subconjunto de proteínas maleables, Figura 11.a y Figura 11.b respectivamente.

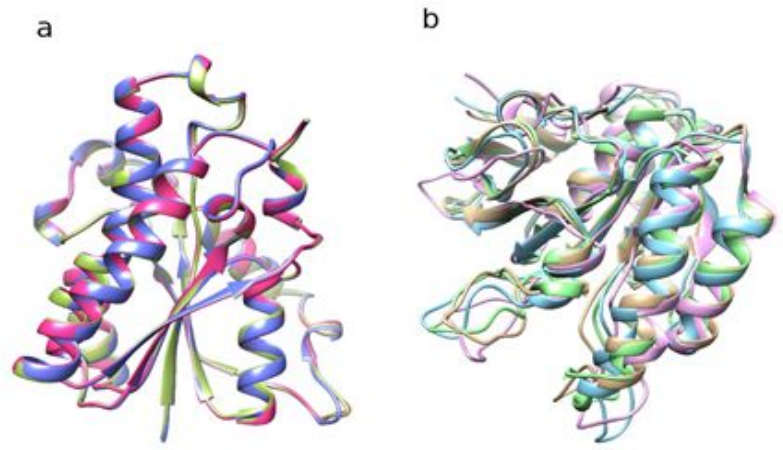

Figura 11. Proteínas depositadas en la base de datos CoDNaS, tomando como base los códigos 2NHQ y 2CVF. a) 2NHQ, proteína rígida (superposición de 3 confórmeros). b) 2CVF, proteína maleable (superposición de 3 confórmeros). Las estructuras están representadas por cintas. La representación fue hecha con Chimera [6].

Si bien la flexibilidad proteica se puede evaluar comparando los múltiples confórmeros de una sola proteína recopilados en CoDNaS, no todas las proteínas tienen un número adecuado de confórmeros. Además, estos son representaciones estáticas de las estructuras 3D.

Adicionalmente, a pesar de lo que se ha avanzado en este área, sigue sin conocerse dónde se almacena la información que indica si una proteína tendrá una pequeña o gran diversidad conformacional. Los estudios basados en secuencias no dieron esta respuesta.

Por otro lado, la dinámica molecular (DM) (*Molecular Dynamics*) es una herramienta que permite la simulación computacional y el estudio del movimiento e interacciones de átomos y moléculas [30]. Este método que se desarrolló originalmente en el campo de la física teórica a fines de la década de 1950, hoy se aplica principalmente en física, química, estudio de materiales y modelado de biomoléculas [31]. Las simulaciones de DM permiten comprender las bases físicas de la estructura y la función de las macromoléculas biológicas mediante el movimiento molecular. La visión inicial de las proteínas como estructuras relativamente rígidas ha sido reemplazada por un modelo dinámico en el que los movimientos internos y los cambios conformacionales resultantes desempeñan un papel esencial en su función. Las simulaciones pueden proporcionar detalle respecto a los movimientos de partículas individuales en el tiempo generando una trayectoria de coordenadas atómicas en función del tiempo. Estas simulaciones permiten abordar preguntas

específicas sobre las propiedades de un sistema modelo dinámico, a veces más fácilmente que los experimentos en el sistema real, ayudando a descartar experimentos que, por sus costos, no tiene sentido realizarlos, orientando, de este modo, a programar experimentos pero teniendo en cuenta los datos obtenidos previamente *in silico*. Los datos obtenidos en la escala de dinámica molecular son muy importantes ya que muchas veces la función de las biomoléculas depende de interacciones a nivel nanoscópico revelando detalles que son de sumo interés a nivel estructural, es allí donde DM ayuda a brindar respuestas [30].

En DM, para estudiar los movimientos físicos de los átomos y las moléculas, se permite que los mismos interactúen durante un período de tiempo fijo, dando una visión de la evolución dinámica del sistema. En la forma más sencilla de simulación, las trayectorias de los átomos y las moléculas se determinan resolviendo numéricamente las ecuaciones de movimiento de Newton para un sistema de partículas en interacción, donde las fuerzas entre las partículas y sus energías potenciales a menudo se calculan utilizando potenciales interatómicos descriptos mediante mecánica clásica. Debido a que los sistemas moleculares consisten en un gran número de partículas, para determinar analíticamente las propiedades de tales sistemas complejos, las simulaciones de DM utiliza métodos numéricos [31].

En DM se resuelven las ecuaciones de movimiento de Newton para un sistema de N átomos que interactúan [32]:

$$m_i \frac{\partial^2 r_i}{\partial t^2} = F_i,\ i = 1,\ ...,\ N$$

Las fuerzas son las derivadas negativas de una función potencial $V(\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_N)$ [32]:

$$F_i = -\frac{\partial V}{\partial r_i}$$

De esta manera a partir de la energía potencial y mediante las ecuaciones de Newton se puede estimar la posición en cada paso de tiempo. Así, los programas computacionales que hacen simulación por DM resuelven las ecuaciones de Newton en pasos de tiempo predeterminados, manteniendo el sistema con la temperatura y la presión controlada, y escriben las coordenadas atómicas en un archivo de salida a intervalos regulares. Las coordenadas en función del tiempo son la trayectoria del sistema. Dado que la estructura de la que se parte es estática, hay varios pasos a realizar para que pueda analizarse una estructura en movimiento. De este modo, luego de varios pasos previos, cuando el sistema está listo para simular un experimento y se inicia la

simulación, al cabo de un tiempo se alcanzará un estado de equilibrio, y es en esta condición de equilibrio en la trayectoria donde se pueden extraer propiedades macroscópicas del sistema [32].

Resolver las ecuaciones de Newton implica que DM utiliza mecánica clásica para describir el movimiento de los átomos. En DM se usa campos de fuerza conservadores que son en función solamente de las posiciones de los átomos. Esto significa utilizar la aproximación de Born-Oppenheimer que considera que los electrones se ajustan instantáneamente al movimiento de los átomos, de este modo se simplifica el sistema ya que no se debe calcular en cada paso el movimiento de los electrones [32]. Entonces los métodos que usan campos de fuerza calculan la energía solamente a partir de las posiciones de los núcleos. La mecánica molecular se basa en un modelo simple de interacciones donde contribuyen: el estiramiento de uniones, la apertura o cierre de los ángulos y la rotación alrededor de los enlaces simples. Los campos de fuerzas más sencillos pueden describirse mediante cinco componentes que describen las fuerzas intra e intermoleculares en el sistema que se desea simular. Así para el cálculo de la energía potencial en función de las posiciones Ep (r) se suman las energías de unión y las energías de no unión. Las energías de unión se modelan mediante: 1- un potencial armónico para la energía de enlace, 2- un término que modela mediante un potencial armónico la energía de flexión angular, 3- un término que modela la energía de rotación alrededor de un enlace mediante una función coseno (energía torsional). De esta manera, por ejemplo, se aplican penalidades en el cálculo de la energía cuando las uniones o los ángulos se alejan de sus posiciones de equilibrio o de referencia. Luego las energías de no unión se modelan utilizando un potencial de Coulomb para las interacciones electrostáticas y un potencial de Lennard-Jones para las interacciones de Van der Waals:

$$E_P(r) = E_{unión} + E_{no\ unión}$$

$$E_P(r) = E_{enlace} + E_{angular} + E_{torsión} + E_{elect} + E_{Van\ der\ Waals}$$

Hay varios pasos a seguir para hacer dinámica de proteínas:

1- Obtener la estructura tridimensional de la molécula que se desea simular, un recurso fundamental en este paso es el PDB.

2- Una vez obtenida la estructura es necesario evaluar la calidad de la misma.

3- Luego verificar si lo que tenemos en el archivo de coordenadas es la unidad biológica o la unidad asimétrica. Si no hay coincidencia entre ambas, hay que obtener la unidad biológica.

4- Verificar si hay residuos perdidos y tomar una decisión al respecto, por ejemplo, si son el N-term y el C-term y no son importantes para la pregunta que se desea contestar, se puede decidir no incluirlos; de otro modo hay que reconstruirlos.

5- Decidir si se dejan o no todas o al menos algunas de las aguas que pueden venir de la estructura cristalina (ya que pueden ser importantes para ciertos procesos o interacciones); identificar la presencia y la importancia de iones y ligandos.

6- Una vez que se decide cuál es la estructura que se desea simular, hay que agregar los hidrógenos, que no siempre están presentes en las estructuras tridimensionales.

7- Luego, se coloca la proteína en una “caja” lo suficientemente grande y se la completa con moléculas de agua y iones. Los iones se agregan para neutralizar el sistema o para lograr una concentración específica de los mismos requerida para un experimento. La caja elegida, en general, es un cubo aunque existen otras formas geométricas que también llenan el espacio (el dodecaedro rómbico o el octaedro truncado son algunos ejemplos). Que la forma de la caja llene el espacio es importante porque, para minimizar los efectos de borde, se aplicarán a la caja de aguas/iones con la proteína localizada en el centro condiciones periódicas de contorno (PBC) [32]. Para esto, la caja del sistema se replica en las tres dimensiones, considerando la longitud de la misma y los radios de corte para las funciones de cálculo de energía, de modo que durante este cálculo una molécula no se vea afectada por su imagen periódica.

8- Una vez obtenida la caja y aplicadas las PBC, se procede a la instancia de minimización de la energía. Este proceso permite relajar la estructura que proviene, por ejemplo, de un cristal, donde hay interacciones proteína-proteína que están dadas en el cristal pero que no son reales si esa proteína es una proteína en solución. La superficie de energía potencial de un sistema como el de las proteínas es compleja. Tiene un mínimo global y un número muy grande de mínimos locales, donde todas las derivadas de la función de energía potencial con respecto a las coordenadas son cero y todas las segundas derivadas son no negativas. Desafortunadamente, la dimensionalidad del espacio de configuración y el número de mínimos locales es muy grande de modo que no existe ningún método de

minimización que garantice la determinación del mínimo global en un período de tiempo computacionalmente posible [32].

Hay varios métodos de minimización, los más utilizados son: el del descenso más pronunciado (*steepest descent)* y el del "gradiente conjugado" (*conjugate gradient*). En ambos casos se van modificando las coordenadas de los átomos en función de la dirección del gradiente negativo. Para *steepest descent*, el tamaño del paso se ajusta de manera que la búsqueda sea rápida (el movimiento siempre es cuesta abajo y no tiene en cuenta la historia de los pasos anteriores). Este es un método simple y robusto, pero su convergencia puede ser bastante lenta, especialmente en las proximidades del mínimo local. El método de gradiente conjugado de convergencia más rápida usa información de gradiente de los pasos anteriores. En general, los descensos más pronunciados lo acercarán al mínimo local más cercano muy rápidamente, mientras que los gradientes conjugados lo acercarán al mínimo local, pero se desempeñan peor lejos del mínimo [32].

9- Luego de la minimización se realizan pasos de equilibración, lo que permite incorporar al sistema temperatura y presión.

10- Finalizada la equilibración, se procede a comenzar con la simulación de producción.

En la Figura 12 se muestra un esquema de flujo global para DM [32]. Cada ejecución de DM requiere como entrada un conjunto de coordenadas iniciales y, opcionalmente, velocidades iniciales de todas las partículas involucradas. Las velocidades, de no estar presentes, se asignan mediante una distribución de velocidades de Maxwell-Boltzmann generada a partir de números al azar.

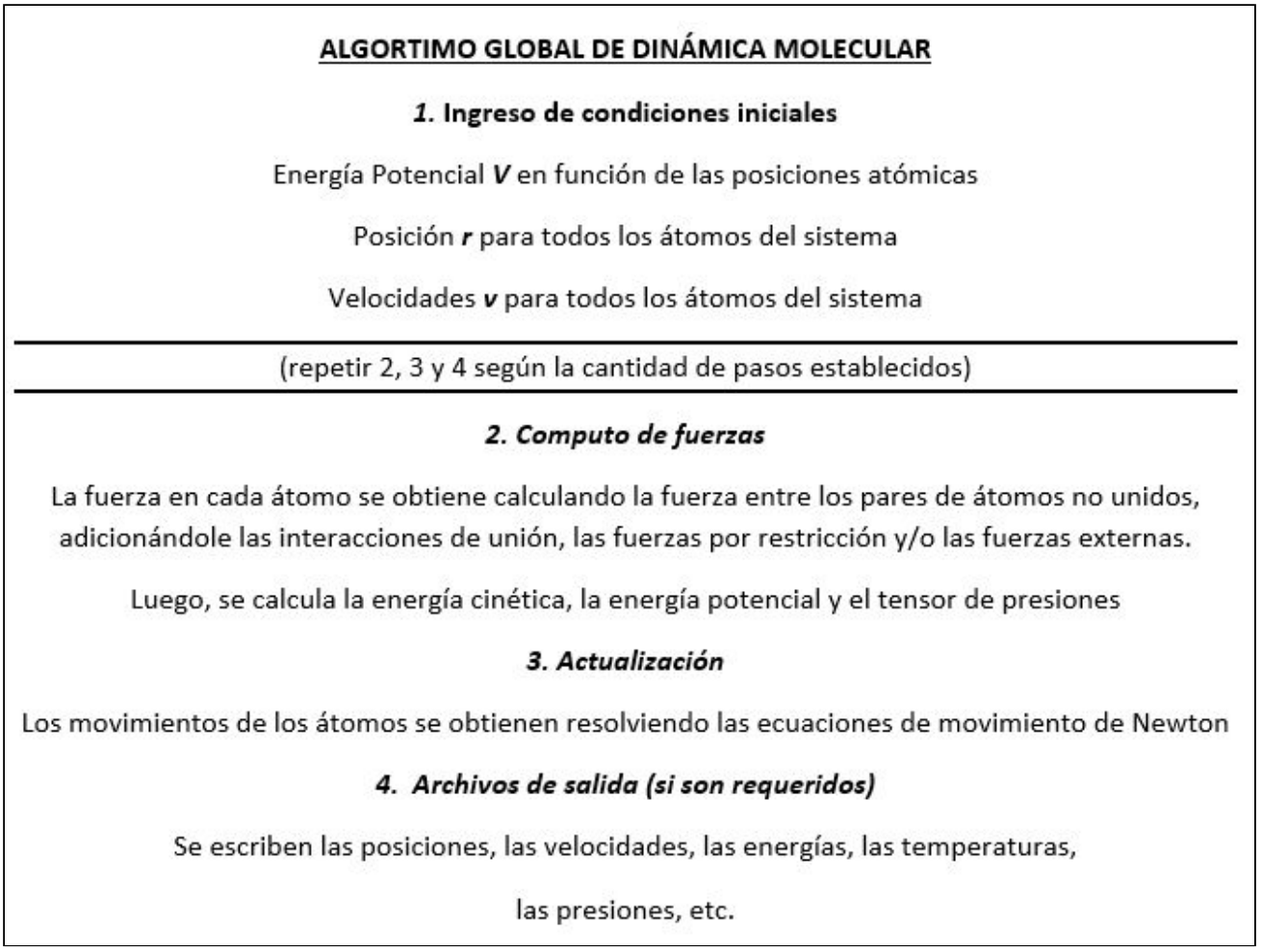

Figura 12. El algoritmo global de DM. Imagen modificada de [32].

Existen tres tipos de aplicaciones de métodos de simulación en el área macromolecular. La primera utiliza la simulación simplemente como medio para muestrear el espacio de configuración. Esto está involucrado en la utilización de la dinámica molecular para determinar o refinar estructuras con datos obtenidos de experimentos. La segunda utiliza simulaciones para obtener una descripción del sistema en equilibrio, incluidas las propiedades estructurales y de movimiento, y los valores de los parámetros termodinámicos. La tercera usa simulaciones para examinar la dinámica real: representar correctamente el desarrollo del sistema a lo largo del tiempo. Es decir, mediante esta técnica se puede evaluar la evolución dinámica de un sistema en un período de tiempo en una escala nanométrica [30]. DM ya se ha utilizado para analizar la flexibilidad de las proteínas en diferentes condiciones, como la flexibilidad térmica, mutaciones, entre otros [33][34].

Se encuentran disponibles diferentes programas que permiten el desarrollo de las simulaciones de DM. Entre ellos, GROMACS (Máquina de Groningen para la simulación química - *Groningen MAchine for Chemical Simulation*). La herramienta de simulación GROMACS se desarrolló en la Universidad de Groningen, Países Bajos, a principios de los años noventa. El software fue escrito en ANSI C. GROMACS es considerado el programa más rápido para la simulación de dinámica molecular. Si bien no tiene un campo de fuerza propio, es compatible con los campos de fuerza GROMOS, OPLS, AMBER y ENCAD, además de que puede manejar modelos de restricciones flexibles. El programa es versátil, ya que los campos de fuerza pueden ser agregados por el usuario, las funciones tabuladas se pueden especificar y los análisis se pueden personalizar fácilmente. Se incorporan dinámicas de equilibrio y determinaciones de energía libre. GROMACS está en el dominio público y se distribuye (con código fuente y documentación) bajo la Licencia Pública General de GNU. Se encuentra mantenido por un grupo de desarrolladores de las Universidades de Groningen, Uppsala y Estocolmo, y el Instituto Max Planck para la Investigación de Polímeros en Maguncia [35]. Su sitio web es http://www.gromacs.org.

El presente trabajo tiene como fin “aprender para predecir”: tal como el oráculo daba una idea del futuro a los griegos de la antigüedad, se busca utilizar la biología computacional como herramienta predictiva. Para esto, es necesario comprender la información que otorgan los programas informáticos y aprender sobre los datos arrojados por los mismos para poder obtener conclusiones. La idea central de este proyecto es investigar la posibilidad de que la información de la movilidad de las proteínas esté codificada en su estructura terciaria, es decir, si conociendo al menos una estructura de la proteína, podremos predecir si pertenece al grupo de proteínas flexibles o rígidas.

¿Cuál es el alcance de esta investigación?: un poco de números

La base de datos de secuencias de proteínas, Uniprot (Universal Protein Resource, https://www.uniprot.org/) cuenta con 150.000.000 proteínas. Es decir se conoce la secuencia de ese número de proteínas repartidas en todos los reinos de la vida (estadística de Uniprot: https://www.uniprot.org/statistics/TrEMBL). De estas, solo 40.137 tienen estructura conocida (coordenadas depositadas en el pdb), de las cuales 21.152 poseen más de un confórmero determinado (estadística de CoDNaS) [27][28].

Por lo tanto, hay 18.985 estructuras en el pdb para las que se conoce un solo confórmero, para las cuales sería de utilidad inferir, conociendo una sola estructura, a qué grupo respecto a la movilidad pertenecen esas proteínas.

Si a la vez pensamos en las restantes secuencias para las que no se conoce ninguna estructura (149.959.863), si se determinara una estructura de cada una, habría esa cantidad de proteínas para las que sería útil poder predecir su grupo estructural.

Objetivos

Para encontrar respuestas acerca de la funcionalidad de proteínas y su relación con la diversidad estructural, hay una vasta cantidad de recursos computacionales disponibles: bases de datos, programas de simulación, herramientas de visualización y modelado y servidores web con aplicación de algoritmos específicos. El conjunto de los recursos científicos disponibles sumado a su combinación e interpretación de los resultados, permitirán buscar relaciones entre la estructura y la función proteica desde distintos enfoques. Este trabajo pretende poder encontrar dichas relaciones al combinar la información estructural disponible de proteínas con la información dinámica obtenida mediante simulaciones de dinámica molecular. De esta manera, poder relacionar la estructura, su flexibilidad y su dinámica para comprender la función, sobre todo cuando está altamente relacionada a la conformación.

Partimos de las premisas de que existe una relación entre la secuencia, la estructura, la diversidad conformacional y la función en las proteínas, que muchas proteínas no tienen una estructura única en su estado nativo y que dicho ensamble de estructuras puede ser evaluado mediante la dinámica de los confórmeros nativos. Dado esto, la hipótesis es que conociendo uno de esos estados, se podrá predecir la existencia de otros necesarios para que la proteína sea funcional, mediante simulaciones de dinámica molecular. Esto es de suma utilidad ya que, como se mencionó anteriormente, para la mayoría de las proteínas no se conocen los confórmeros, no porque no los tenga, sino porque el proceso de obtención es altamente costoso y largo.

Se seleccionarán entonces grupos de proteínas tanto rígidas como móviles a partir de la base de datos CoDNaS y se comparará dicha información con la información dinámica obtenida por simulaciones de DM con el objetivo de hallar una relación entre conformación y dinámica. De esta manera, si se obtienen resultados que den cuenta de dicha relación, se podría desarrollar un predictor que permita dar cuenta de la flexibilidad de una proteína aun cuando ésta tenga muy pocos o un sólo confórmero determinado y depositado en el PDB.

Materiales y Métodos

Set de proteínas: PDB y CoDNaS

Para obtener las estructuras se utilizaron las siguientes bases de datos: PDB (*Protein Data Bank*, https://www.rcsb.org/) y CoDNaS (*Conformational Diversity of Native State in proteins*, http://ufq.unq.edu.ar/codnas/). Esta última contiene información de confórmeros por lo que nos permite elegir proteínas móviles y rígidas. Teniendo dicha clasificación, se procedió a la descarga de las mismas.

Para no introducir ningún sesgo *a priori*, el grupo de proteínas elegidas, es independiente de la clase, tamaño o estructura de las proteínas. Se eligieron teniendo en cuenta únicamente la diferencia conformacional medida en Armstrongs entre los dos confórmeros más alejados (diferencia máxima), la media entre todos los confórmeros, y que poseyera un número de confórmeros mayor a 4. Este número de confórmeros se ha demostrado que es suficiente para asegurar una buena muestra del espacio estructural de una proteína [6]. El set de proteínas utilizadas fue el siguiente:

- Rígidas:
    - 1A26: proteína “Poli (ADP-Ribose) polimerase” de *Gallus gallus* (gallo). 7 confórmeros. Monómero. RMSD: mín 0.1 - máx 0.45 - promedio 0.25. Método de obtención: Difracción de Rayos X. Se le quitaron las aguas y el ligando (carba-nicotinamida-adenina-dinucleótido).
    - 1A2T: proteína “Staphylococcal Nuclease” de *Staphylococcus aureus*. 10 confórmeros. Monómero. RMSD: mín 0.08 - máx 0.28 - promedio 0.18. Método de obtención: Difracción de Rayos X. Se ha modificado el aminoácido 23, cisteína modificada (CME), por una cisteína sin modificar. Se quitaron las aguas y uno de los ligandos (timidina-3',5'-difosfato); el ion Ca^{2+} se ha conservado.
    - 1A0G: proteina “D-AMINO acid aminotransferase” de *Bacillus sp.* 18 confórmeros. Dímero. RMSD: mín 0.1 - máx 0.55 - promedio 0.3528. Método de obtención: Difracción de Rayos X. Se le quitaron las aguas y el ligando (4'-dioxi-4'-aminopiridoxal-5'-fosfato).

- 1NHG: proteína “enoyl-acyl carrier reductase” de *Plasmodium falciparum* (Plasmodium que causa malaria en humanos). 10 confórmeros. Tetrámero. RMSD: mín 016 - máx 0.42 - promedio 0.29. Método de obtención: Difracción de Rayos X. Se quitaron las aguas y los ligandos (nicotinamida-adenina-dinucleotida y triclosano).

- Móviles:

  - 5F4V: proteína “Izumo sperm-egg fusion protein 1” de *Homo Sapiens* (humana). 13 confórmeros. Monómero. RMSD: mín 0.38 - máx 2.75 - promedio 1.38. Método de obtención: Difracción de Rayos X. Debido a la existencia de *missing atoms*, se utilizó el software Chimera para poder incorporarlos [36]. Se le quitaron las aguas y el ligando (N-acetil-D-glucosamina).

