Universidad Zaragoza
1542

Trabajo Fin de Grado

Evaluación del mutoma asociado al síndrome de Rett
Evaluating the mutome of Rett syndrome

Autor/es

Sergio Lafuente Avellanas

Director/es

Juan José Galano Frutos
Javier Sancho Sanz

Facultad de Ciencias / Departamento de Bioquímica y Biología Molecular y Celular

Curso 2018-2019

Repositorio de la Universidad de Zaragoza – Zaguan http://zaguan.unizar.es

ÍNDICE

RESUMEN

El síndrome de Rett es un trastorno neurológico que afecta principalmente al desarrollo y maduración de las neuronas, siendo éste exclusivo para las niñas. Se caracteriza por fallos en la coordinación léxica y el uso de las manos para acciones conscientes, apareciendo los primeros de los síntomas entre los 6 y 18 meses de edad. Con el tiempo, los síntomas derivan en efectos tales como autismo sindrómico, ataxia, etc.

Dicho síndrome toma origen en disfuncionalidades de la proteína codificada por el gen MeCP2, provocado por mutaciones deletéreas que afecta directamente a su estructura nativa, y el mal funcionamiento de ésta. Por ello, se provocan efectos tóxicos en la célula por la mala transcripción de proteínas, ya que la función de la proteína afectada es el silenciamiento de otros genes.

Para estudiar el origen de dicho trastorno, se eligieron 10 mutaciones en total a lo largo de la hélice alfa, presente en el dominio MBD de la proteína, entre las cuales, 8 de ellas son patogénicas y 2 benignas. Se modelaron a nivel atómico las mutaciones propuestas y la proteína nativa por medio del método de dinámica molecular, para así optimizar las condiciones (en concreto la temperatura, con un óptimo encontrado a 360K), utilizando un proceso estándar de preparación, calentamiento, equilibrado y producción. Se analizaron los resultados a partir de dos técnicas de análisis: análisis de propiedades generales y clustering basado en RMSD-2D. En las 7 mutaciones con resultados fiables, se ha obtenido un acierto en la predicción del 85%. En vista de los resultados, una mayor fiabilidad y precisión se podrían obtener desarrollando métodos para el estudio de los extremos de la estructura y de la interacción con otras proteínas.

ABSTRACT

Rett syndrome is a neurological disorder that affects mainly to maturation and development of the neurons, being exclusive for girls. The most important effects are failures in the lexical coordination and the use of the hands for conscious actions, with an onset between the first 6 and 18 months of age. Progressively, the symptoms trigger other effects like syndromic autism, ataxia, etc.

The origin of the syndrome are the dysfunctionalities of the protein codified by the MeCP2 gen, provoked by deleterious mutations that affect directly to the native structure of the protein and its subsequent malfunction. Therefore, toxic effects appear in the cell because of a bad transcription of proteins, since the function of the mutated protein is the silencing of other genes.

In order to study the origin of this disorder, 10 mutations, located in the alpha-helix of the MBD domain of the protein, were chosen among which 8 of them are pathogenic, and the other 2 are benign. We simulated the mutants and the original protein by using molecular dynamics, and then optimized the conditions (specially the temperature, whose optimal value is 360K), using the standard procedure of preparation, heating, equilibration and production. The results were analyzed using two different techniques: general property analysis and RMSD-2D-based clustering. Considering the 7 mutants with reliable enough results, a 85% accuracy in the prediction was obtained. These results suggest that a higher reliability and accuracy may be achieved by the development of methods to study the ends of the structure and the interaction of the protein with its partners.

INTRODUCCIÓN

Las proteínas son biomoléculas constituidas por una secuencia de aminoácidos unidos por enlace peptídico, formando una estructura primaria de tipo cadena lineal. Dichas cadenas se organizan formando estructuras más complejas como pueden ser hélices alfa o láminas beta, lo cual se denomina estructura secundaria. Desde el punto de vista estructural las proteínas pueden sufrir alteraciones intrínsecas (dinámica conformacional propia de la proteína nativa) o extrínsecas (consecuencia de una o varias mutaciones), y dichos cambios en su conformación afectan directamente al comportamiento de la proteína en el medio. El buen comportamiento funcional de las proteínas depende en primer lugar de su estructura o conjunto de conformaciones promedio, cuya estabilidad y/o capacidad de interacción con otras biomoléculas no se ha de ver afectada por los factores extrínsecos antes mencionados o el medio en que se desenvuelven (el entorno fisiológico en el caso de las proteínas que residen en el interior celular).