  - 2CVF: proteina “DNA repair and recombination protein radB” de *Thermococcus kodakarensis*. 4 confórmeros. Monómero. RMSD: mín 1.16 - máx 2.16 - promedio 1.795. Método de obtención: Difracción de Rayos X. Se quitaron las aguas.

  - 1S2O: proteina “sucrose-phosphatase” de *Synechocystis sp*. 9 confórmeros. Monómero. RMSD: mín 0.09 - máx 2 - promedio 0.5167. Método de obtención: Difracción de Rayos X. Se le quitó las aguas, con excepción de aquellas tres moléculas que coordinan con la proteína (números en el pdb: 2003, 2030 y 2036) [37]. Se mantuvo el ion Mg^{2+}.

  - 2V1S: proteina “mitochondrial import receptor subunit TOM20 homolog” de *Rattus novergicus* (rata). 21 confórmeros. RMSD: mín 0.21 -máx 2.67 - promedio 1.089. Método de obtención: Difracción de Rayos X. Si bien el cristal cuenta con 14 cadenas, se ha estudiado la unidad biológica, siendo la cadena A la principal, mientras que la cadena H será considerada como solvente. No se agregaron los residuos perdidos debido a que se encontraban en los extremos de la cadena A. Se han quitado las aguas y el péptido de unión (L-cisteína).

  - 3K8D: proteina “3-deoxy-manno-octulosonate cytidylyltransferase” de *Escherichia coli*. 12 confórmeros. Dímero. RMSD: mín 0.04 - máx 2.22 - promedio 1.37. Método de obtención: Difracción de Rayos X. Se quitaron las aguas. Para poder utilizar el

pdb en GROMACS, se modificó el número de átomo del segundo Mg^{2+} (tiene presente dos iones) ya que coincidía con del primero.

- 1NIW: proteína “calmodulin” de *Rattus novergicus* (rata). 24 confórmeros. Monómero. RMSD: mín 0.33 - máx 3.02 - promedio 2.05. Método de obtención: Difracción de Rayos X. Se quitó las aguas. Se mantuvieron los iones Ca^{2+}.

- Móvil extrema, proteína descripta como intrínsecamente desestructurada (IDP):
  - 2FFT: proteína “thylakoid soluble phosphoprotein” de *Spinacia oleracea* (espinaca). 190 confórmeros. Monómero. RMSD: mín 1.4 - máx 3.16 - promedio 2-33. Método de obtención: **Resonancia Magnética**. Se quitaron las aguas. Se incluye esta proteína como un ejemplo de máxima movilidad.

Para realizar las simulaciones, se ha utilizado la unidad biológica de las proteínas mencionadas. Esto implica que, en algunos casos, fue necesario eliminar o agregar cadenas, puesto que lo que presentaba el cristal no siempre coincide con la unidad biológica, es decir que fue necesario reconstruir la unidad biológica.

Como se mencionó en la introducción, CoDNaS utiliza, para realizar los cálculos, cadenas y no estructuras completas. Es por esto que, si bien las simulaciones se realizan utilizando la estructura proteica que se corresponde con la unidad biológica en una caja de aguas con iones, para poder comparar con los resultados de dicha base de datos, se considera las cadenas de una misma proteína en una misma simulación como elementos de estudio individuales.

Las similitudes y diferencias entre estructuras de la misma proteína están representadas con un dendrograma. Un dendograma es la representación de las diferencias entre objetos, donde la longitud de las líneas o ramas, es proporcional a su distancia (en este caso, la diferencia entre estructuras). Las estructuras más similares, se encuentran más cercanas en las “hojas del árbol” (ver Figura 13). Dichas representaciones muestran las distancias, medidas en Å, luego de una superposición estructural de a pares entre los confórmeros. Dos confórmeros pertenecen a un mismo grupo o *cluster* si sus estructuras no difieren más de 0.4 Å en su superposición estructural y están representados dentro del mismo cuadro rojo en las figuras (ver Figura 14). Por el contrario, estructuras más diferentes (distancia > 0.4 Å), se ven en distintos cuadros rojos en las figuras. En Anexo 1 se presentan los dendrograma de todas las proteínas estudiadas. Se puede observar que en las proteínas móviles, las estructuras se dividen en muchos grupos, mientras que, para las

rígidas, la cantidad de grupos es menor, es decir, los confórmeros se diferencian poco entre sí (diferencia inferior a 0.4 Å).

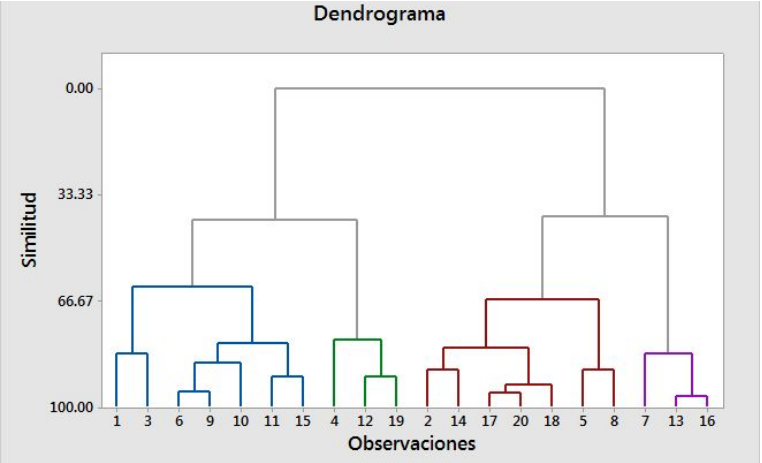

Figura 13. Ejemplo de dendrograma. Hacia las “hojas del árbol”, las estructuras son más similares entre sí. Ej: las rojas son más similares entre sí que con las azules, verdes o violetas. Y en conjunto, las rojas son más similares a las violetas que a azules.

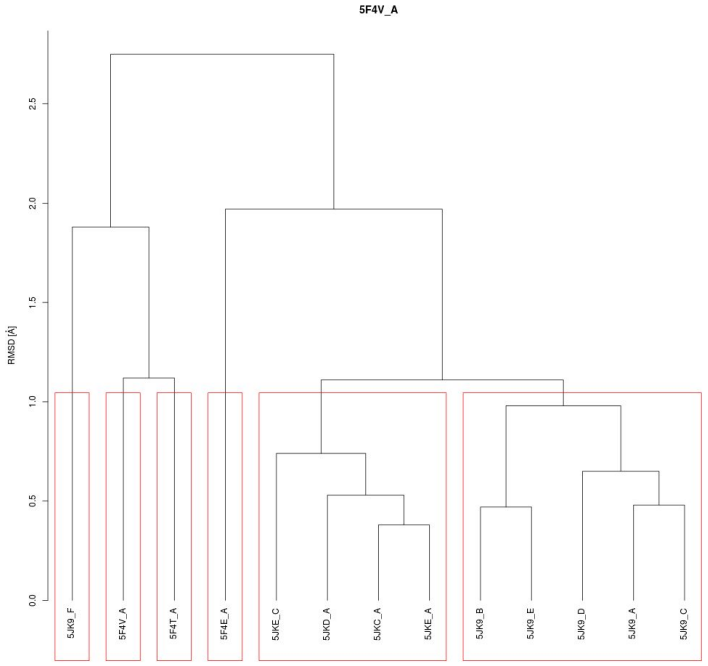

Figura 14. Dendrograma de la proteína con código PDB 5F4V. Luego de la superposición en el espacio de los distintos confórmeros de la misma proteína, todos los que se diferencian menos de 0.4 Å quedan en el mismo grupo (encuadrados en rojo). Imagen obtenida de http://ufq.unq.edu.ar/codnas/index.php.

Dinámica Molecular

Para llevar a cabo las simulaciones de DM de 200ns, se ha utilizado el herramienta de simulación GROMACS.

En Anexo 2 se pueden observar los parámetros establecidos para cada momento de la simulación [38].

Análisis de simulaciones

Una vez realizadas las simulaciones, utilizando el programa GROMACS se obtuvo, para cada una de ellas, la siguiente información:

- RMSD del backbone: *gmx rms* compara dos estructuras al calcular la desviación cuadrática media (*Root-Mean-Square Deviation*) [39]. Cada estructura de una trayectoria se compara con la estructura inicial y con la estructura de 50ns, donde se considera que todas las proteínas están estabilizadas.

- Matriz RMSD: *gmx rms* también permite obtener una matriz de comparación entre una trayectoria consigo misma [39]. De esta forma, se puede ver la semejanza/diferencia entre las estructuras obtenidas durante la simulación de DM. Una vez obtenida la matriz, se la normalizó entre 0 y 1 para obtener la misma distribución de colores en todos los casos. Normalizada, *gmx xpm2ps* crea la matriz a color. Se generaron matrices a partir de la estructura del tiempo 50ns.

- Radio de giro: El radio de giro de una proteína se define como la distancia cuadrática media de cada átomo de la proteína a su centroide. *gmx gyrate* calcula el radio de giro de una molécula en función del tiempo.

- Grupos: *gmx cluster* puede agrupar estructuras usando varios métodos. Las distancias entre estructuras se determinaron a partir de las matrices obtenidas. Se utilizó el método *gromos*, el cual usa el algoritmo descrito por Daura y col. [40]. Éste, cuenta el número de vecinos que se encuentran a cierta distancia (punto de corte o *cut off*). Se simuló con diferentes puntos de corte: 0.20, 0.25 y 0.30 nm. Los archivos de salida obtenidos informan

sobre la cantidad de grupos para cada punto de corte. También se compara contra la estructura del tiempo cero y con el tiempo 50ns.

Esta función genera también un archivo de coordenadas con múltiples modelos, donde cada modelos es un representante de un grupo. Se seleccionaron las proteínas móviles 5F4V, 2V1S, 1NIW y 2CVF para representar visualmente con VMD [41] los modelos obtenidos a partir de 50 ns con un punto de corte de 0.30 nm. Para esto se usó el script splitmultiframepdb.tcl para VMD que toma el archivo de coordenadas con los distintos modelos y los representa visualmente [42].

- Factor B: El factor de temperatura (también llamado valor de temperatura, factor B, valor B o factor de Debye-Waller), brevemente, es un factor que puede aplicarse al término de dispersión de rayos X para cada átomo (o para grupos de átomos) que describe el grado de dispersión de la densidad electrónica. Mientras que la teoría es que el factor B indica la verdadera movilidad estática o dinámica de un átomo, también puede indicar dónde hay errores en la construcción del modelo o regiones flexibles de la proteína.

  El factor B se da por: $Bi = 8\pi^2 U_i^2$, donde U_i^2 es el desplazamiento cuadrado medio del átomo i. A medida que U aumenta, el factor B aumenta y la contribución del átomo a la dispersión disminuye. Si los átomos están posicionados incorrectamente en el modelo, sus factores B tenderán a ser más altos que el de los átomos ubicados correctamente. Los factores B pueden tomarse como indicadores del movimiento vibratorio relativo de diferentes partes de la estructura. Los átomos con bajos factores B pertenecen a una parte de la estructura que está bien ordenada. Los átomos con grandes factores B generalmente pertenecen a una parte de la estructura que es muy flexible. Cada registro ATOM (formato de archivo PDB) de una estructura cristalina depositada en el banco de datos de proteínas contiene un factor B para ese átomo. Se mide en unidades de $Å^2$.

  gmx rmsf calcula la *root mean square fluctuation* (RMSF) durante la trayectoria para cada átomo, obteniendo valores respecto de la fluctuación alrededor de una posición promedio. A su vez, también permite convertir estos valores de RMSF en factores B que son comparables al factor B o factor de temperatura cristalográfico. Las imágenes que representan los factores B fueron realizadas con Pymol (TM) [43].

Gráficos

Los gráficos presentados en este trabajo fueron elaborados con QtGrace, versión de Grace para Windows, Open Source Pymol para linux, Visual Molecular Dynamics (VMD) para linux versión 1.9.3 y con MATLAB, versión R2016a.

Resultados y Discusión

La dinámica de las estructuras en función del tiempo permite evaluar distintos parámetros:

RMSD

En bioinformática, la desviación cuadrática media (RMSD) de las posiciones atómicas de una macromolécula, es la medida de la distancia promedio entre los átomos (generalmente los átomos del *backbone*) de las moléculas superpuestas, en este caso, proteínas [44].

En las comparaciones de conformaciones (estructuras en general) de proteínas, se mide habitualmente la similitud en la estructura tridimensional por el RMSD de las coordenadas atómicas de los Cα, después de la superposición óptima de cuerpo rígido [44]. En el caso del RMSD en DM, se analiza la diferencia entre las proteínas de los diferentes instantes de tiempo con respecto a un estado de referencia definido por el usuario, como por ejemplo, la situación inicial, la proteína en el instante anterior o la estructura del instante en que se considera que está equilibrada que está equilibrada, una estructura promedio. Los instantes o unidades de tiempo mínimas de muestreo son decididos por el usuario antes de la simulación.

Se espera que al superponer estructuras de proteínas rígidas tomadas a distintos puntos de tiempo de la trayectoria, no se vean grandes diferencias, ya que una proteína ya clasificada como rígida no ha mostrado diferencias entre las estructuras de los distintos confórmeros que están depositados en el PDB.

Por el contrario, se espera que las estructuras tomadas a distintos puntos del tiempo de la trayectoria de proteínas móviles, tengan mayores diferencias en su RMSD ya que los distintos confórmeros depositados en el PDB dan muestras de la mayores diferencias entre ellas.

En la Figura 15.a y 15.b se pueden visualizar las curvas de RMSD respecto a la estructura inicial de la simulación en función de las estructuras obtenidas en la trayectoria a lo largo del tiempo de simulación para las proteínas rígidas y para las proteínas móviles, respectivamente. En la Figura 15.c se incorpora la proteína 2FFT (IDP), la cual, al tener valores tan alejados respecto al resto de las proteínas móviles, interfiere en la presentación de los resultados de la Figura 15.b, impidiendo poder llevar a cabo una comparación con la Figura 15.a. En la Figura 16.a y 16.b se pueden visualizar las curvas de RMSD respecto a la estructura de 50 ns (considerada equilibrada) en

función del tiempo de simulación de las proteínas rígidas y de las proteínas móviles, respectivamente.

Las curvas de RMSD se presentan desde el nanosegundo 50, instante de tiempo donde se puede considerar que los sistemas se encuentran equilibrados (un cuarto del tiempo total de simulación). En ambas figuras se puede observar una amplitud mayor de las oscilaciones (que son la diferencia en distancia del *backbone* al superponer dos estructuras) de las proteínas móviles respecto a las rígidas. Esto significa que los confórmeros se diferencian más entre sí en el primer grupo de proteínas. En las rígidas, los *backbone* permanecen más inmóviles, lo que se traduce en oscilaciones pequeñas del RMSD (Figura 15.a y Figura 16.a). La altura en que se encuentran las curvas, no es relevante ya que solo da cuenta de las diferencias de esas estructuras respecto a la estructura de referencia (llamada estructura semilla o *"seed"*). El dato relevante es la oscilación de la curva una vez que el sistema está equilibrado.

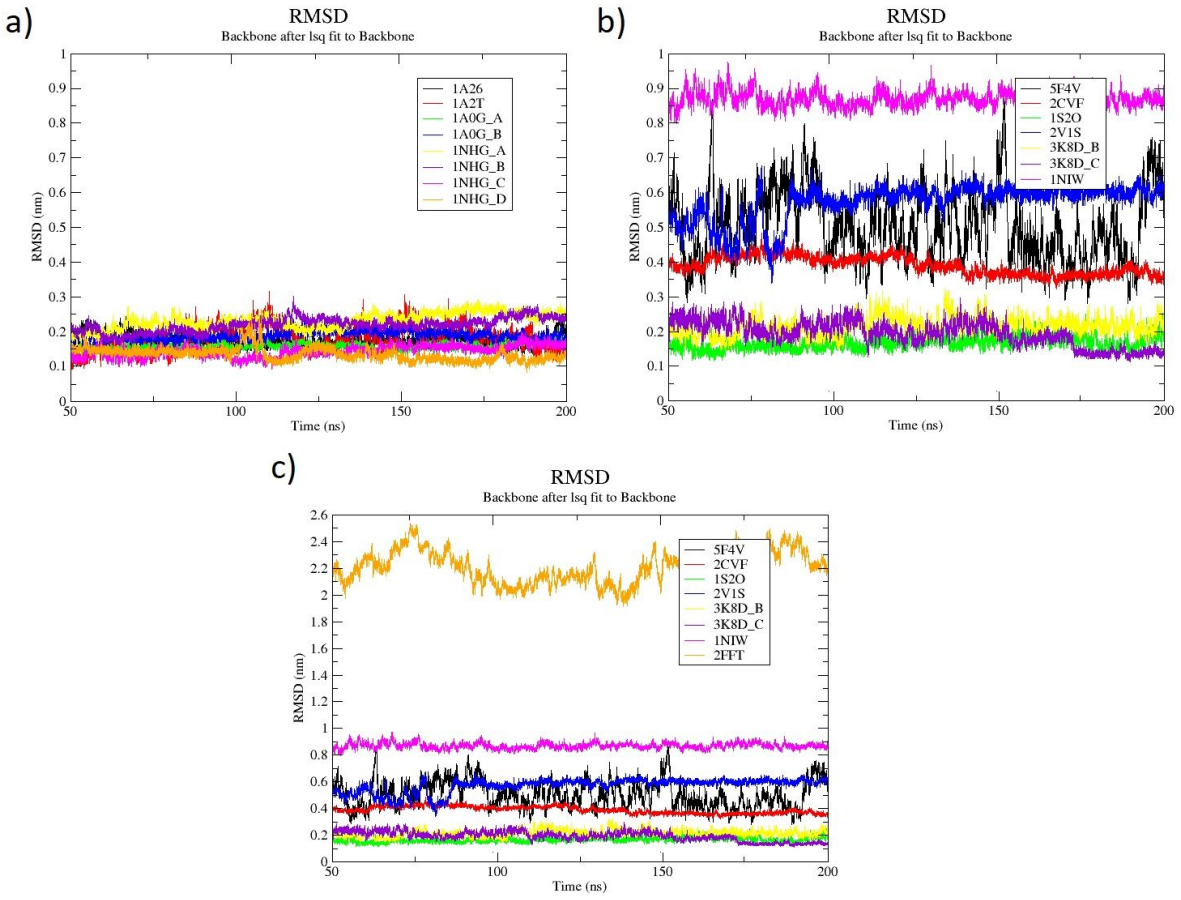

Figura 15. RMSD en función del tiempo de simulación de las proteínas respecto a estructura inicial. a) Proteínas rígidas. b) Proteínas móviles sin 2FFT (IDP). c) Proteínas móviles con 2FFT (IDP).

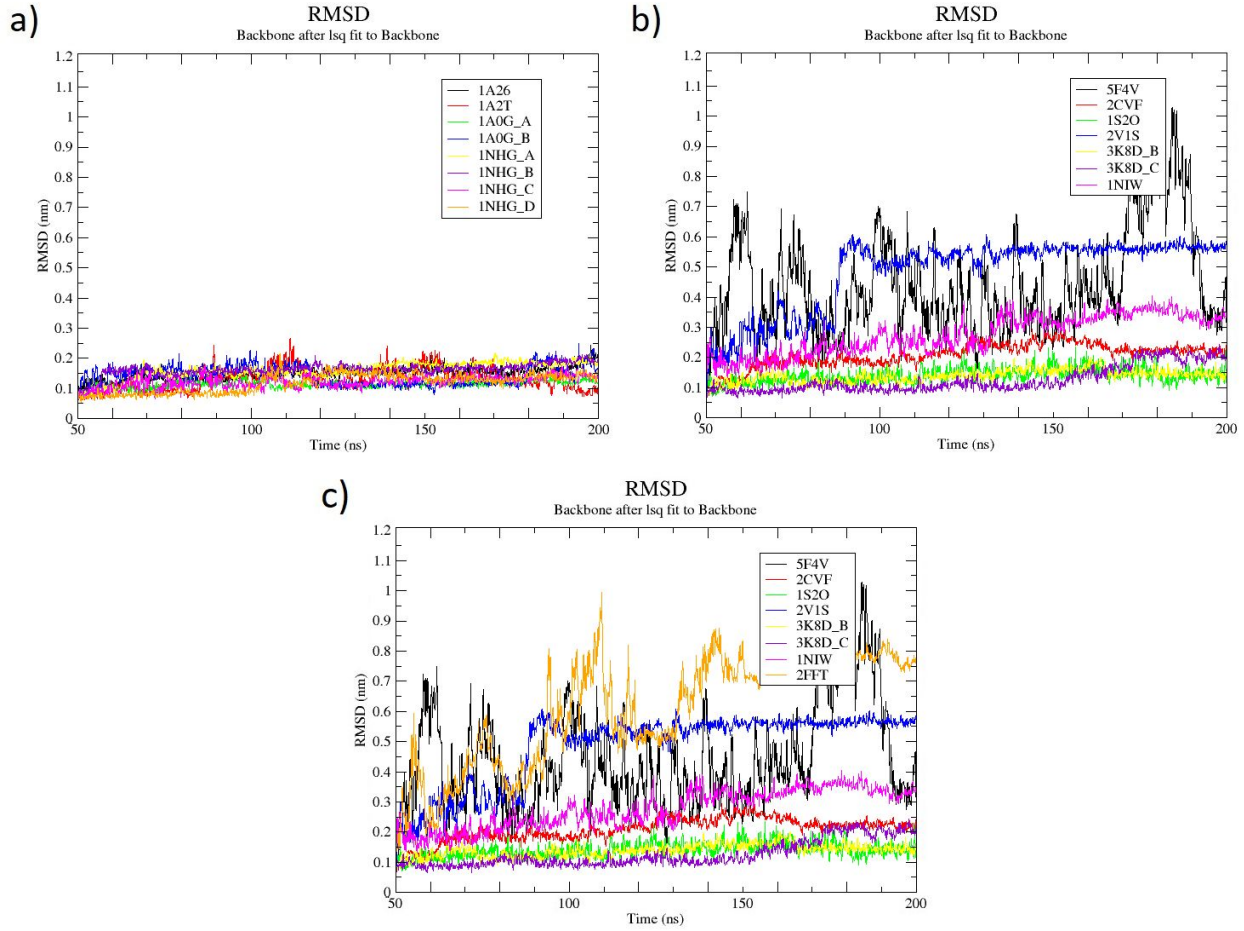

Figura 16. RMSD en función del tiempo de simulación de las proteínas respecto a la estructura de tiempo 50ns. a) Proteínas rígidas. b) Proteínas móviles sin 2FFT (IDP). c) Proteínas móviles con 2FFT (IDP).

Otra manera de representación de la RMSD son los gráficos de violín (*violin plots*) que se observan en las Figuras 17 y 18. Los gráficos de violín son una combinación entre los diagramas de cajas y bigotes *(box plots)* y los gráficos de densidad (*density plots)*. Nos dan una idea más certera acerca de la distribución de los valores (densidad) con el ensanchamiento de la caja. Además, permiten visualizar si la distribución es homogénea, bimodal o multimodal.

Las figuras 17 y 18 muestran la distribución de RMSDs a lo largo de la corrida respecto a la estructura inicial y a la estabilizada (a los 50 ns), respectivamente. Se pueden observar varias diferencias entre los grupos de proteínas: las proteínas rígidas (Figura 17.a y 18.a) oscilan en un

rango muy acotado, entre 0.05 y 0.25 nm, es decir, la diferencia entre los valores máximos y mínimos (incluyendo los outliers) es menor a 0.25 nm. También se observa que la distribución de valores es bastante similar a una distribución normal (forma del violín). Por el contrario, en la Figura 17.b y Figura 18.b se puede observar que la variación de los valores de RMSD de las proteínas móviles es en un rango mayor, entre 0.1 y 1 nm, y que la distribución de valores no es tan homogénea, mostrando algunas distribuciones más irregulares.

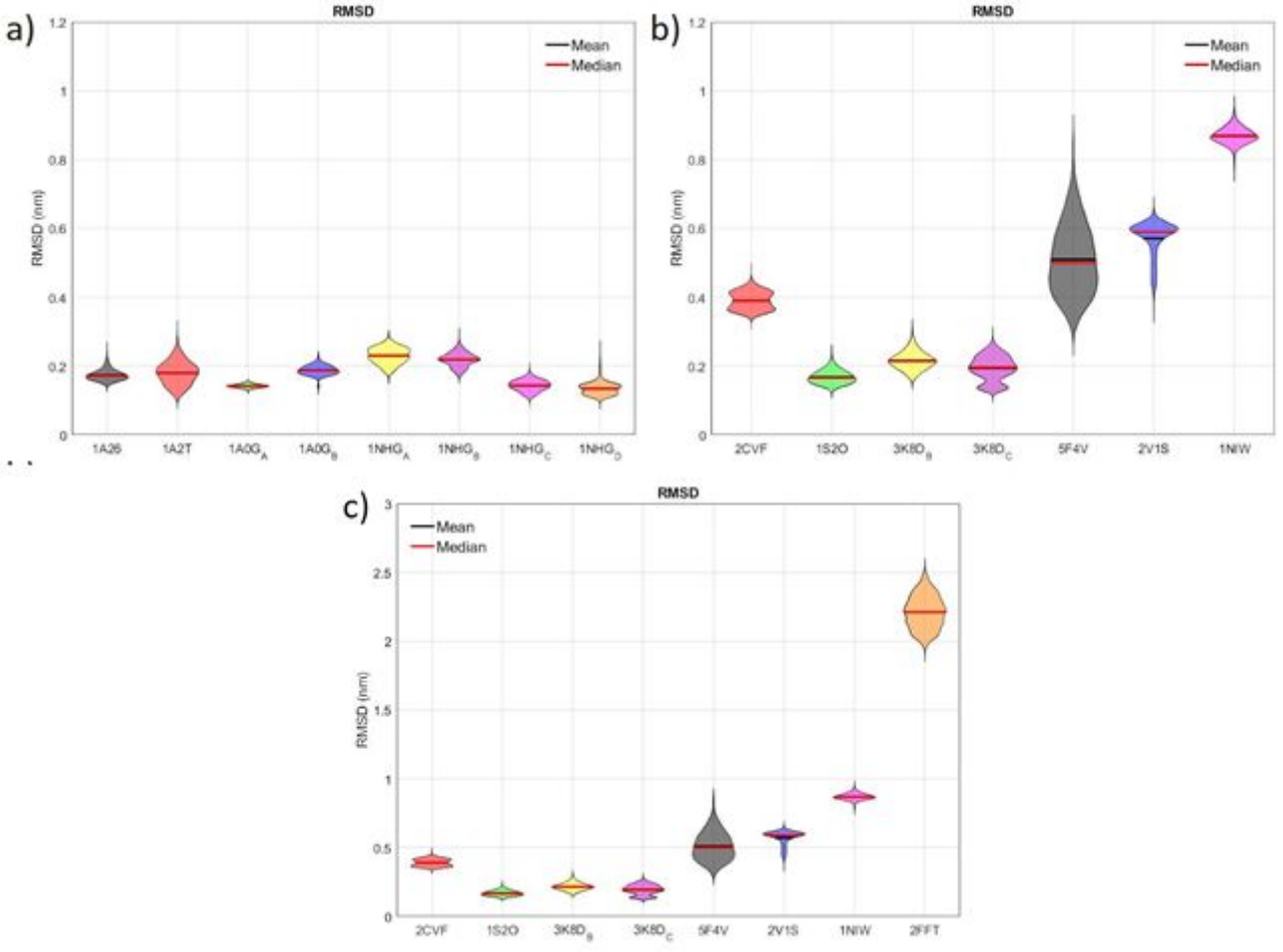

Figura 17. Representación en Box Plot de violín, de los valores de RMSD obtenidos respecto a estructura inicial. a) Proteínas rígidas. b) Proteínas móviles sin 2FFT (IDP). c) Proteínas móviles con 2FFT (IDP).

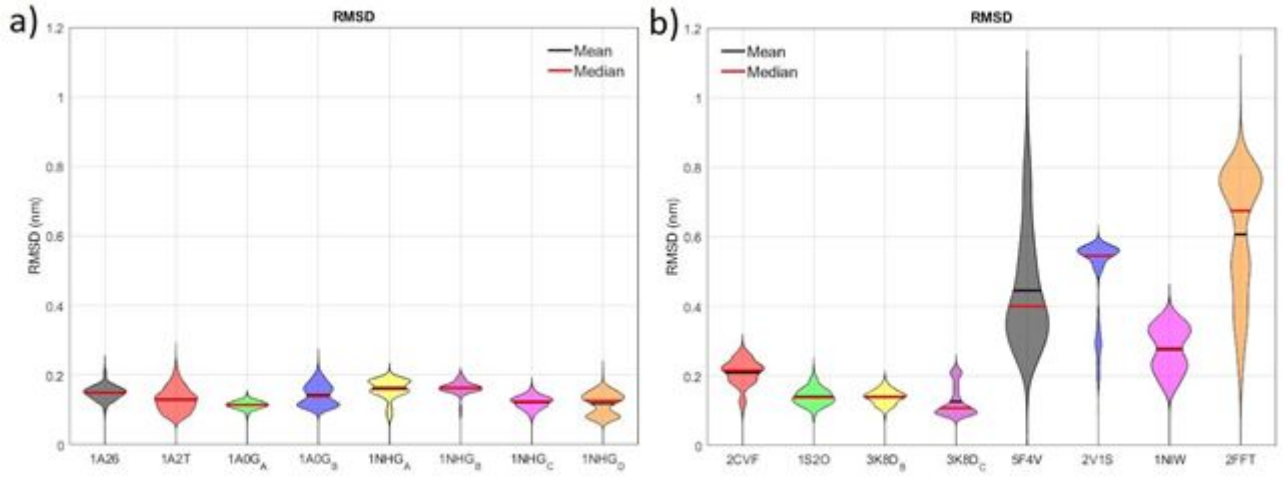

Figura 18. Representación en Box Plot de violín, de los valores de RMSD obtenidos respecto a la estructura a los 50 ns. a) Proteínas rígidas. b) Proteínas móviles con 2FFT (IDP).

Dentro del set de proteínas móviles, se puede diferenciar dos subgrupos:

- proteínas cuyos valores de RMSD tienen oscilaciones mayores a 0.25 nm (tomando como valores iniciales y finales las punta del violín): 5F4V, 2V1S, 1NIW y 2FFT,
- proteínas cuyos valores de RMSD tienen oscilaciones alrededor de 0.25 nm: 2CVF, 1S2O, 3K8D - cadena B y C.

Este último subgrupo muestra un comportamiento similar al de las proteínas rígidas (Figura 17.a y 18.a). Para entender este comportamiento, miramos en detalle las estructuras en cada uno de los casos. Las observaciones que se realizan en esta sección y las hipótesis acerca de los mecanismos que generan estos comportamientos, será aplicable a los resultados obtenidos por los distintos métodos.

La proteína 1S2O se incluyó en el set de proteínas móviles dado que tiene 9 confórmeros y que los RMSD mínimo y máximo son de 0.09 y 2 Å respectivamente. Si bien posee 9 confórmeros, se agrupan en 2 grupos, uno con una única estructura y otro con las 8 restantes. Mirando la tabla de RMSD de a pares de todos los confórmeros con todos, vemos que 1S2O_A (confórmero sólo en un grupo) tiene diferencias superiores a 1 Å con todas las demás estructuras, mientras que las otras entre sí tienen diferencias menores a 0.2 Å. Esto sugiere que probablemente 1S2O sea una proteína rígida y que una de las estructura se ha diferenciado de las demás debido a algún error o

condición experimental particular. Otra posible hipótesis es que estamos frente a una proteína que tiene solo dos posiciones posibles, y que no contamos con suficientes confórmeros de una de las posiciones para asegurarlo. Pero en los dos escenarios, se trataría de una proteína rígida.

3K8D posee 12 confórmeros con RMSD min, max y promedio de 0.04, 2.2 y 1.37 Å respectivamente, por lo que siendo ciego a otros datos, es incluída en el grupo de proteínas móviles. Sin embargo, un análisis en detalle nos permite ver que al superponer los 12 confórmeros, se trata de una proteína que sólo tiene 2 conformaciones posibles (ver Figura 19). Tiene un subdominio fijo, y el otro se mueve en forma de bisagra como cuerpo rígido entre dos posiciones. Algunos confórmeros están en una y los demás en la otra, pero dentro de cada posición son rígidos (todos, independiente del PDB que vengan, se superponen muy bien). Dependiendo de la estructura que se haya usado para iniciar la dinámica, se quedará en su estructura posible sin moverse de ella. En este caso podemos asegurar que se trata de una proteína con posiciones fijas, sin posibilidad de moverse libremente en un espacio continuo. Cada confórmero es rígido en su posición.

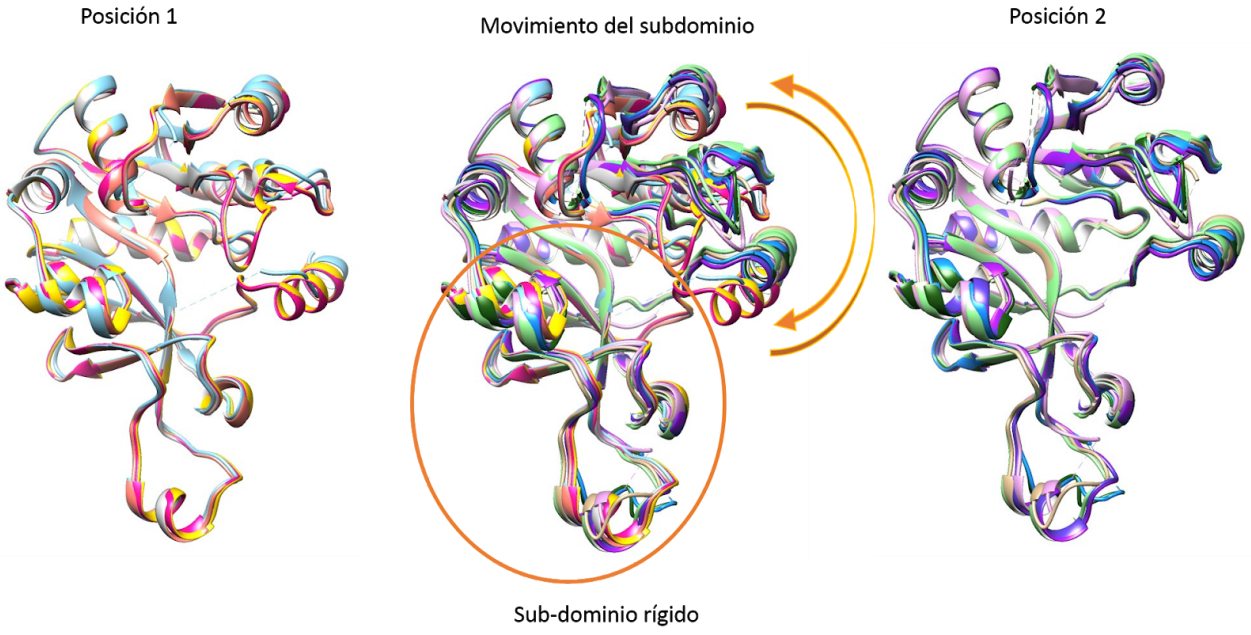

Figura 19. Panel central: superposición de los 12 confórmeros proveniente de los PDBs 1NH1, 3K8D y 3K8E, que son todas las estructuras posibles de la proteína 3-deoxy-manno-octulosonate cytidylyltransferase. Panel izquierdo: confórmeros en posición 1. Panel derecho: confórmeros en posición 2. El círculo muestra el subdominio que se mantiene fijo en los 12 confórmeros. Las flechas indican el movimiento en forma de bisagra del otro subdominio. Imagen realizada con Chimera [5].

Analizando la proteína 2CVF, la superposición de sus confórmeros revela que aunque es rígida en su mayoría, posee 2 loops de gran movilidad (ver Figura 20). La movilidad del loop afecta el valor de RMSD.

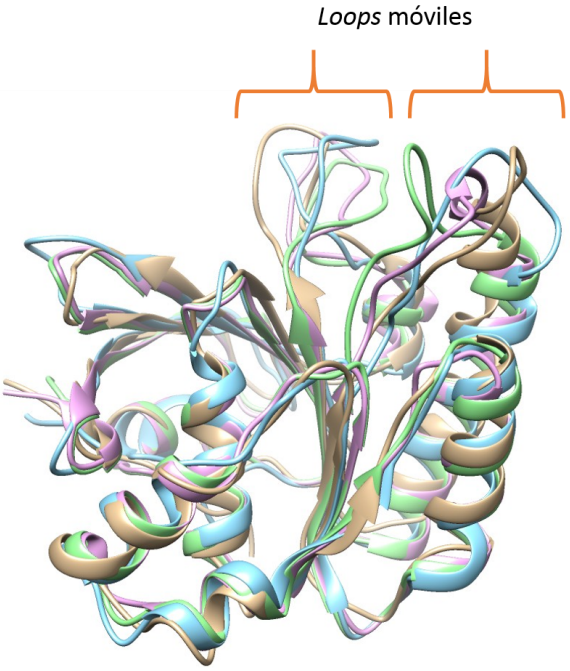

Figura 20. Superposición de los 4 confórmeros provenientes de los PDBs 2CVF y 2CVH. Imagen realizada con Chimera [5].

Solo por comparación, la figura 21 muestra un ejemplo de proteína rígida y uno de proteina móvil. En la rígida (panel izquierdo), a pesar de poseer loops muy extensos, los distintos confórmeros de la proteína se superpone en su totalidad. En el panel derecho se puede observar que, en las proteínas que llamamos móviles, la superposición de los confórmeros no es perfecta, sino que los confórmeros recorren el espacio de manera continua dentro de un rango de Armstrongs.

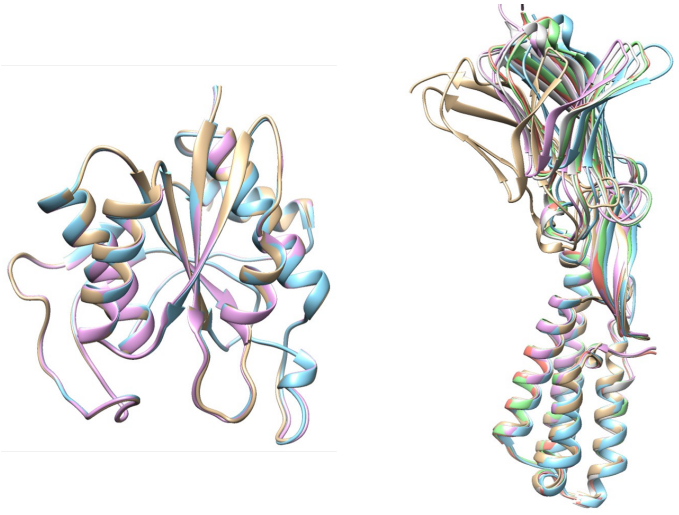

Figura 21. Panel izquierdo, ejemplo de proteína rígida: superposición de 3 confórmeros provenientes de los PDBs 1AUG, 3RNZ y 3RO0. Panel derecho, ejemplo de proteína móvil: superposición de los conformeros proveniente de los cristales 5F4V; 5F4E; 5jk9; 5JKC; 5JKD y 5JKE. Imagen realizada con Chimera [5].

Matriz de RMSD

Otra representación de las diferencias en valores de RMSD es la “Matriz RMSD”: una manera de ver las diferencia entre todas las estructuras de una DM (en lugar de comparar siempre con la estructura de referencia), es superponer las estructuras presentes en cada momento de la DM. Aquí, cada estructura se compara con las estructuras obtenidas para cada tiempo, así los ejes X e Y se representan de 0 a 150 ns (de 0 a 150000 ps en la figura). En la diagonal cada estructura de cada tiempo se compara con sí misma por lo tanto el color de la diagonal es azul aunque por el tamaño del gráfico a veces no llegue a distinguirse en las proteínas muy móviles. Como es de esperar la imagen es simétrica, es decir que la información que se encuentra por arriba de la diagonal es la misma que la que se encuentra por debajo. En cada punto el color representa las diferencias entre las estructuras de cada tiempo. La Figura 22 permite observar las matrices RMSD normalizadas y representadas a color de las proteínas rígidas (Figura 22.a) y proteínas móviles (Figura 22.b). En el eje X e Y está representado el tiempo de la trayectoria de DM (ps). El color de la intersección de dos puntos de tiempo es un indicador de la distancia en Armstrongs al superponer las estructuras presentes en ese tiempo de simulación.

Con esta comparación, otra vez vemos que las diferencias entre los confórmeros de las proteínas móviles es mayor a la de las rígidas.

Estos resultados son llamativos ya que en las simulaciones de DM no se incluye información acerca de las diferencias entre confórmeros que se han hallado al cristalizar la proteína más de una vez. Esto deja ver que al partir de una única estructura, las proteínas denominadas “rígidas” por la determinación estructural experimental, recorren un espacio conformacional menor al simular una trayectoria mediante dinámica molecular. Mientras que las proteínas que consideramos móviles, demuestran visitar conformaciones más diferentes. Estos son los primeros indicios de que se podría predecir teóricamente, partiendo de una estructura, a qué grupo pertenece esa proteína.

Al igual que en el apartado anterior, se puede distinguir los mismos subgrupos entre las proteínas móviles: para 5F4V, 2V1S, 1NIW y 2FFT, en sus matrices predominan los colores de verde a rojo (valores de RMSD superiores a 0.40 nm), mientras que para 2CVF, 1S2O y 3K8D (ambas cadenas), los colores de las matrices se encuentran, principalmente, entre los azules y celestes (valores de RMSD inferiores a 0.40 nm). Sin embargo hay cierta diferencia notoria con las proteínas rígidas que deberá ser estudiada con mayor profundidad, pero que escapa al alcance del presente trabajo.

Es importante destacar que la matriz de 1S2O (Figura 22.b.b) es, en gran proporción, de color azul, como ocurre con las rígidas (valores de RMSD inferiores a 0.30 nm). Este comportamiento (más similar a las proteínas rígidas que a las móviles) se puede explicar, como se mencionó antes, con el hecho que, como se ve en su dendrograma de grupos (Anexo 1), aunque el RMSD máximo es grande (lo que ha hecho que se considere dentro del grupo de “móviles”), el grupo mayoritario contiene todas las estructuras excepto una. Esta última, podría ser el resultado de condiciones muy alejadas a las biológicas o que, dado una circunstancia particular del experimento que se hizo para su cristalización, esa estructura haya adoptado una conformación muy diferente a las nativas.

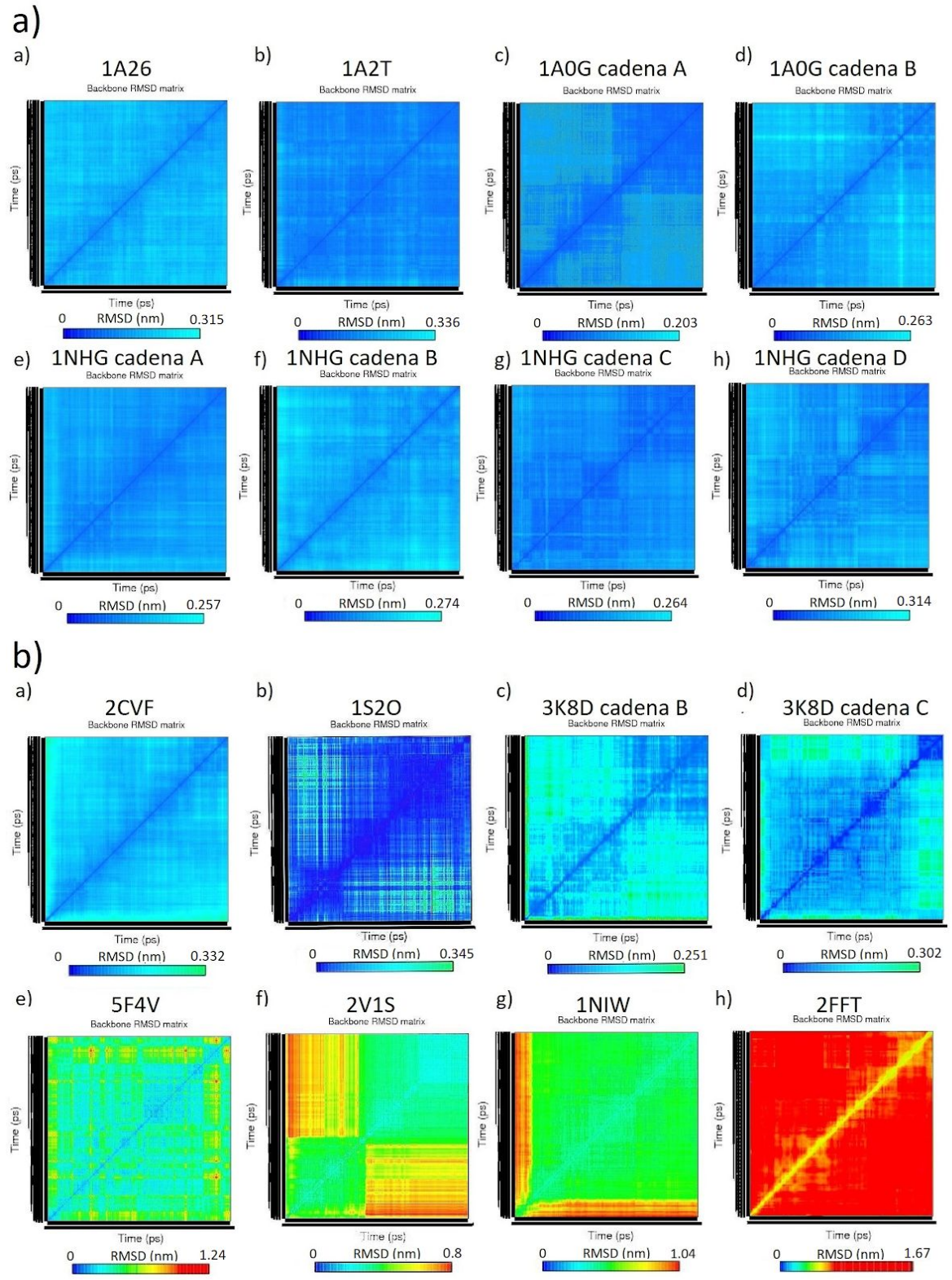

Figura 22. Matrices RMSD normalizadas y a color a partir de 50ns. a) Proteínas rígidas b) Proteínas móviles (2FFT es una IDP).

Radio de giro (Rg)

El radio de giro es una manera de medir la compactación de la estructura de la proteína: un valor bajo de Rg sugiere una proteína más compacta y de manera opuesta, un valor alto de Rg sugiere un compactamiento más relajado [45]. Un valor constante de Rg a lo largo de las trayectorias de DM muestra que la estructura espacial de la proteína es estable a lo largo de la DM, por el contrario, proteínas en movimiento darían un valor cambiante de Rg a lo largo de la trayectoria.

Se representó el radio de giro de las proteínas rígidas y móviles en las Figuras 23 y 24.

El radio de giro de las distintas proteínas va entre 1 y 3 nm. Esto se debe a que el Rg depende del tamaño de la proteína, al no estar normalizado por el número de aminoácidos (proteínas más grandes tienen Rg más grandes). Al igual que el RMSD, lo destacable para analizar es la amplitud de los cambios representados como trayectorias (Figura 23) y gráficos de violín (Figura 24).

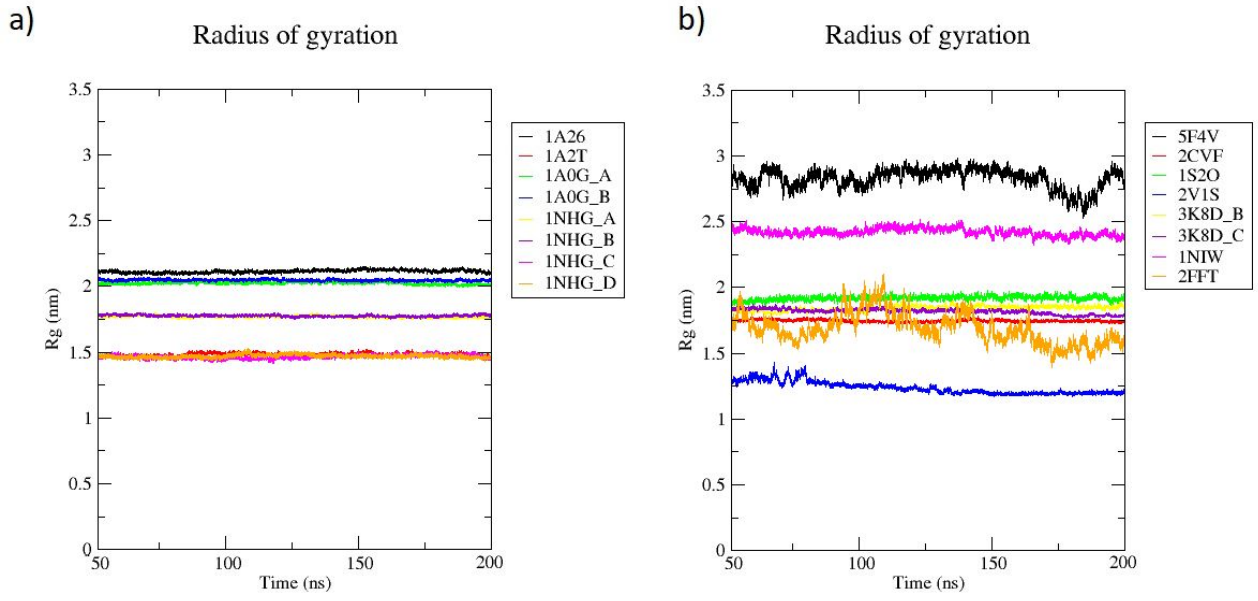

Figura 23. Radio de giro en función del tiempo de las proteínas estudiadas. a) Proteínas rígidas. b) Proteínas móviles.

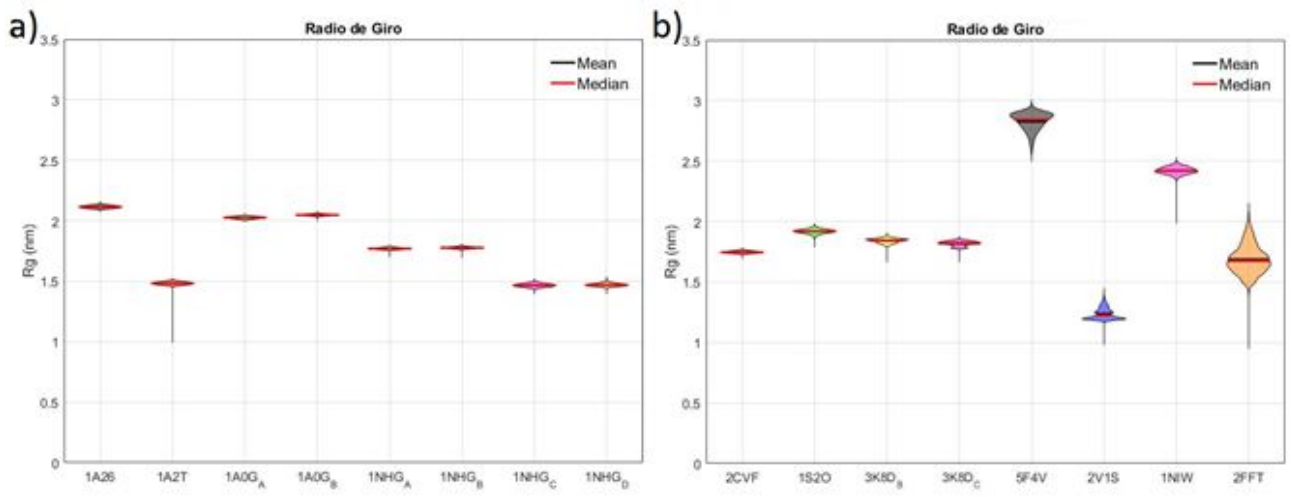

Figura 24. Representación en Box Plot de violín, de los valores de Radio de Giro obtenidos. a) Proteínas rígidas. b) Proteínas móviles.

En las figuras se puede observar que las proteínas móviles tienen una mayor fluctuación de los Rg que las rígidas. El rango de los Rg de todas las rígidas (Figura 24.a) es inferior a 0.20 nm. En cambio, respecto a las móviles (Figura 24.b), otra vez, consistentemente con los resultados anteriores, se diferencian los mismos dos subgrupos: cuatro proteínas (2CVF, 1S2O y 3K8D, cadena B y C) tienen valores de Rg con oscilaciones del mismo rango que las rígidas (0.20 nm).

Agrupamiento de estructuras

El agrupamiento de proteínas se utiliza para asociar aquellas proteínas similares [46]. Las proteínas que difieren en una distancia menor al valor de punto de corte establecido al superponerlas estructuralmente, serán consideradas del mismo grupo (estructuras vecinas). Si por el contrario, difieren en más, se ubicarán en grupos diferentes. En DM, este agrupamiento permite reducir la cantidad de representaciones de la proteína simulada en los diferentes instantes de tiempo, permitiendo observar cuánto varía dicha molécula a lo largo de la ejecución de la dinámica.

Cuanto mayor sea la cantidad de grupos encontrados en una simulación, mayor movilidad tiene la proteína a lo largo del tiempo adopta diferentes disposiciones espaciales. Una proteína rígida, con poca movilidad, conlleva a que las diferentes “instantáneas” de la simulación sean muy similares entre sí, obteniéndose poca cantidad de grupos.

Para presentar los resultados de una manera clara, en la Tabla 1 se detalla la cantidad de grupos obtenidos para cada punto de corte (0.20, 0.25 y 0.30 nm) por cada proteína para el tiempo de referencia 0 (Tabla 1.a) y 50 ns (Tabla 1.b).

a

Rígidas, tiempo 0ns	1A26	1A2T	1A0G_A	1A0G_B	1NHG_A	1NHG_B	1NHG_C	1NHG_D
Punto de corte (nm)								
0.20	2	2	1	1	2	2	2	2
0.25	1	1	1	1	1	1	1	1
0.30	1	1	1	1	1	1	1	1
Móviles, tiempo 0ns	**5F4V**	**2V1S**	**1NIW**	**2FFT**	**3K8D_B**	**3K8D_C**	**1S2O**	**2CVF**
Punto de corte (nm)								
0.20	86	70	60	695	4	3	3	2
0.25	38	28	30	446	1	1	1	1
0.30	19	13	18	207	1	1	1	1

b

Rígidas, tiempo 50ns	1A26	1A2T	1A0G_A	1A0G_B	1NHG_A	1NHG_B	1NHG_C	1NHG_D
Punto de corte (nm)								
0.20	2	1	1	1	1	1	1	1
0.25	1	1	1	1	1	1	1	1
0.30	1	1	1	1	1	1	1	1
Móviles, tiempo 50ns	**5F4V**	**2V1S**	**1NIW**	**2FFT**	**3K8D_B**	**3K8D_C**	**1S2O**	**2CVF**
Punto de corte (nm)								
0.20	50	31	14	263	1	1	1	3
0.25	25	10	6	164	1	1	1	1
0.30	11	6	3	81	1	1	1	1

Tabla 1. Cantidad de grupos encontrados por proteína según el punto de corte utilizado. a) Tomando como referencia el tiempo 0ns. b) Tomando como referencia el tiempo de 50ns.

Como se esperaba, las proteínas rígidas presentan uno o dos grupos (según el punto de corte utilizado), característica de su baja movilidad durante sus simulaciones de DM tanto para el tiempo de referencia 0 ns como para 50 ns. Por otro lado, respecto a las proteínas móviles, nuevamente se pueden distinguir dos grupos: cuatro proteínas con una cantidad de grupos muy similar a las rígidas, y cuatro proteínas con varios grupos.

En la Figura 25 se muestran los modelos obtenidos para cada grupo, punto de corte de 0.30 nm, en distintos colores de las proteínas 5F4V, 2V1S, 1NIW y 2CVF. Las primeras tres pertenecen al segmento de proteínas móviles (con más de un grupo) mientras que la proteína 2CVF representa al segmento de proteínas asignadas a móviles pero con un solo grupo, similar a las rígidas. La proteína 2FFT se excluye de esta figura por el exceso de modelos que presenta.

Este resultado, al igual que la matriz de RMSD no deja de sorprender ya que, al agrupar millones de estructuras producto de una DM de 200 ns, se obtienen pocos grupos con las proteínas rígidas y un número significativamente mayor con las proteínas móviles, sin haber agregado información dinámica ni estructural de la proteína, más que las coordenadas de un solo confórmero. Esto reproduce teóricamente el resultado experimental.

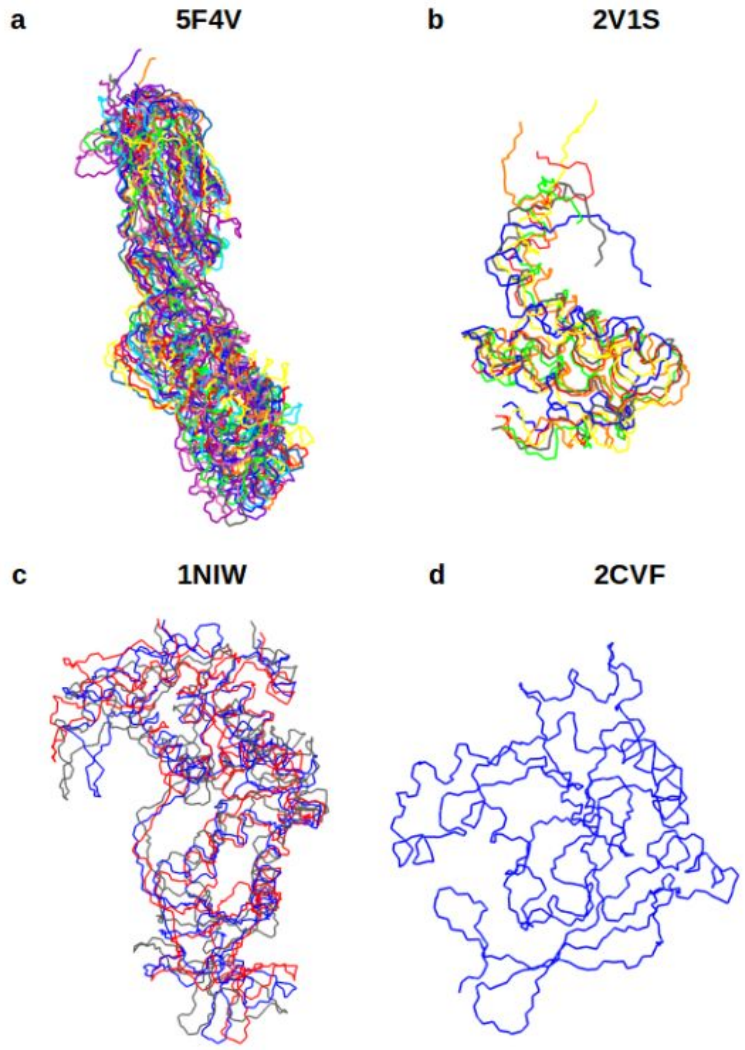

Figura 25. Representación visual del *backbone* los distintos modelos obtenidos para cada grupo obtenido para las proteínas a) 5F4V (11 modelos), b) 2V1S (6 modelos), c) 1NIW (3 modelos) y d) 2CVF (1 modelos), tomando el radio de corte en 0.30 nm y tiempo de referencia 50ns. Imágenes obtenidas con VMD.

Factor B

El factor B representa un valor de la movilidad de los residuos presentes en una proteína. Como mencionamos, los factores B pueden tomarse como indicadores del movimiento de los átomos en la estructura y, por consiguiente, con su flexibilidad.

Para las proteínas estudiadas, el factor B de temperatura de sus carbonos alfas se encuentra dentro del rango 1-600. Debido a esto, se configuró 1 como límite inferior y 600 como límite superior. A continuación se indica el número de aminoácidos de las proteínas con valores por fuera del rango 1-600 para sus carbonos alfa que generalmente coinciden con sus extremos amino o carboxilo terminal:

- 5F4V (móvil): 235 aminoácidos, tiene 8 aminoácidos (3%) por fuera del rango.
- 2V1S (móvil): 73 aminoácidos, tiene 5 aminoácidos (7%) por fuera del rango.
- 1NIW (móvil): 296 aminoácidos, tiene 3 aminoácidos (1%) por fuera del rango.
- 2CVF (móvil): 220 aminoácidos, tiene 2 aminoácidos (1%) por fuera del rango.
- 2FFT (IDP): 84 aminoácidos, tiene 70 aminoácidos (83%) por fuera del rango (este caso especial se incluye para ver un comportamiento extremo de movilidad)

En las figuras 26 a 41 Se representan los histogramas de los factores B y la representación en cinta (*cartoon*) de la estructura correspondiente.

En la Figura 26 se representa el histograma obtenido para 2FFT (IDP), proteína con alta movilidad. Se puede observar que 2FFT tiene un comportamiento extremo donde el 83 % de sus aminoácidos están por fuera del rango elegido para todas las otras proteínas (1-600).

En la Figura 27 se muestra el histograma obtenido para la proteína móvil (5F4V). Esta proteína si bien comparte el rango (1-600) el resto de las proteínas incluso con las proteínas rígidas, la mayoría de sus aminoácidos están por encima del valor de 100. Los valores bajos en el rango de 0-100 (7%) los comparte con otras proteínas móviles como 2V1S (34%, Figura 29) y 1NIW (57%, Figura 29). Las figuras 33 a 40 muestran los histogramas obtenidos para las proteínas rígidas, en todos los casos el porcentaje de aminoácidos con factores B dentro del rango 0-100 va de 83% (1NHG_A) a 97% (1A0G_A y 1A0G_B), mostrando que son todas proteínas rígidas. En cambio, las figuras 30, 31, 32 y 33 muestran los histogramas obtenidos para las proteínas 1S2O, 2CVF, 3K8D_B y 3K8D_C respectivamente que se esperaba fueran móviles, sin embargo sus factores B

se encuentran, al igual que en las rígidas, mayoritariamente en el rango 0-100 (de 83% 2CVF a 95% 3K8D_A). Resultados que confirman los datos obtenidos previamente.

Junto a los gráficos de histogramas, se presentan representaciones de cinta de las proteínas coloreadas por su factor B, donde la escala de colores representa el factor B calculado para cada aminoácido de la cadena proteica: de azul a rojo, según valores crecientes.

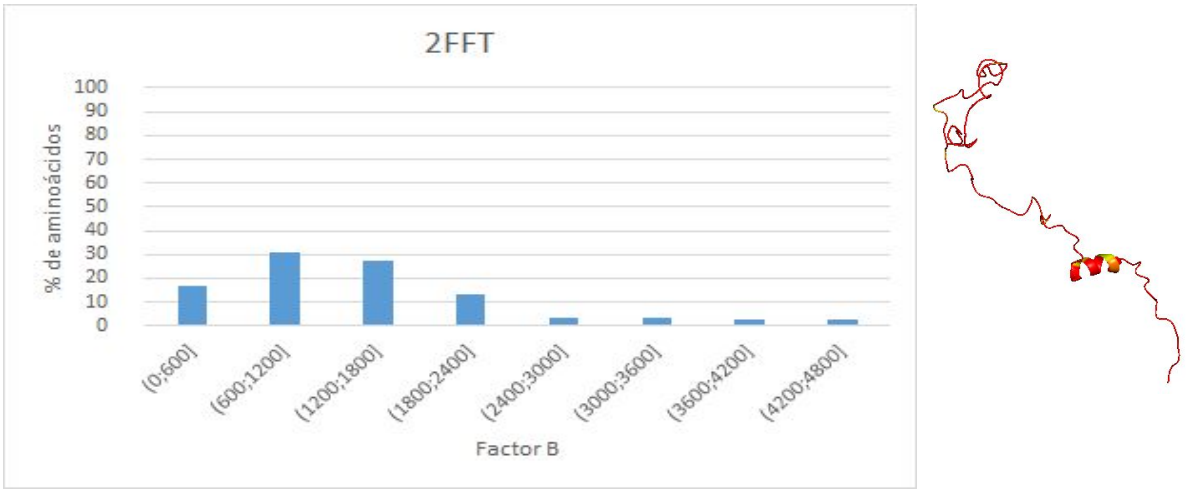

Figura 26. Proteína 2FFT (IDP). Izquierda, histograma B. Derecha: representación cinta o *cartoon* de las proteínas coloreadas por su factor B.

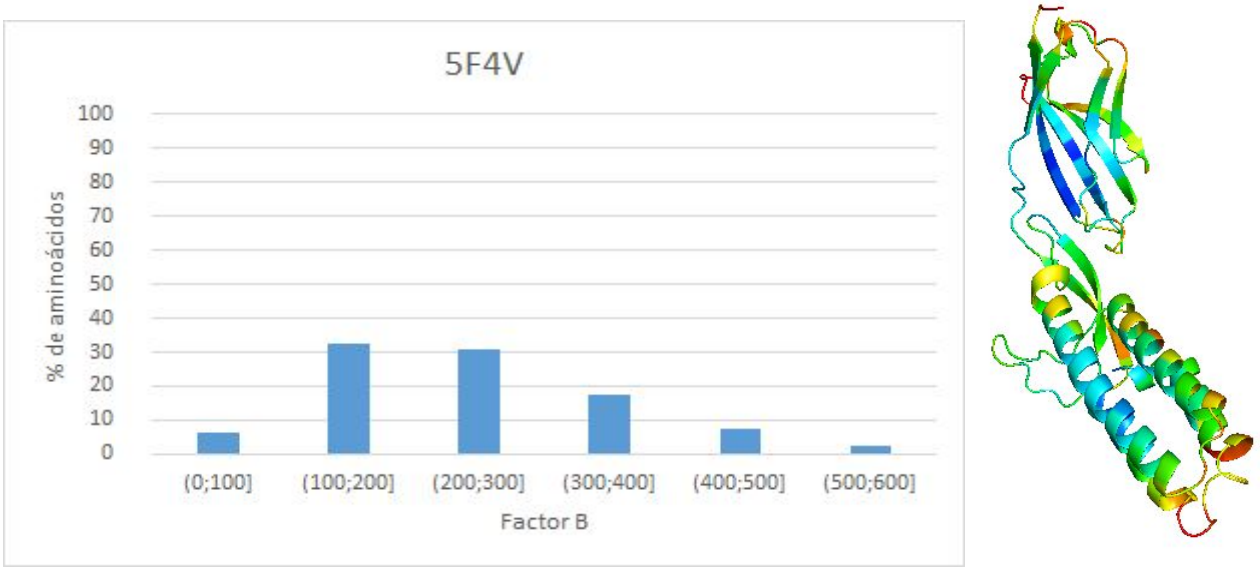

Figura 27. Proteína 5F4V (móvil). Izquierda, histograma B. Derecha: representación cinta o *cartoon* de las proteínas coloreadas por su factor B.

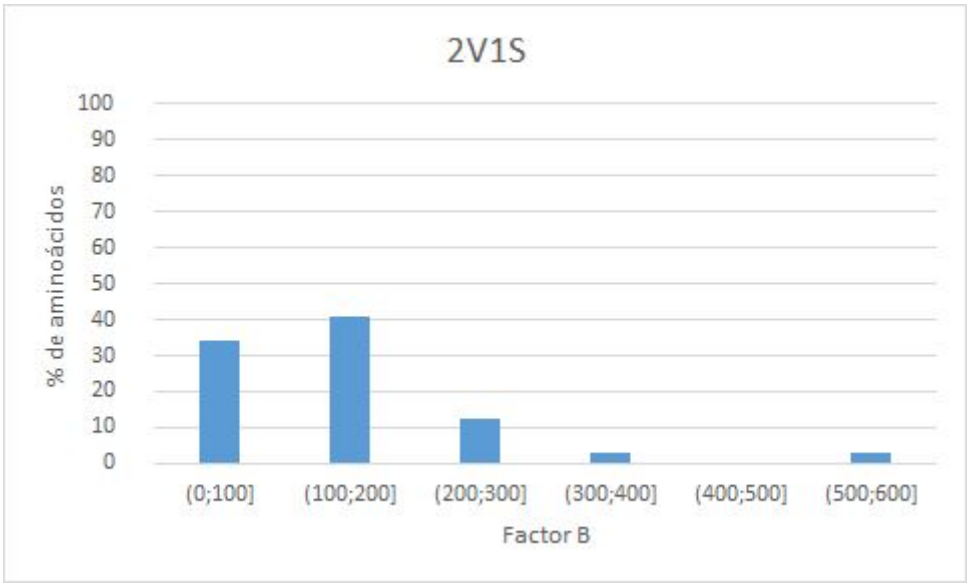

Figura 28. Proteína 2V1S (móvil). Izquierda, histograma B. Derecha: representación cinta o *cartoon* de las proteínas coloreadas por su factor B.

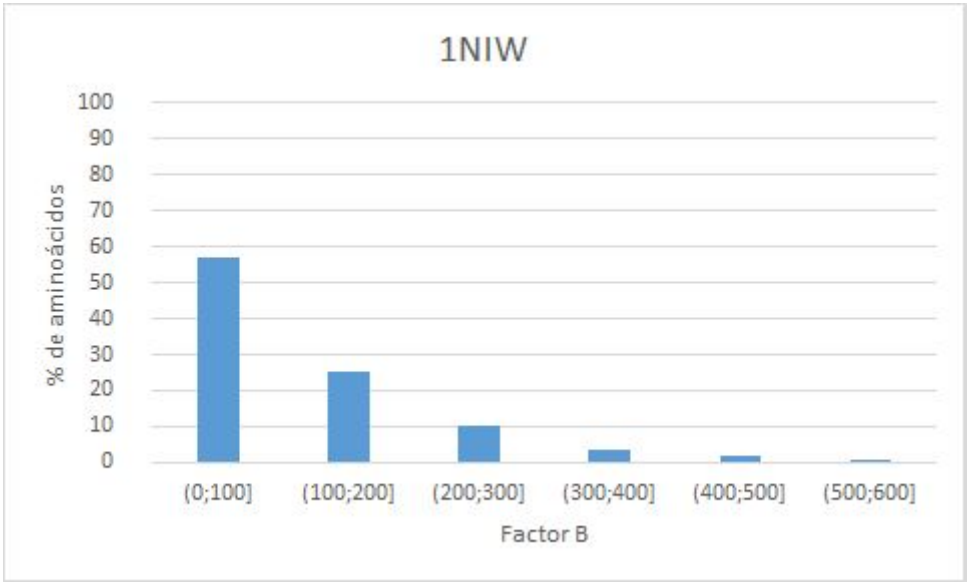

Figura 29. Figura A3.1. Proteína 1NIW (móvil). Izquierda, histograma B. Derecha: representación cinta o *cartoon* de las proteínas coloreadas por su factor B.

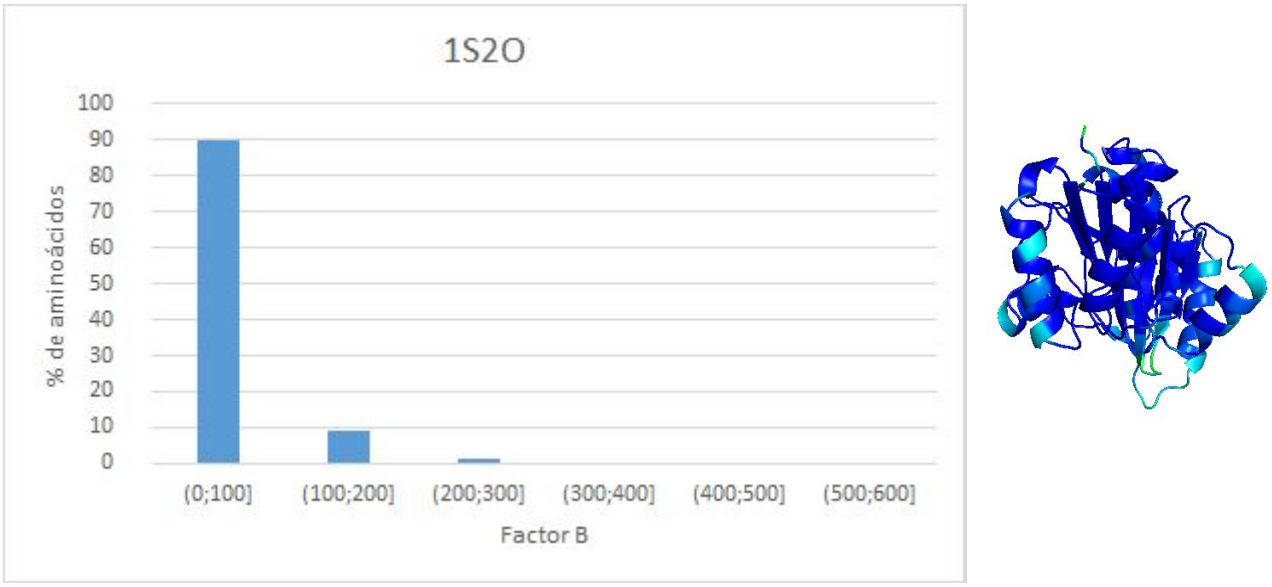

Figura 30. Proteína 1S2O (rígida, posiblemente maleables). Izquierda, histograma B. Derecha: representación cinta o *cartoon* de las proteínas coloreadas por su factor B.

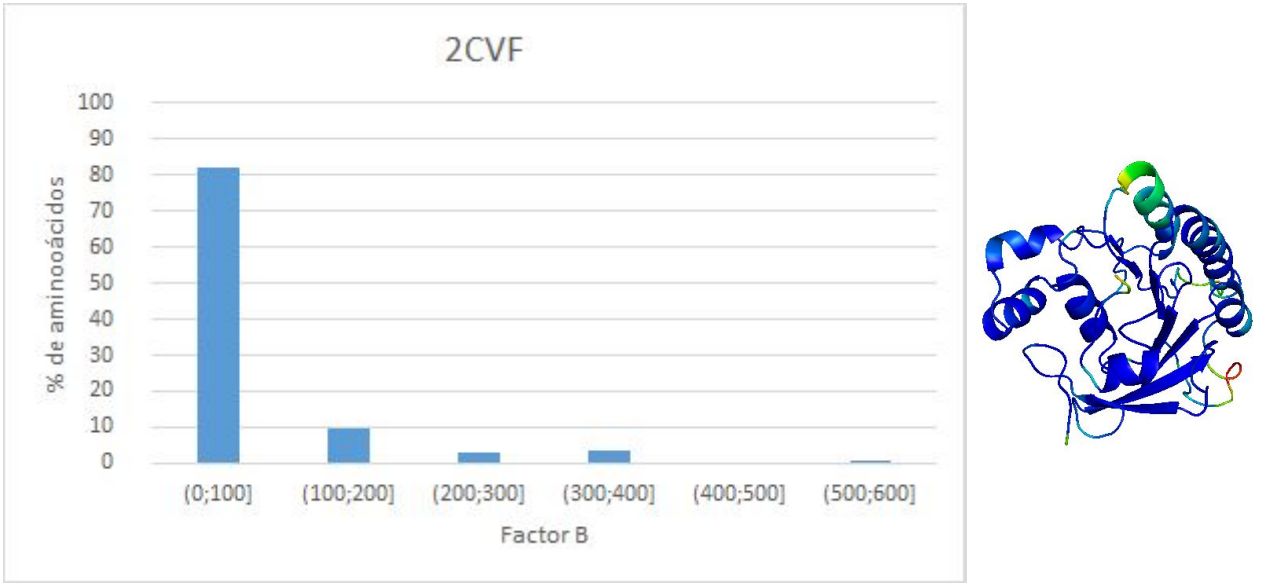

Figura 31. Proteína 2CVF (rígida, posiblemente maleables). Izquierda, histograma B. Derecha: representación cinta o *cartoon* de las proteínas coloreadas por su factor B.

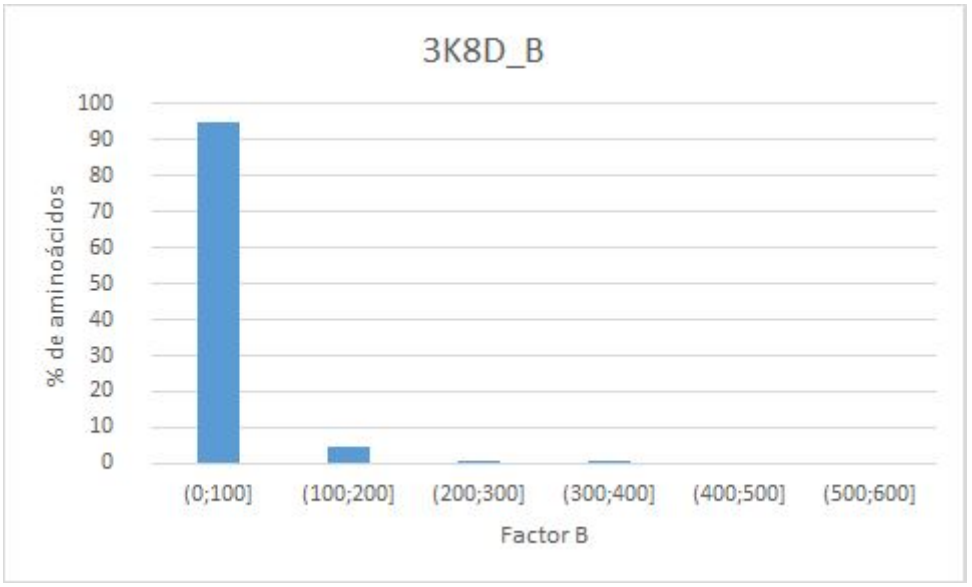

Figura 32. Proteína 3K8D, cadena B (rígida, posiblemente maleables). Izquierda, histograma B. Derecha: representación cinta o *cartoon* de las proteínas coloreadas por su factor B.

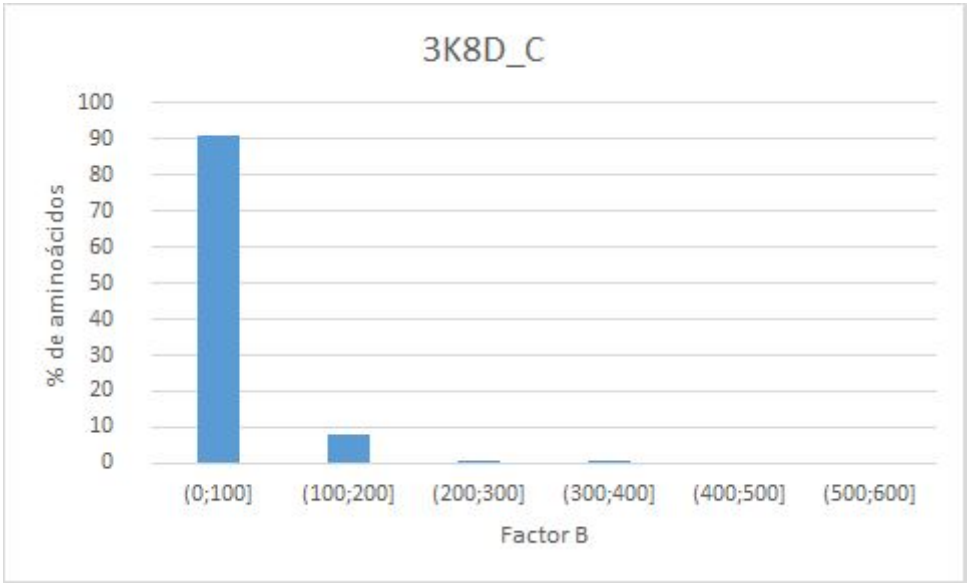

Figura 33. Proteína 3K8D, cadena C (rígida, posiblemente maleables). Izquierda, histograma B. Derecha: representación cinta o *cartoon* de las proteínas coloreadas por su factor B.

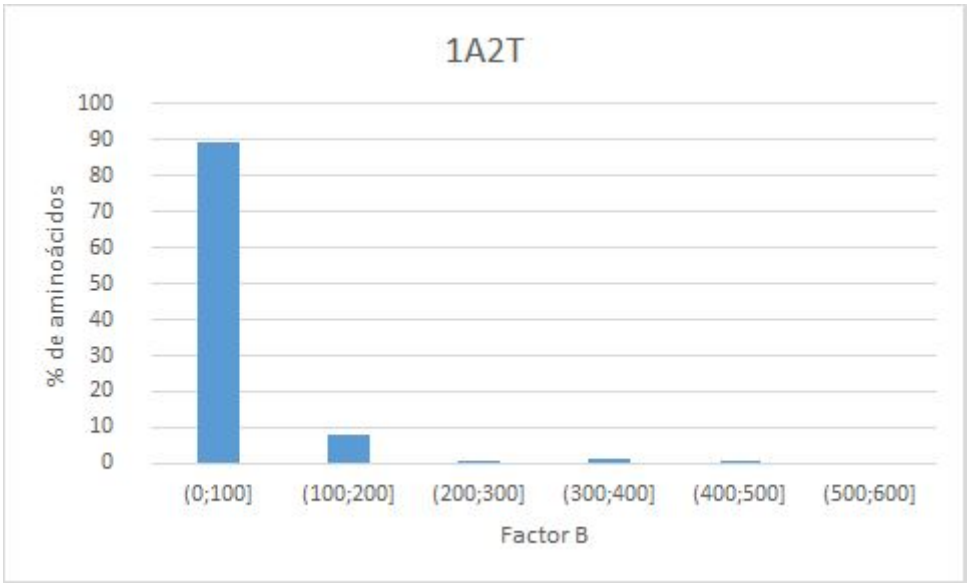

Figura 34. Proteína 1A2T (rígida). Izquierda, histograma B. Derecha: representación cinta o *cartoon* de las proteínas coloreadas por su factor B.

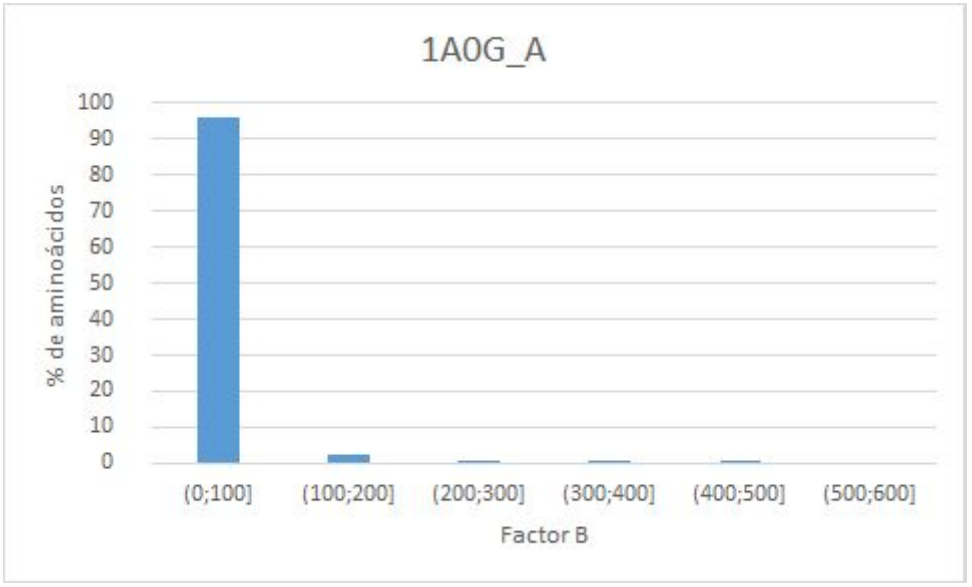

Figura 35. Proteína 1A0G, cadena A (rígida). Izquierda, histograma B. Derecha: representación cinta o *cartoon* de las proteínas coloreadas por su factor B.

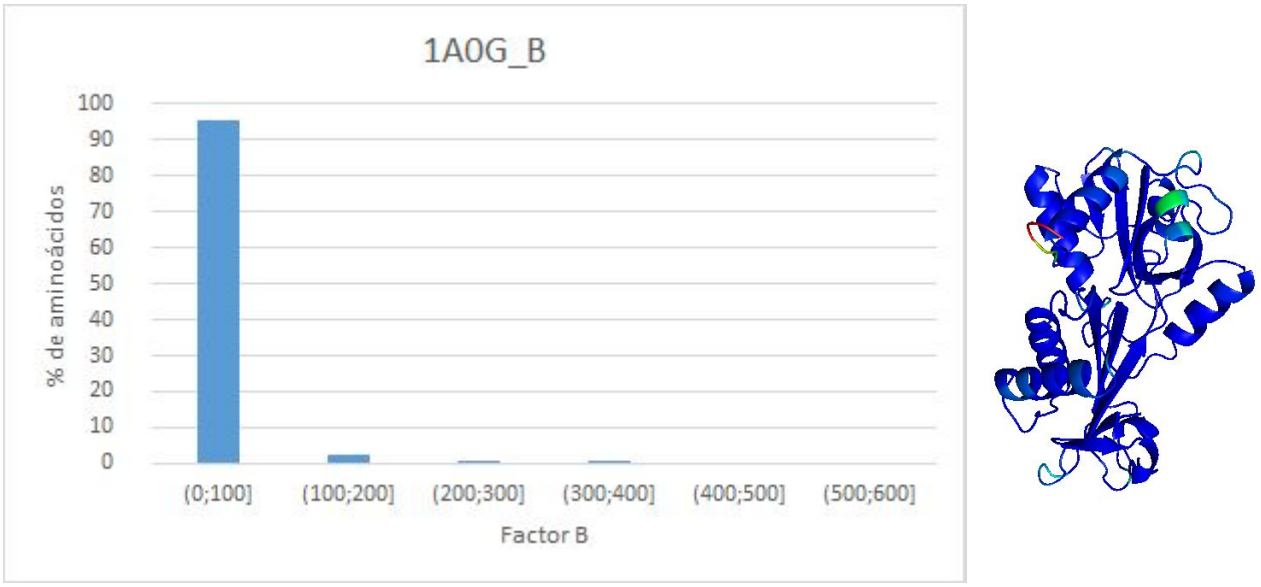

Figura 36. Proteína 1A0G, cadena A (rígida). Izquierda, histograma B. Derecha: representación cinta o *cartoon* de las proteínas coloreadas por su factor B.

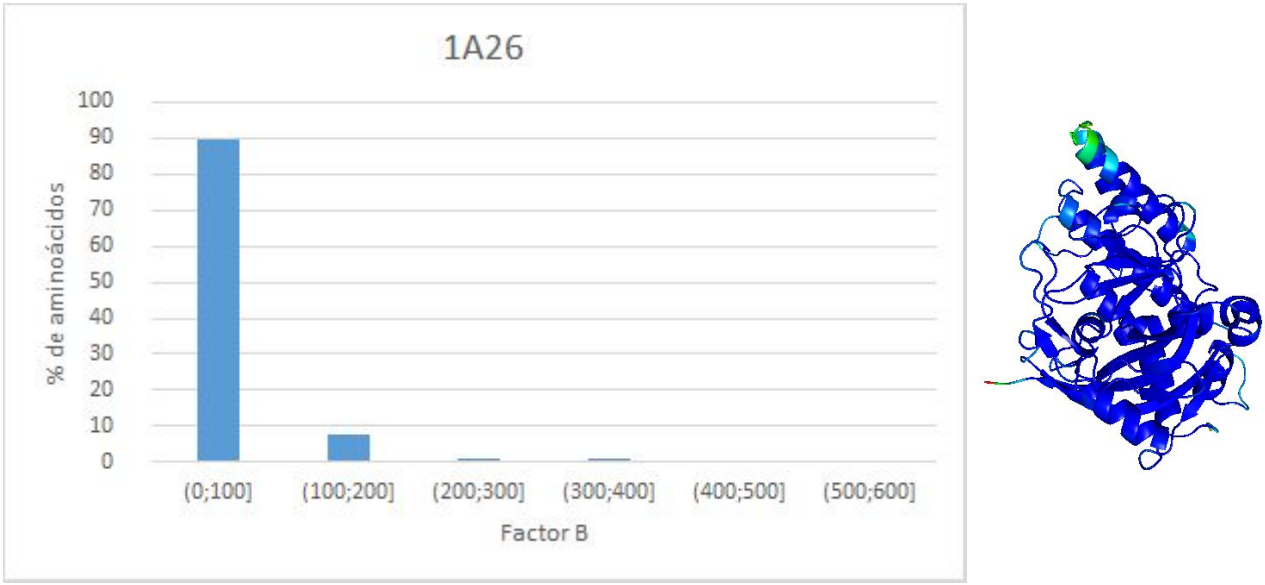

Figura 37. Proteína 1A26 (rígida). Izquierda, histograma B. Derecha: representación cinta o *cartoon* de las proteínas coloreadas por su factor B.

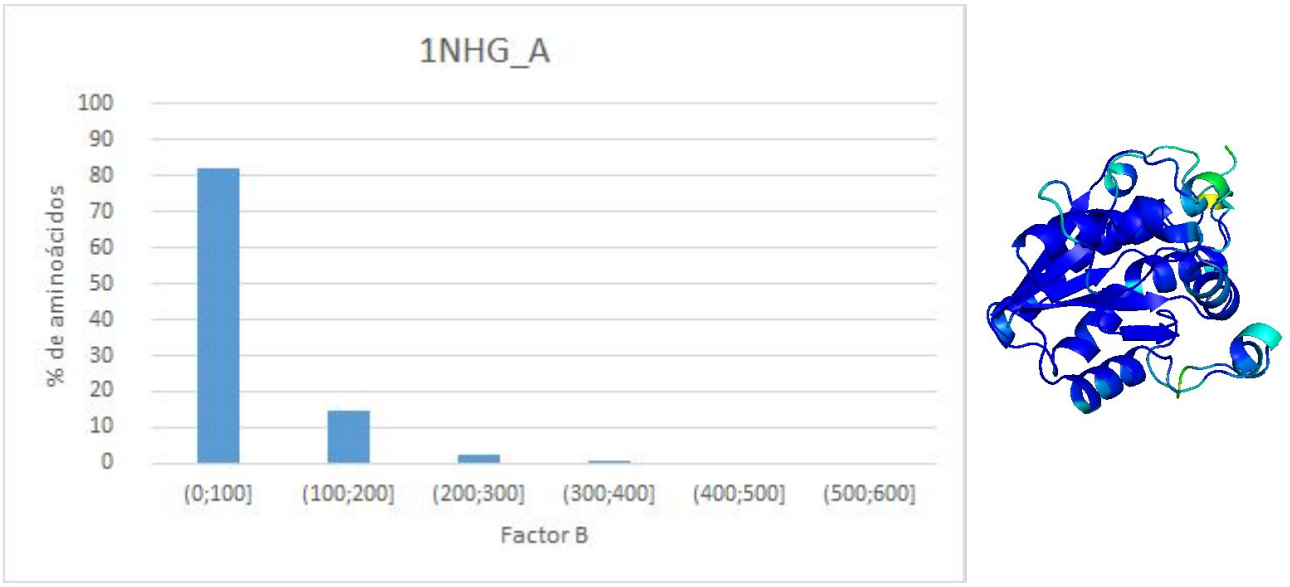

Figura 38. Proteína 1NHG, cadena A (rígida). Izquierda, histograma B. Derecha: representación cinta o *cartoon* de las proteínas coloreadas por su factor B.

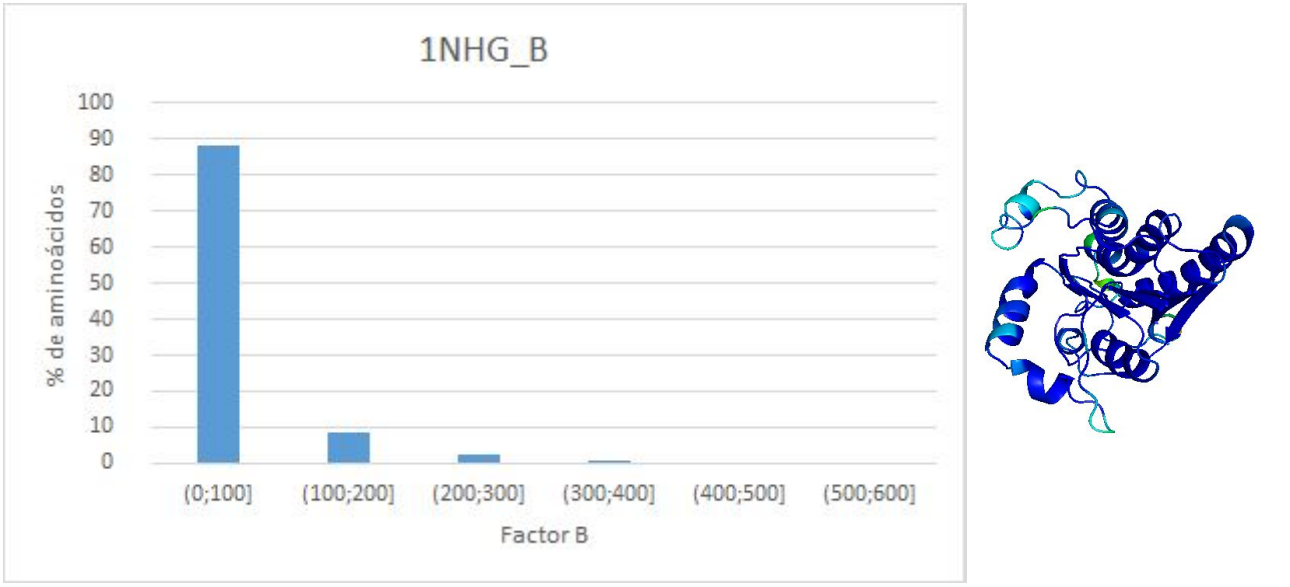

Figura 39. Proteína 1NHG, cadena B (rígida). Izquierda, histograma B. Derecha: representación cinta o *cartoon* de las proteínas coloreadas por su factor B.

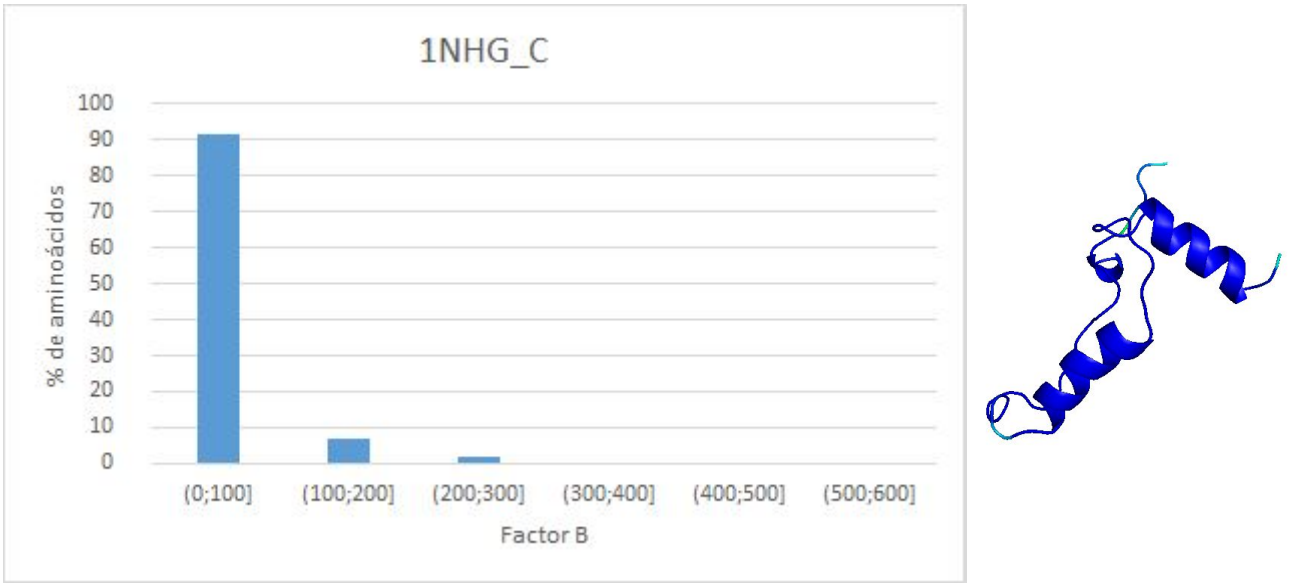

Figura 40. Proteína 1NHG, cadena C (rígida). Izquierda, histograma B. Derecha: representación “cartoon” de las proteínas coloreadas por su factor b.

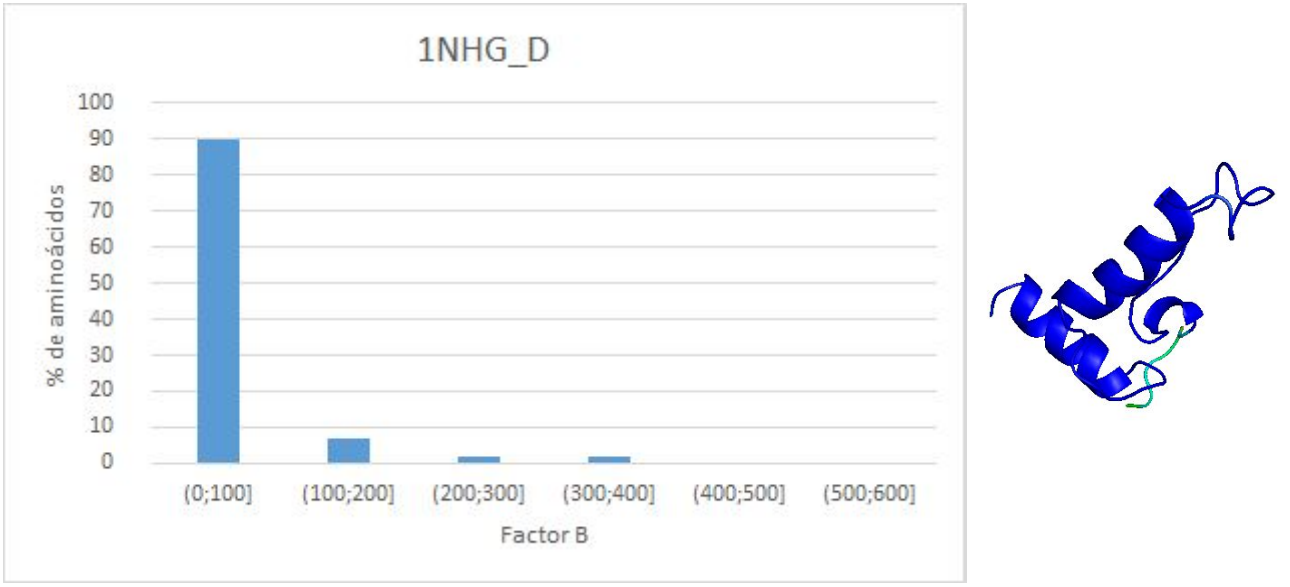

Figura 41. Proteína 1NHG, cadena D (rígida). Izquierda, histograma B. Derecha: representación cinta o *cartoon* de las proteínas coloreadas por su factor B.

Discusión

Las proteínas no son cuerpos rígidos, sino que poseen una cierta movilidad que es indispensable para su correcto funcionamiento. Sin embargo, la magnitud de estos movimientos no es universal: algunas proteínas se mueven más y otras mantienen su cadena casi inmóvil a lo largo del tiempo. Para responder la pregunta de dónde está almacenada la información a cerca de la movilidad de una proteína, hemos realizado simulaciones por DM a distintos grupos de proteínas clasificadas según su movilidad comprobada por RX. En el trabajo de Monzón y col. [27] clasifican las proteínas en tres clases según la movilidad de la estructura y grado de desorden. Las clases son las que llamaron “rígidas" (RMSD promedio entre los confórmeros = 0.83 Å), no tiene regiones desordenadas, muestra una baja diversidad conformacional. Los dos subconjuntos adicionales contienen regiones desordenadas, pero con una composición y comportamiento estructural diferencial. Las proteínas parcialmente desordenadas tienen en promedio el 67% de sus confórmeros con regiones desordenadas y RMSD promedio entre confórmeros de 1.1 Å, mayor número de bisagras y regiones desordenadas más largas. Las proteínas llamadas “maleables” tienen en promedio solo el 25% de los confórmeros desordenados y RMSD promedio de 1.3 Å. Este último grupo muestra además cavidades flexibles afectadas en tamaño por la presencia de regiones desordenadas y mayor diversidad de ligandos afines.

En el presente trabajo hemos hecho una simplificación y solo hemos considerados a las proteínas en dos grandes clases, como “rígidas” y “móviles” con el objeto de comprobar si un confórmero particular es capaz de reproducir todos los confórmeros conocidos experimentalmente.

Los distintos escenarios serían que, dado uno solo de los confórmeros para iniciar la dinámica, no se observen todos los demás posibles en la trayectoria (que no aparezcan algunas de las estructuras vistas). Otra posibilidad sería que dado un confórmero para iniciar la dinámica, se observen otros que no se han observado en las estructuras disponibles (que aparezcan estructuras no vistas). Por último cabe la posibilidad de que dado un confórmero inicial, se reproduzcan todos los conocidos.

La hipótesis es que si la información de la movilidad de la proteína está en su estructura, cada estructura de partida de la misma proteína podría tener información diferente, por lo tanto, una estructura no sería suficiente para conocer el espacio conformacional de esa proteína. Si la información acerca de la movilidad está codificada en la secuencia, cualquiera de los confórmeros alcanzaría para conocer el espacio conformacional completo de una proteína.

Para los cálculos realizados en este trabajo sólo se utilizó la estructura de uno de los confórmeros como dato inicial (información).

Para nuestro asombro, resultó que las proteínas se comportan en las dinámicas reproduciendo la tendencia que se observa en las “fotos” de los cristales.

Las proteínas consideradas “rígidas” han fluctuado alrededor de 0.20 nm respecto a la estructura de referencia (< 0.20 ± 0.01 nm) y 0.30 nm respecto a las estructuras de cada instante de la trayectoria (0.30 ± 0.05 nm), manteniendo un radio de giro con oscilaciones inferiores a 0.20 nm (< 0.20 ± 0.05 nm). El radio de giro casi constante, denota el hecho que la proteína se mantiene compacta a lo largo de la trayectoria. Este dato está en coincidencia con la cantidad de grupos encontrados al superponer las estructuras de los distintos instantes de tiempo que, si se establecen puntos de corte entre 0.20 y 0.30 nm, se puede afirmar que el número máximo no supera a 2. Observando sus valores de Factores B, un porcentaje superior al 80% de los aminoácidos presenta valores menores a 100.

Dentro de las proteínas “móviles”, se pueden apreciar dos subgrupos. El primero se caracteriza por tener fluctuaciones respecto a la estructura de referencia, valores máximos encontrados para las fluctuaciones respecto a las estructuras en cada instante de la simulación, amplitud de radio de giro y porcentaje de aminoácidos con valores de Factor B inferiores a 100 iguales que las proteínas rígidas (< 0.20 ± 0.01 nm, 0.30 ± 0.05 nm, < 0.20 ± 0.05 nm y > 80% de aminoácidos, respectivamente), diferenciándose levemente en la cantidad de grupos obtenidos: 4. Dado que para la elección de las proteínas en cada categoría solo hemos tenido en cuenta los RMSD mínimo, máximo y promedio, sin examinar y considerar la población de confórmeros disponibles, creemos que estas proteínas han sido consideradas erróneamente dentro del grupo de móviles. Creemos que son proteínas rígidas que por algún error experimental o condición extrema de la cristalización, una o unas pocas estructuras se diferencian de la población general y dan esos valores de RMSD. Consideramos que esas pocas estructuras diferentes de la población, son un artefacto y el análisis en detalle de estos casos nos aproxima a confirmar la hipótesis. Es remarcable que la técnica nos permite reconocer estos casos. Lo que sería sorprendente es que la técnica nos permita distinguir entre proteínas parcialmente desordenadas y maleables, los dos grupos que hemos compactado en nuestra categoría de móviles para simplificar. Esto sin duda será un tema para continuar con la investigación.

Para el segundo subgrupo, la movilidad de la estructura se hace evidente en todos los valores. Los valores de RMSD respecto a estructura de referencia son mayores a 0.20 nm (> 0.20 ± 0.01 nm), el

máximo valor RMSD entre estructuras en cada momento de la DM es superior a 0.40 nm (> 0.40 ± 0.01 nm): las matrices de RMSD entre estructuras de distinto punto de tiempo muestran que las diferencias son mayores que la observada para las proteínas rígidas (colores hacia el rojo). El rango de valores de Rg supera los 0.20 nm (> 0.20 ± 0.05 nm) y se logran diferenciar más de 4 grupos al superponer las estructuras de distintos momentos de la DM, considerando puntos de corte entre 0.20 y 0.30 nm. Para los valores de Factor B, la distribución de porcentajes no se concentra en valores inferiores a 100.

Teniendo en cuenta los comportamientos mencionados, observamos que se pueden correlacionar las trayectorias obtenidas por DM (entendida como las distintas estructuras que va adoptando la proteína a lo largo del tiempo) con la clasificación de proteínas descripta por Monzón y col. [27] basándose en las estructuras estáticas "fotos" disponibles en el PDB:

- 1A26, 1A2T, 1A0G (cadenas A y B), 1NHG (cadenas A, B, C y D) y 1S2O son proteínas rígidas;
- 2CVF y 3K8D (cadenas B y C) son proteínas rígidas, posiblemente maleables (esta clasificación no ha sido tomada en cuenta en este trabajo);
- 5F4V, 2V1S, 1NIW y 2FFT son proteínas móviles.

Conclusión

Dado los resultados expuestos, se puede establecer las siguientes características para clasificar a las proteínas según su flexibilidad:

- ➢ Proteínas Rígidas:
  - ○ Rango de valores de RMSD respecto a estructura de referencia: < 0.20 ± 0.01 nm;
  - ○ Máximo valor RMSD entre estructuras en cada momento de la trayectoria: 0.30 ± 0.05 nm;
  - ○ Amplitud del rango de valores de Rg: < 0.20 ± 0.05 nm;
  - ○ Cantidad máxima de grupos encontrados para punto de corte entre 0.20 y 0.30 nm: 2.
  - ○ Factor B: más del 80% de aminoácidos con valores inferiores a 100.
- ➢ Proteínas Móviles:
  - ○ Rango de valores de RMSD respecto a la estructura de referencia: > 0.20 ± 0.01 nm;
  - ○ Máximo valor RMSD entre estructuras en cada momento de la trayectoria: > 0.40 ± 0.01 nm;
  - ○ Amplitud del rango de valores de Rg: > 0.20 ± 0.05 nm;
  - ○ Cantidad máxima de grupos encontrados para punto de corte entre 0.20 y 0.30 nm: > 4.
  - ○ Factor B: distribución de los aminoácidos entre los valores 1 a 600.
- ➢ Proteínas Rígidas, posiblemente maleables: Estas proteínas tienen un comportamiento que no es atribuible ni a las totalmente móviles ni a las totalmente rígidas. Si bien es apresurado establecer una clasificación y requiere el análisis de un mayor número de proteínas, posiblemente estemos ante el comportamiento de las proteínas maleables.
  - ○ Rango de valores de RMSD respecto a estructura de referencia: < 0.20 ± 0.01 nm;

- Máximo valor RMSD entre estructuras en cada momento de la DM: 0.30 ± 0.05 nm;
- Amplitud del rango de valores de Rg: < 0.20 ± 0.05 nm;
- Cantidad máxima de grupos encontrados para punto de corte entre 0.20 y 0.30 nm: < 4.
- Factor B: más del 80% de aminoácidos con valores inferiores a 100.

De esta forma, se puede establecer que las simulaciones de dinámica molecular permiten, conociendo uno de los estados nativos de proteínas, predecir la existencia de otros estados utilizando la información de los valores de RMSD (respecto a una estructura de referencia y entre todas las estructuras obtenidas de la simulación), radio de giro, cantidad de grupos al superponer las estructuras de los distintos puntos de tiempo y la distribución de Factor B entre los aminoácidos de la cadena proteica. Estas propiedades se pueden obtener con el programa de simulación como el utilizado en el presente trabajo (GROMACS). Con esto, queda dilucidada la relación entre conformación y dinámica: una proteína móvil da confórmeros a lo largo del tiempo suficientemente diferentes para no coincidir a menos de 0.4 Å al superponerse, mientras que, por otro lado, una proteína rígida, permanece como tal a lo largo de toda la trayectoria.

Si bien no podemos asegurar que la información dinámica de una proteína está codificada en su secuencia, estos resultados sugieren que esa es una posibilidad.

La conclusión general es que sería posible, teniendo un solo confórmero depositado en las bases de datos, dar cuenta de la flexibilidad de una proteína. Los resultados del presente trabajo sugieren que podríamos ser capaces de predecir la dinámica de una proteína conociendo una estructura, con las implicancias funcionales de ese conocimiento. Esto supliría la incompletitud dinámica del PDB (Marino-Buslje y col. [47]) y sugiere enfocar los esfuerzos en conseguir una estructura por proteína.

Las proteínas, además de poder adoptar conformaciones mínimamente diferentes, pertenecen a una familia de proteínas homólogas donde los distintos miembros ejercen la misma función (o similar) en distintas especies. Las familias de proteínas pueden ser muy extensas y los homólogos tener un rango amplio de similitud de secuencia. Habiendo homólogos altamente idénticos (ej: 80% o más) o marginalmente idénticos (ej 30% o menos). Un paso posterior de análisis sería correr dinámicas moleculares iniciándose con estructuras de distintas proteínas de la familia, ej: con homólogos 90, 70, 50 y 30%, con el objeto de conocer si la información dinámica es inherente a

una familia de proteínas o es particular de cada proteína. Si los homólogos también tienen el mismo patrón dinámico, ¿hasta dónde se mantiene evolutivamente?, ¿tendrán el mismo patrón los homólogos 80% y los 30% idénticos? La información interconectada de la evolución, secuencia, estructura, función y dinámica de las proteínas aún plantea muchos interrogantes por resolver.

Bibliografía

[1] D. L. Nelson, M. M. Cox, “Lehninger, Principles of Biochemistry”, W.H. Freeman and Company, New York, 2005, Cuarta Edición, Capítulo 1.

[2] “What is a protein?”, PDB-101: Educational Portal of PDB, https://pdb101.rcsb.org.

[3] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, P. Walter, “Molecular Biology of the cell”, Garland Science, New York, 2008, Quinta Edición, Capítulo 3.

[4] B. Adhikari, J Cheng, “Protein Residue Contacts and Prediction Methods”, Methods Mol Biol, 1415:463-476, 2016.

[5] E. F. Pettersen, T. D. Goddard, C. C. Huan, G. S. Couch, D. M. Greenblatt, E. C. Meng, T. E. Ferrin, “UCSF Chimera: A visualization system for exploratory research and analysis”, J Comput Chem, 13:1605-16012, 2004.

[6] D. J. Zea, A. M. Monzón, G. Parisi, C. Marino-Buslje, “How is structural divergence related to evolutionary information?”, Mol Phylogenet Evol, 127:859-866, 2018.

[7] A. M. Monzon, M. S. Fornasari, D. J. Zea, G. Parisi, “Exploring Protein Conformational Diversity”, Methods Mol Biol, 1851:353-365, 2018.

[8] P. Salahuddin, “Protein Folding in the Cell”, J. Biochem Mol Biol Res 1:123-130, 2015.

[9] J. C. Lin, H. L. Liu, “Protein conformational diseases: from mechanisms to drug designs”, Curr Drug Discov Technol., 3:145-53, 2006.

[10] A. M. Ellisdon, S. P. Bottomley, “The role of protein misfolding in the pathogenesis of human diseases”, IUBMB Life,56:119-123, 2004.

[11] P. Sweeney, H. Park, M. Baumann, J. Dunlop, J. Frydman, R. Kopito, A. McCampbell, G. Leblanc, A. Venkateswaran, A. Nurmi, R. Hodgson, “Protein misfolding in neurodegenerative diseases: implications and strategies”, Transl Neurodegener, 6:6, 2017.

[12] E. Reynaud, “Protein Misfolding and Degenerative Diseases”, Nature Education, 3:28, 2010.

[13] T. M. Picknett, S. Brenner, “X-Ray Crystallography”, Encyclopedia of Genetics, 2154, 2001.

[14] S. Hardinger, “Guide to Understanding X-ray Crystallography”, Department of Chemistry & Biochemistry, University of California, Tutorial 73, 2018.

[15] Q. Teng, “Structural Biology: Practical NMR Applications”, Springer Science+Business Media, 65-101, 2013.

[16] J. Pietzsch, “Protein folding technology”, Nature Publishing Group, 2002

[17] P. Broadwith, “Explainer: What is cryo-electron microscopy”, Chemistry World, 2017.

[18] P. Brzezinski, “Scientific Background on the Nobel Prize in Chemistry 2017: The development of cryo-electron microscopy”, The Royal Swedish Academy of Sciences, 2017.

[19] D. Cressey, E. Callaway, “Cryo-electron microscopy wins chemistry Nobel”, Nature: International weekly journal of science, 2017.

[20] “PDB File Format - Contents Guide Version 3.30”, World Wide Protein Data Bank, 2012.

[21] H.M. Berman, K. Henrick, H. Nakamura, “Announcing the worldwide Protein Data Bank”, Nat Struct Biol 10:980, 2003.

[22] F. C. Bernstein, T. F. Koetzle, G. J. Williams, E. F. Jr Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, M. Tasumi, “The Protein Data Bank: a computer-based archival file for macromolecular structures”, J Mol Biol, 112(3):535-42, 1977.

[23] H. M. Berman, T. Battistuz, T. N. Bhat, W. F. Bluhm, P. E. Bourne, K. Burkhardt, Z. Feng, G. L. Gilliland, L. Iype, S. Jain, P. Fagan, J. Marvin, D. Padilla, V. Ravichandran, B. Schneider, N. Thanki, H. Weissig, J. D. Westbrooka, C. Zardecki, “The Protein Data Bank”, Biological Crystallography, D58: 899-907, 2002.

[24] S. Dutta, R. Kramer Green, C. L. Lawson, “Introduction to Biological Assemblies and the PDB Archive”, Educational portal PDB-101, https://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/biological-assemblies.

[25] J. Marchetti, A. M. Monzón, S. C. E. Tosatto, G. Parisi, M. S. Fornasari, “Ensembles from Ordered and Disordered Proteins Reveal Similar Structural Constraints during Evolution”, J Mol Biol, 431:1298-1307, 2019.

[26] C. Marino-Buslje, A. M. Monzón, D. J. Zea, M. S. Fornasari, G. Parisi, “On the dynamical incompleteness of the Protein Data Bank”, Brief Bioinform, 20:356-359. 2019.

[27] A. M. Monzón, E. Juritz, M. S. Fornasari, G. Parisi, “CoDNaS: a database of conformational diversity in the native state of proteins”, Bioinformatics, 29:2512-2514, 2013.

[28] A. M. Monzón, C. O. Rohr, M. S. Fornasari, G. Parisi, “CoDNaS 2.0: a comprehensive database of protein conformational diversity in the native state”, Database (Oxford), pii: baw038, 1-8, 2016.

[29] A. M. Monzón, D. J. Zea, M. S. Fornasari, T. E. Saldaño, S. Fernandez-Alberti, S. C. E. Tosatto, G. Parisi, “Conformational diversity analysis reveals three functional mechanisms in proteins”, PLOS Comput Biol., 13:e1005398, 2017.

[30] M Karplus, J. A. McCammon, “Molecular dynamics simulations of biomolecules”, Nat Struct Biol, 9:646652, 2002.

[31] A. Basile, K. Ghasemzadeh, “Current Trends and Future Developments on (Bio-) Membranes: Silica Membranes - Preparation, Modelling, Application, and Commercialization”, Elsevier, Ed. 1, 2017.

[32] M. Abraham, B. Hess, D. van der Spoel, E. Lindahl, “GROMACS - Groningen Machine for Chemical Simulations: User Manual Version 5.0.4”, 2014.

[33] Y. W. Dong, M. L. Liao, X. L. Meng, G.N. Somero, “Structural flexibility and protein adaptation to temperature: Molecular dynamics analysis of malate dehydrogenases of marine molluscs”, Proc Natl Acad Sci USA. 115:1274-1279, 2018.

[34] G. R. C. Pereira, A. N. R. Da Silva, S. S. Do Nascimento, J. F. De Mesquita, “In silico analysis and molecular dynamics simulation of human superoxide dismutase 3 (SOD3) genetic variants”, J Cell Biochem, 120:3583-3598, 2018.

[35] D. Van Der Spoel, E.Lindahl, B. Hess, G. Groenhof, A. E. Mark, H. J. C. Berendsen, “GROMACS: Fast, flexible, and free”, Journal of Computational Chemistry, 26:1701-1718, 2005.

[36] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, T. E. Ferrin, “UCSF Chimera--a visualization system for exploratory research and analysis”, J Comput Chem, 25:1605-1612, 2004.

[37] S. Fieulaine, J. E. Lunn, F. Borel, J. L. Ferrer, “The Structure of a Cyanobacterial Sucrose-Phosphatase Reveals the Sugar Tongs That Release Free Sucrose in the Cell”, Plant Cell, 17:2049-2058, 2005.

[38] G. A. Kaminski, R. A. Friesner, J. Tirado-Rives, W. L. Jorgensen. Evaluation and Reparametrization of the OPLS-AA Force Field for Proteins via Comparison with Accurate Quantum Chemical Calculations on Peptides. J Phys Chem B, 105:6474-6487. 2001.

[39] V. N. Maiorov, G. M. Crippen, “Size-independent comparison of protein three-dimensional structures”, Proteins: Structure, Function, Bioinformatics, 22:273-283, 1995.

[40] X. Daura, K. Gademann, B. Jaun, D. Seebach, W. F. van Gunsteren, A. E. Mark, “Peptide Folding: When Simulation Meets Experiment”, Angewandte, 38:236-240, 1999.

[41] W. Humphrey, A. Dalke, K. Schulten, "VMD - Visual Molecular Dynamics", J. Molec. Graphics, Vol. 14, pp. 33-38, 1996.

[42] J. Stone. “script splitmultiframepdb.tcl” para VMD, https://www.ks.uiuc.edu/Research/vmd/script_library/scripts/splitmultiframepdb/.

[43] Open-Source PyMOL is Copyright (C) Schrodinger LLC. PyMOL is a trademark of Schrodinger, LLC. https://github.com/schrodinger/pymol-open-source/blob/master/LICENSE.

[44] I. Kufareva, R. Abagyan, “Methods of protein structure comparison”, Methods Mol Biol, 857:231–257, 2009.

[45] S. Dastmalchi, M. Hamzeh-Mivehroud, B. Sokouti, “Methods and Algorithms for Molecular Docking-Based Drug Design and Discovery (Advances in Chemical and Materials Engineering)”, IGI Global, Ed 1, 2016.

[46] L. Zaslavsky, S. Ciufo, B. Fedorov, T. Tatusova, “Clustering analysis of proteins from microbial genomes at multiple levels of resolution”, BMC Bioinformatics, 17:276, 2016.

[47] C. Marino-Buslje, A. M. Monzón, D. J. Zea, M. S. Fornasari, G. Parisi, “On the dynamical incompleteness of the Protein Data Bank”, Brief Bioinform, Vol. 20, pp 356–359, 2019.

[48] H. J. C. Berendsen, J. R. Grigera and T. P. Straatsma, The missing term in effective pair potentials, Journal of Physical Chemistry 91:6269-6271, 1987.

[49] B. Hess, H. Bekker H, H. J. C. Berendsen, J. G. E. M. Fraaije. LINCS:a linear constraint solver for molecular simulations. J Comput Chem 18:1463-1472, 1997.

[50] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren WF, J. Hermans J. Interaction models for water in relation to protein hydration. In: Pullman B (ed) Intermolecular forces. Reidel, Dordrecht, pp 341-342, 1984.

Anexos

Anexo 1

A continuación, se presentan los dendrogramas de las proteínas estudiadas obtenidos de http://ufq.unq.edu.ar/codnas/index.php. Luego de la superposición en el espacio de los distintos confórmeros de la misma proteína, todos los que se diferencian menos de 0.4 Å quedan en el mismo grupo (encuadrados en rojo).

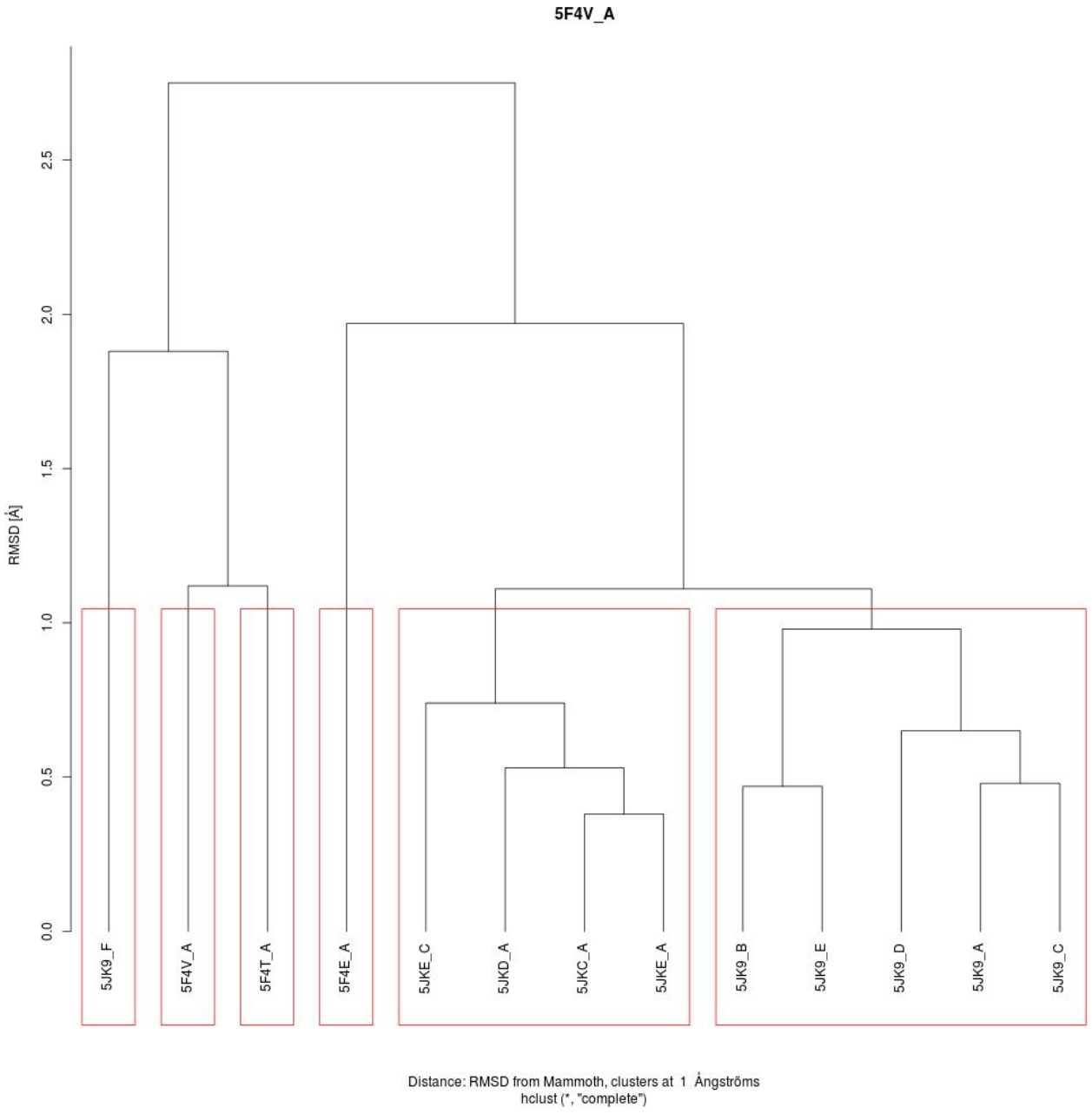

Figura A1.1. Dendrograma de proteína 5F4V.

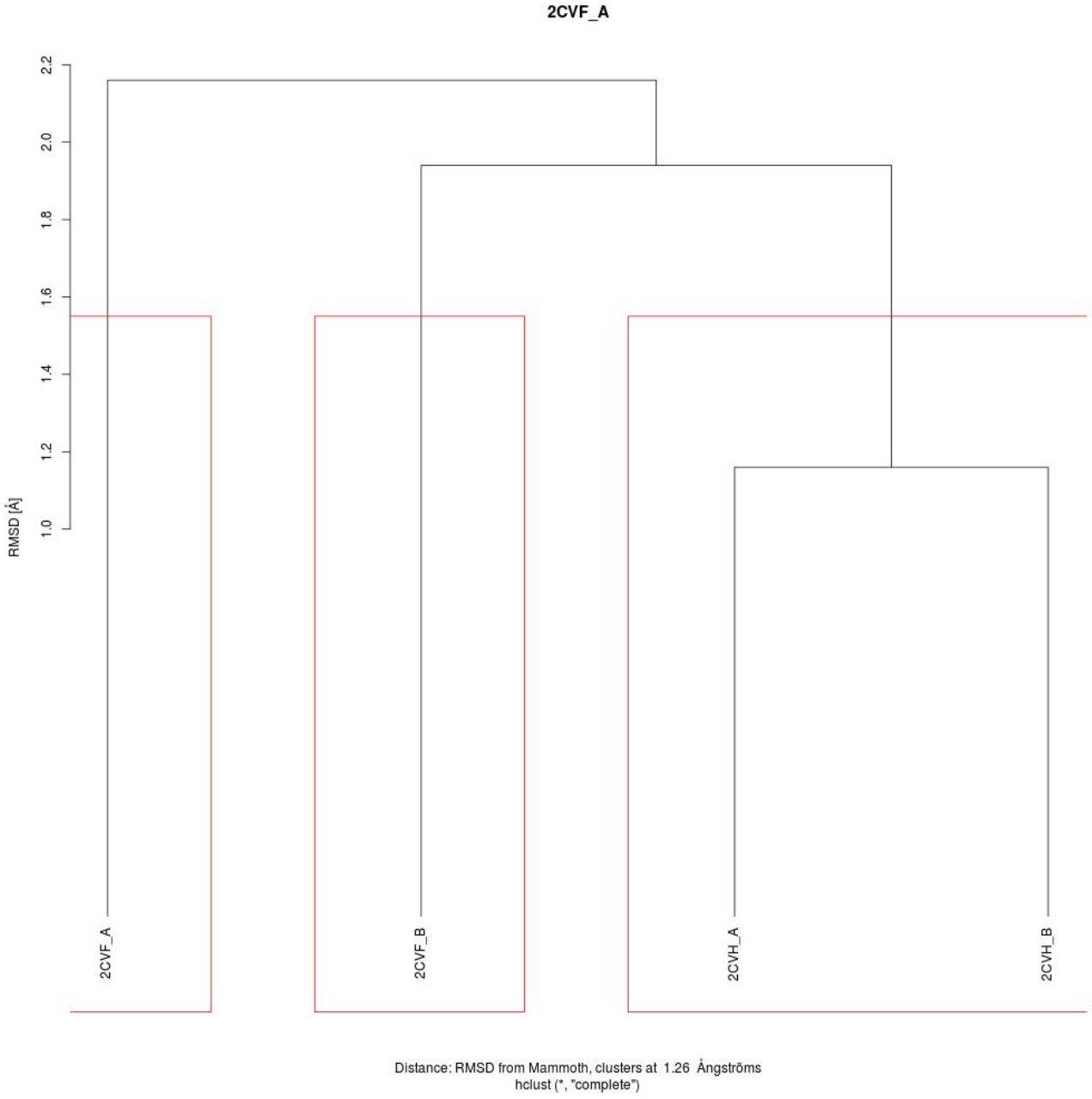

Figura A1.2. Dendrograma de proteína 2CVF.

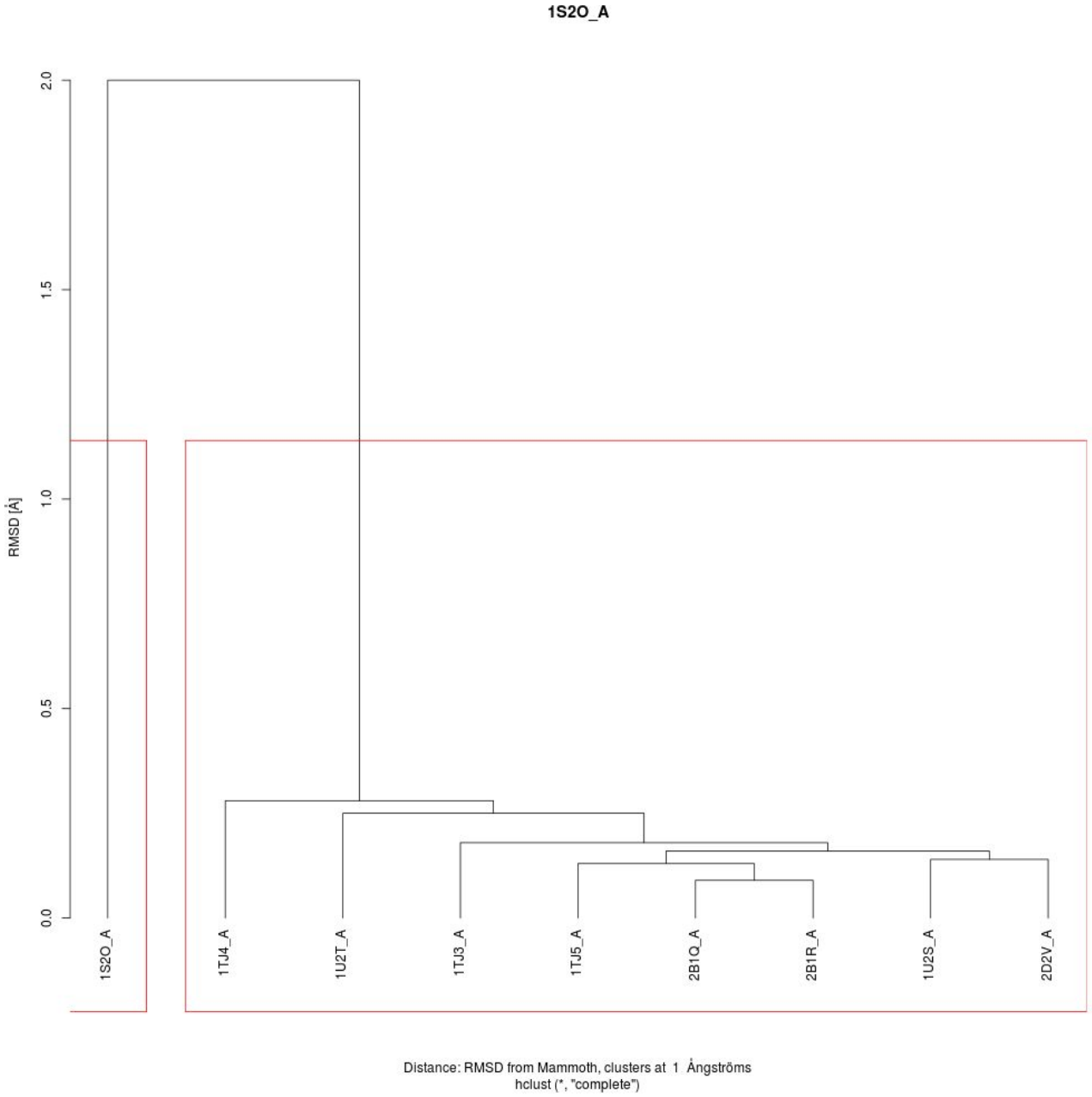

Figura A1.3. Dendrograma de proteína 1S2O.

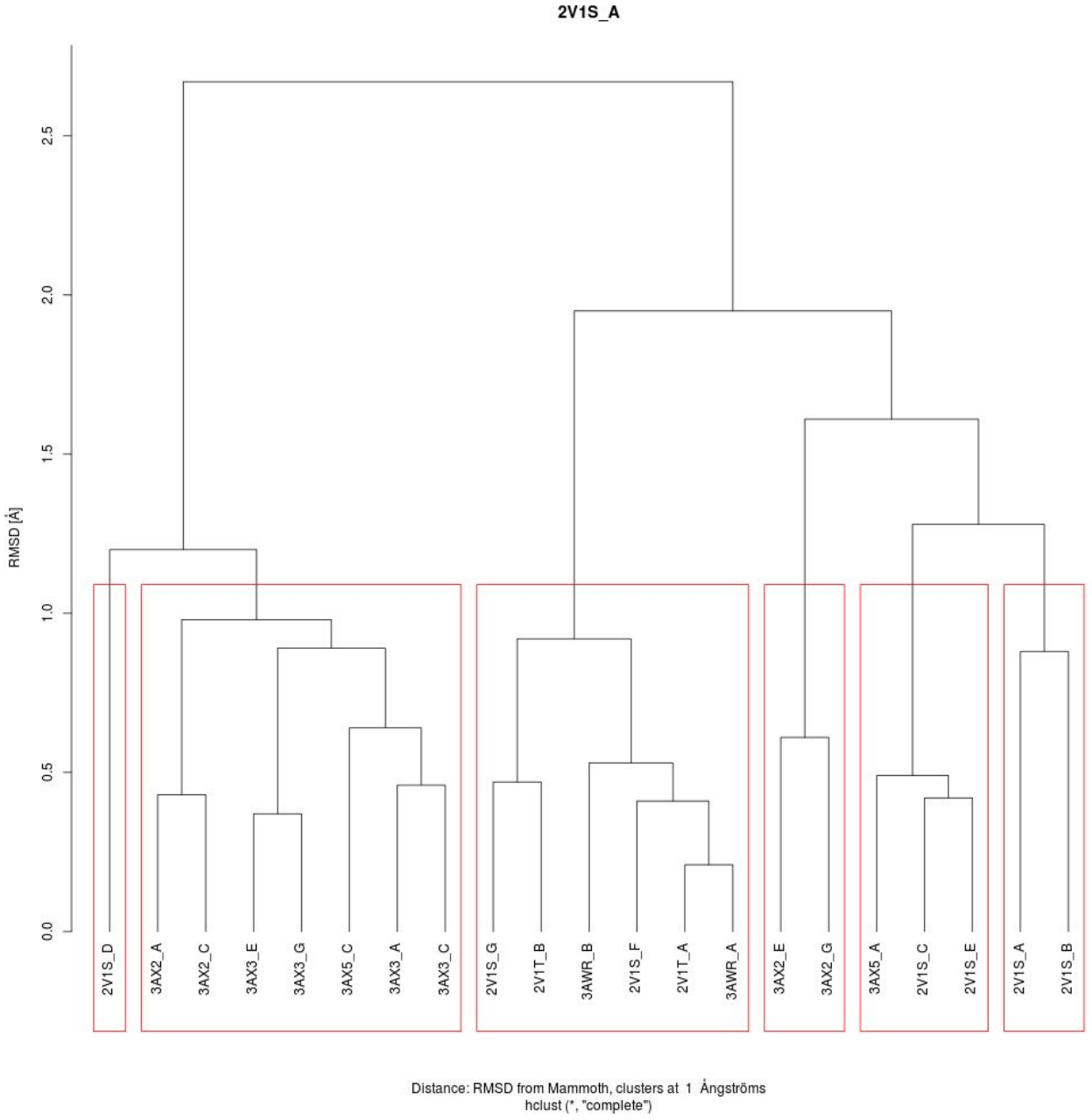

Figura A1.4. Dendrograma de proteína 2V1S.

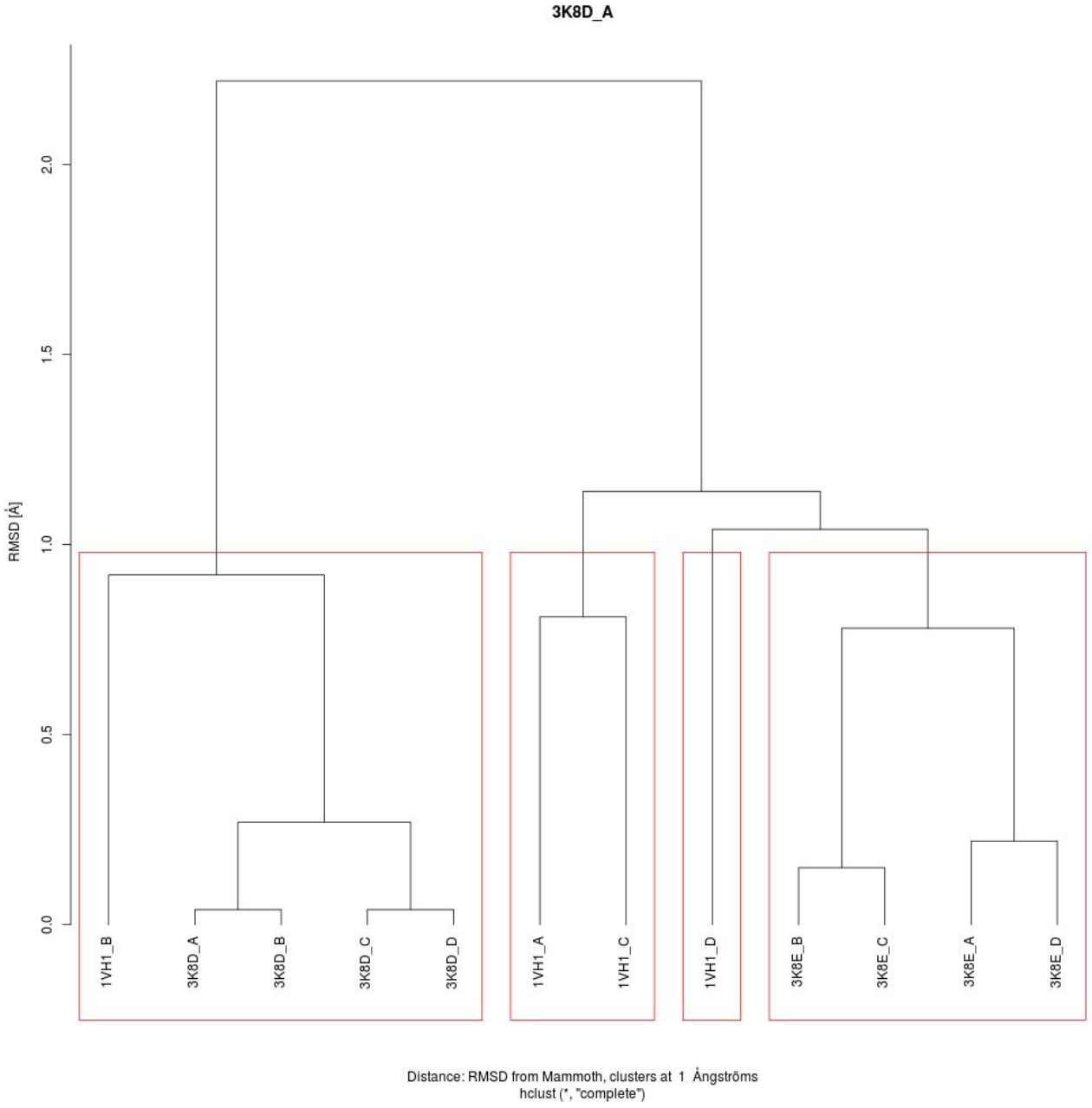

Figura A1.5. Dendrograma de proteína 3K8D.

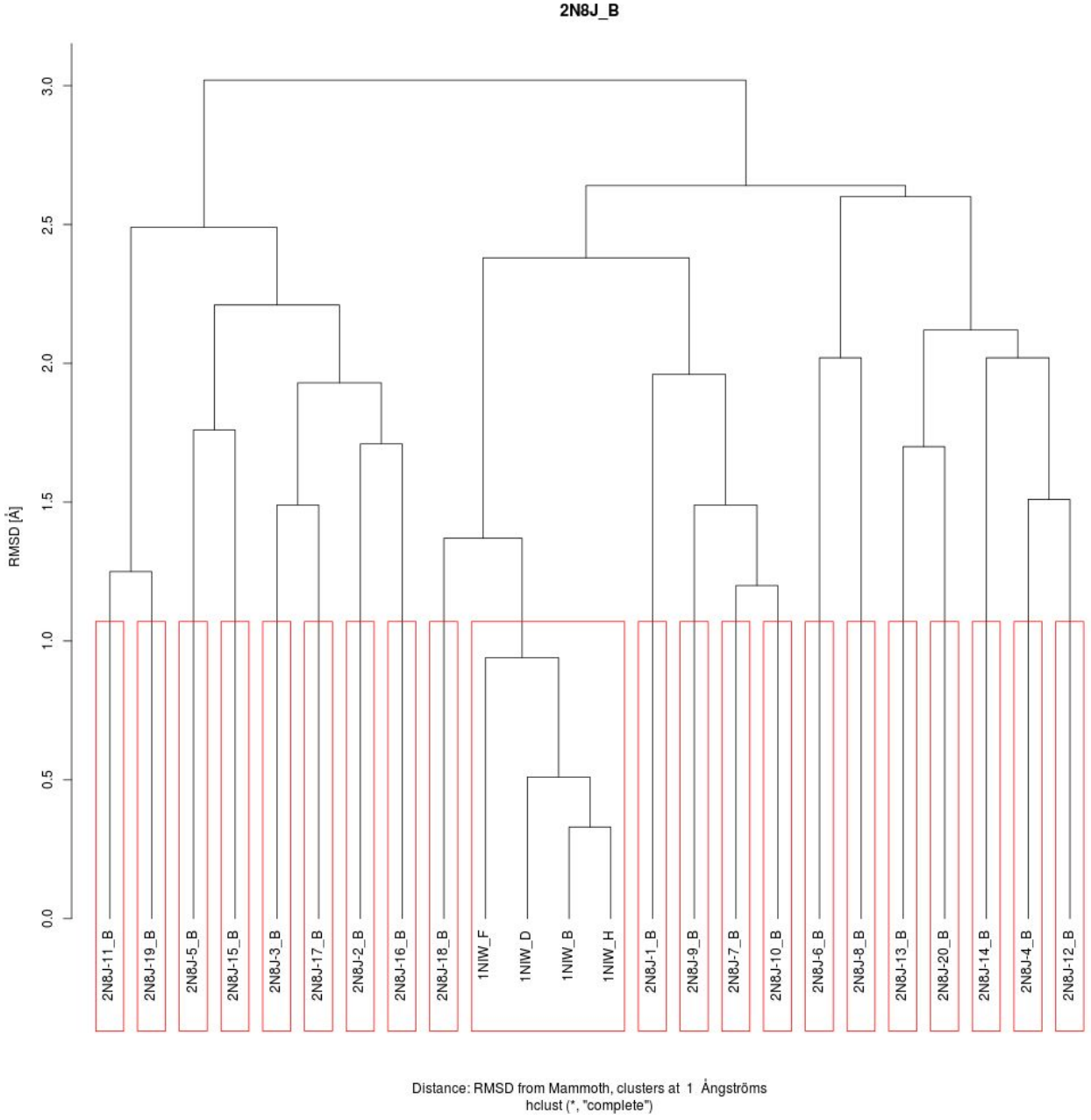

Figura A1.6. Dendrograma de proteína 1NIW.

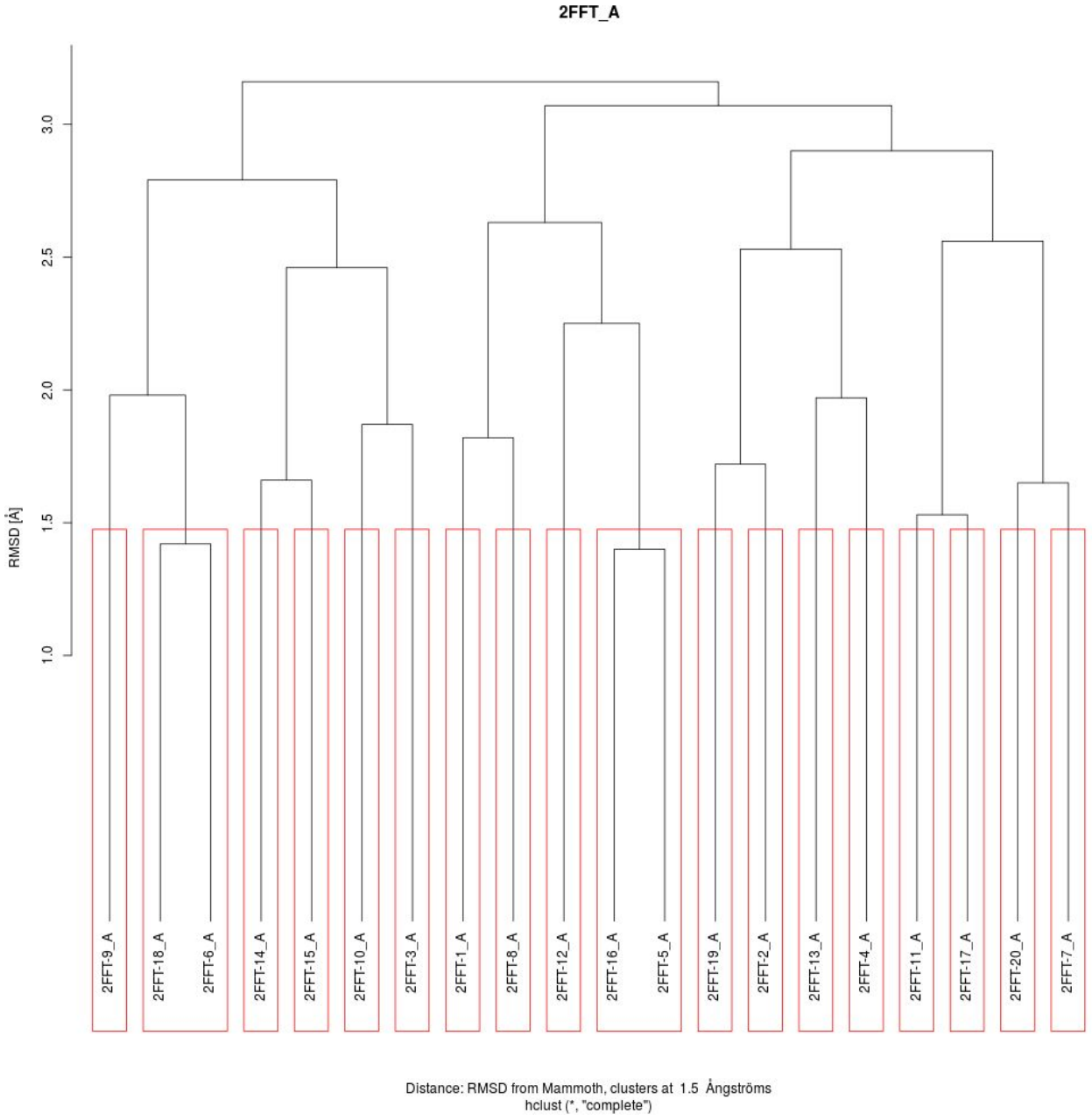

Figura A1.7. Dendrograma de proteína 2FFT.

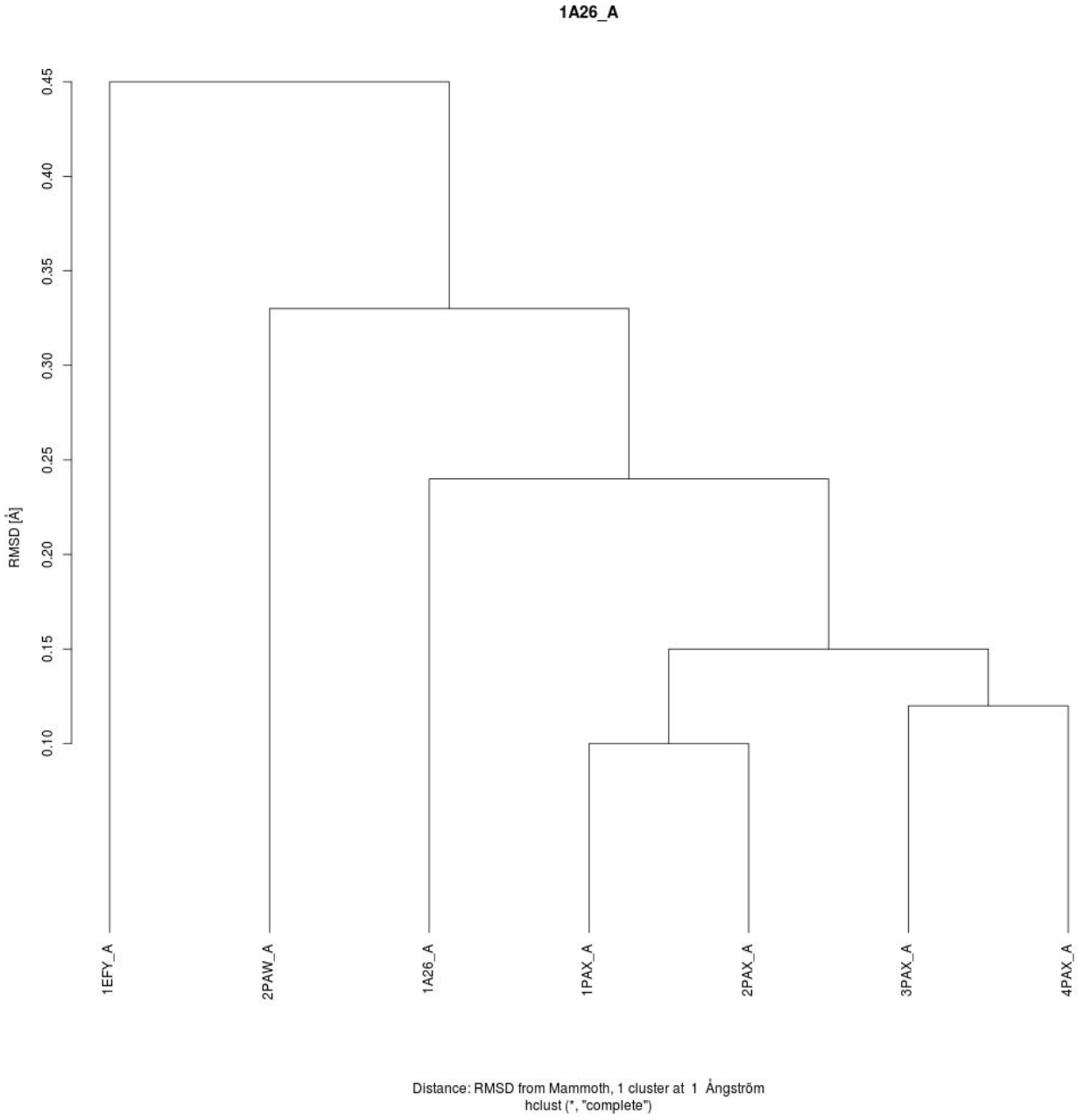

Figura A1.8. Dendrograma de proteína 1A26.

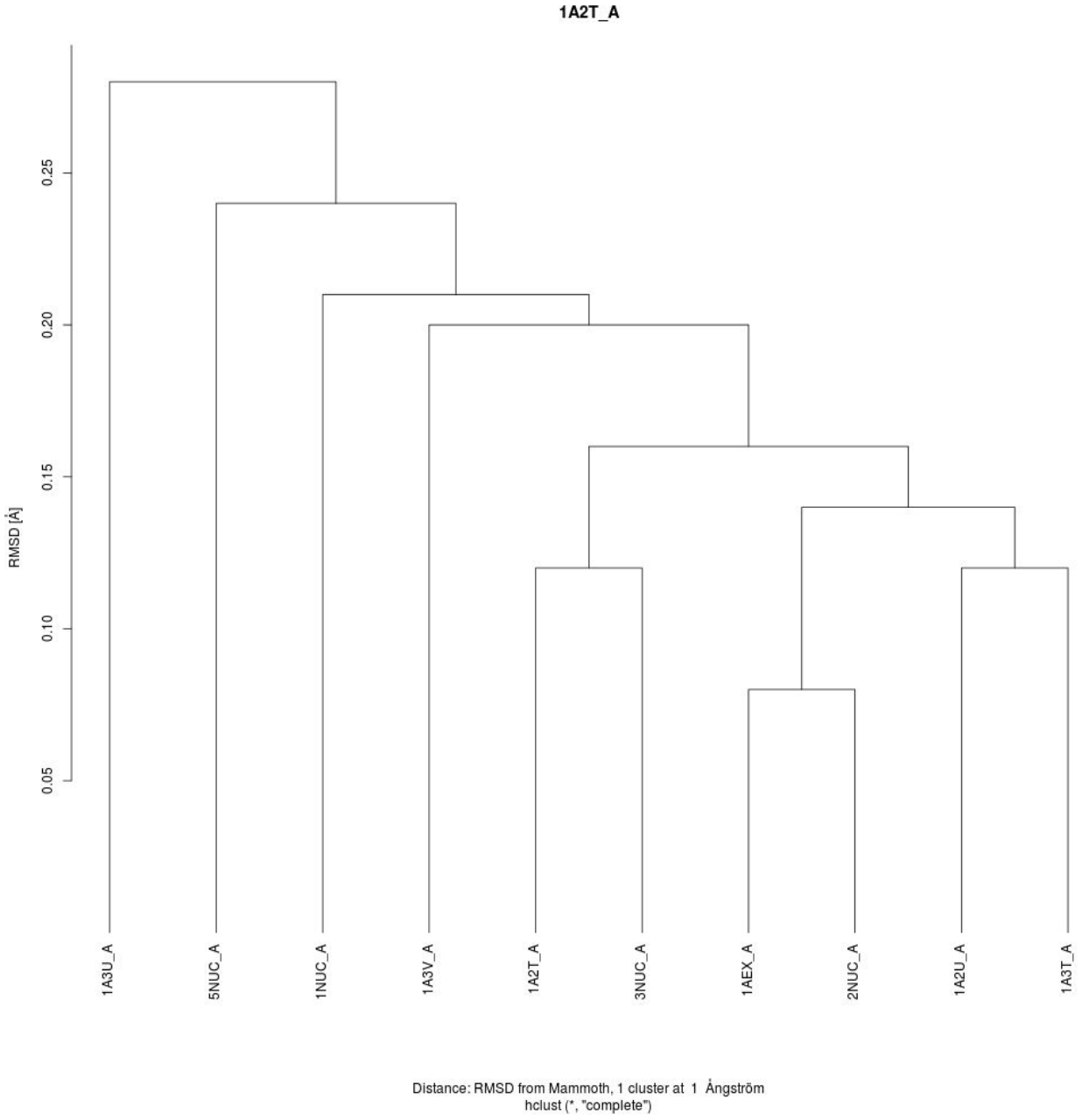

Figura A1.9. Dendrograma de proteína 1A2T.

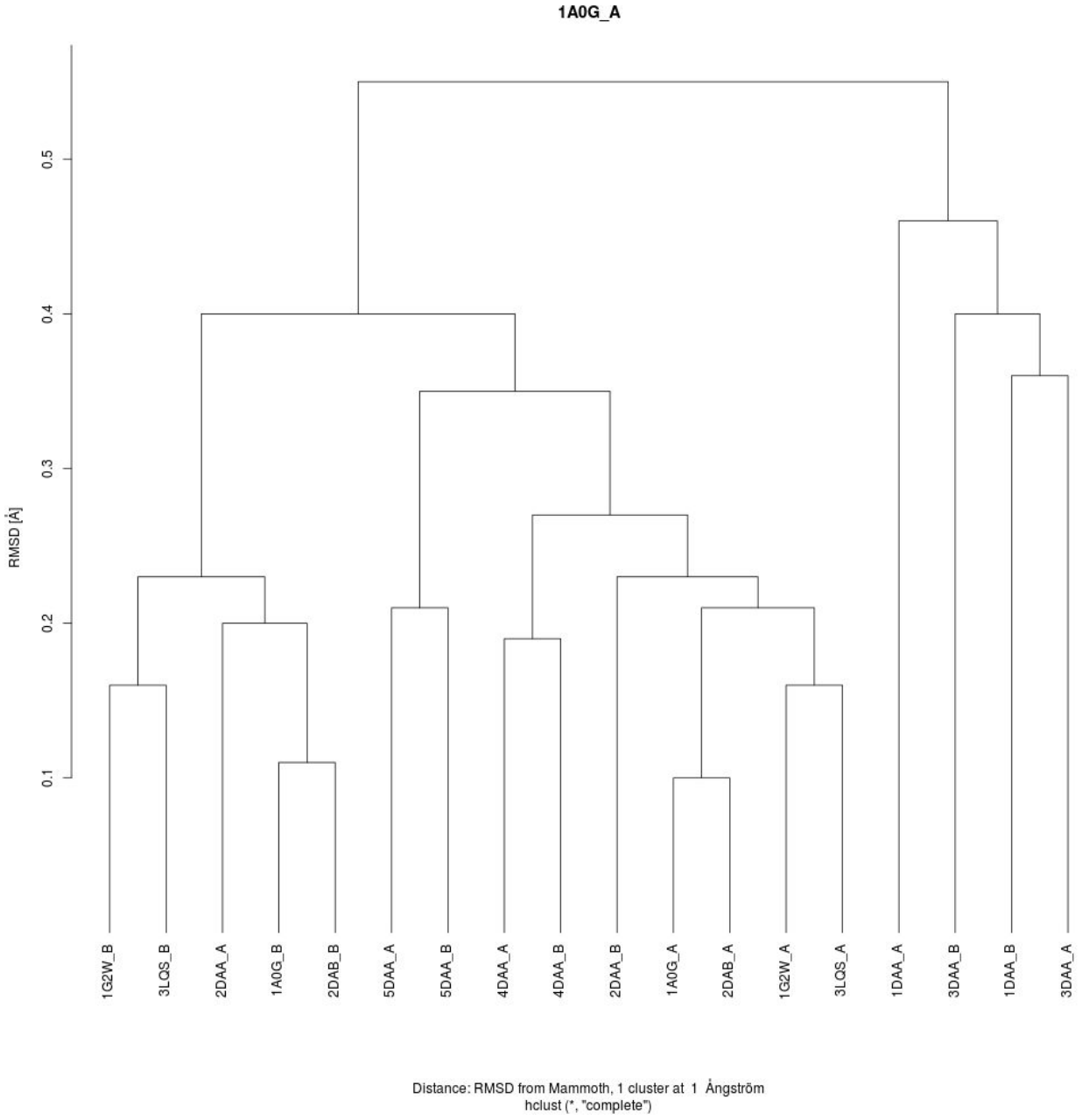

Figura A1.10. Dendrograma de proteína 1A0G.

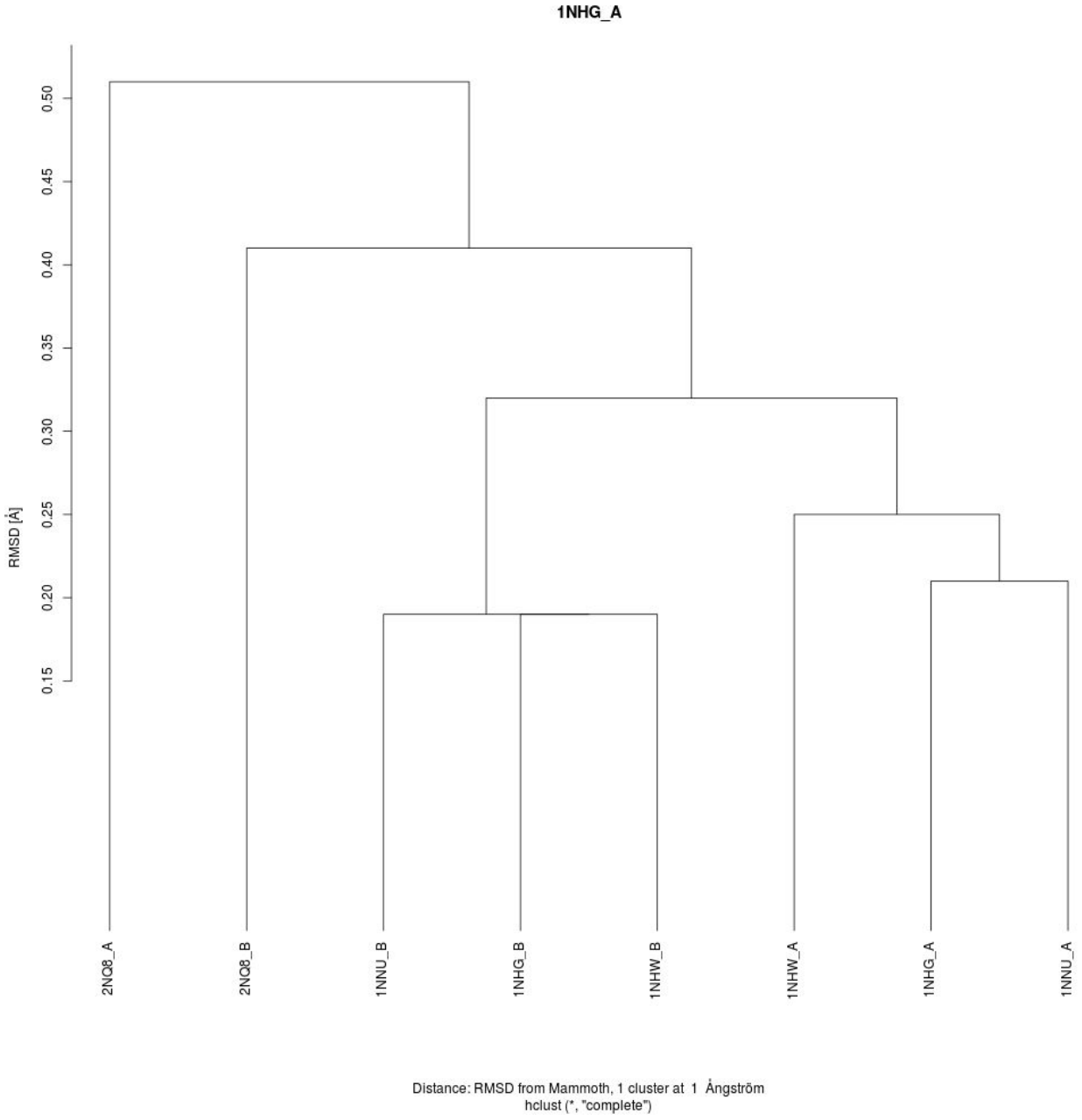

Figura A1.11. Dendrograma de proteína 1NHG.

Anexo 2

Los parámetros establecidos son los siguientes:

Se utiliza all-atom OPLS force field y spce como modelo de aguas [38,48]. Para la minimización se utiliza el algoritmo *steepest descent*, Nº máximo de pasos 50000, Fuerza máxima <1000.0 kJ / mol / nm: emtol = 1000.0, paso de minimización = 0.01, rcoulomb = 1.0, Van der Waals: rvdw = 1.0, pbc = xyz. Para la primera parte de la equilibración se utilizó un ensamble NVT o canónico (nº de partículas, volumen y temperatura constantes). Se aplican velocidades, se usa un integrador leap-frog, tiempo = 100 ps, paso de tiempo = 2 fs, rcoulomb = 1.0, Van der Waals: rvdw = 1.0, pbc = xyz, LINCS [49], coulombtype = PME (Particle Mesh Ewald), T = 300 K, tau_t = 0.1, método = velocity rescaling thermostat [50]. Para la segunda parte se utilizó un ensamble NPT (nº de partículas, presión y temperatura constantes), donde no se generan nuevas velocidades, y se agrega la presión con los siguientes parámetros: método = Parrinello-Rahman, tipo = isotrópico, tau_p = 2.0, ref_p = 1.0, compressibility = 4.5e-5. Para la simulación de la trayectoria se usa un integrador leap-frog, tiempo = 100 ps, paso de tiempo = 2 fs, rcoulomb = 1.0, Van der Waals: rvdw = 1.0, pbc = xyz, LINCS, coulombtype = PME (Particle Mesh Ewald), T = 300 K, tau_t = 0.1, método = velocity rescaling thermostat. Se utilizó un ensamble NPT, no se generan nuevas velocidades, método = Parrinello-Rahman, tipo = isotrópico, tau_p = 2.0, ref_p = 1.0, compressibility = 4.5e-5. Tiempo de simulación = 200 ns.