El síndrome de Rett toma origen en disfuncionalidades presentes en la proteína MeCP2 (Methyl CpG Binding Protein 2) debidas en la mayoría de los casos a mutaciones en su secuencia primaria de aminoácidos. El impacto de mutaciones deletéreas en MeCP2 normalmente impide que la proteína adquiera su estructura nativa o hace que se vea afectada su estabilidad intrínseca de manera significativa y en consecuencia su funcionamiento, llegando a provocar efectos tóxicos en la célula por acumulación de errores en la transcripción de proteínas, ya que MeCP2 se encarga del silenciamiento de otros genes involucrados.[1]

Varios son los métodos a través de los cuales puede estudiarse el comportamiento de las proteínas frente a mutaciones, tanto experimental como computacionalmente. Entre estos últimos uno de los más ampliamente utilizados son los métodos de dinámica molecular (DM). El reciente avance de la tecnología ha propiciado el aumento significativo en la potencia de cálculo, y ha incentivado el desarrollo de nuevos métodos y técnicas de DM, así como de los modelos físicos (campos de fuerza y modelos de agua) que subyacen estas aplicaciones de modelado biomolecular. Tales avances han sido de gran utilidad para el estudio de sistemas biológicos y han permitido complementar resultados obtenidos a través de la experimentación *in vivo / in vitro*, en particular cuando es difícil o imposible obtener información a partir de la misma. Los resultados obtenidos a partir de métodos computacionales han sido de tal impacto que su uso se ha extendido por campos de la química o la biología, como son las químicas inorgánica y orgánica o las biologías molecular y estructural.[2]

Los métodos de simulación por DM han sido ampliamente utilizados para el estudio de estabilidad y dinámica de proteínas, por su exactitud y la posibilidad de modelización en solvente explícito, la inclusión de iones, de temperatura, entre otros que permiten simular condiciones fisiológicas. En los últimos años ha habido un aumento de la fiabilidad y la precisión en los campos de fuerzas y los modelos de agua explícita utilizados en los programas de simulaciones de DM, así como una mejora en los métodos de muestreo conformacional. En un reciente estudio realizado en el laboratorio donde este trabajo tiene lugar, por Vladimir Espinosa y colaboradores, se llevó a cabo un estudio para el análisis de la estabilidad intrínseca sobre todo el espacio mutacional del dominio LA5 perteneciente al receptor de LDL relacionado con la enfermedad de la hipercolesterolemia.[3] Tal análisis fue realizado a partir de la realización y análisis de trayectorias de DM, combinado con información sobre las interacciones proteína-proteína de esta proteína con sus parejas fisiológicas. En el trabajo se logró predecir de forma satisfactoria la patogenicidad de 49 (de un total de 50) mutaciones conocidas en su momento, superando así el poder predictivo en relación con el efecto de las mutaciones mostrado por otras herramientas de predicción populares entre la comunidad científica, como PMUT[4], Condel[5] y Polyphen2[6], que están basadas en elementos más primarios como son la secuencia de aminoácidos, o propiedades estructurales derivadas de la proteína rígida.

OBJETIVOS

El objetivo general de este proyecto es predecir la estabilidad del dominio de unión a *DNA* (*MBD, por sus siglas en inglés*) de la proteína MeCP2 -cuya estructura es perfectamente conocida- frente a varias mutaciones localizadas en un subdominio concreto del mismo.

Este subdominio de MBD consiste en una hélice alfa (la única que contiene el subdominio MBD), sobre la que se escogieron 10 mutaciones distintas utilizando diferentes bases de datos como Clinvar o RettBASE (ver apartado de Materiales y Métodos), entre las cuales 8 han sido reportadas como patogénicas, y 2 como benignas. Se supone desde el principio del proyecto que la patogenicidad de la mutación proviene de una desestabilización de la proteína, y por tanto que sea de tipo benigno no es causa de ningún efecto desestabilizante proveniente de la alteración de la secuencia lineal de aminoácidos, pero no tiene por qué ser de esta manera.

Para dar cumplimiento al objetivo general antes planteado, proponemos los siguientes objetivos específicos:

1) Modelar a nivel atómico las mutaciones propuestas y la proteína nativa utilizando el método de simulación de dinámica molecular con solvente explícito.

2) Determinar las condiciones idóneas de simulación, en particular la temperatura, que permitan estudiar el efecto que provocan las mutaciones en un tiempo razonable y sin que condicionen el resultado final.
3) Aplicar y evaluar metodología *ad-hoc* de análisis de trayectorias de DM desarrollada previamente en el grupo ProtMol del Departamento de Bioquímica y Biología Computacional, para establecer el impacto sufrido por el dominio MBD frente a las mutaciones propuestas en este estudio.

CAPÍTULO I: MARCO TEÓRICO

1.1. Síndrome de Rett

El síndrome de Rett es un trastorno genético neurológico y del desarrollo que afecta al desarrollo y maduración de las neuronas, exclusivo en niñas, que provoca retrasos mentales y pérdida de coordinación motriz.[7] Su incidencia es de 1 por cada 10 000 mujeres a nivel mundial, y en España afecta hoy en día a aproximadamente 6 000 niñas.[8] A nivel clínico, el cuadro de la enfermedad se caracteriza por pacientes con fallos en la coordinación léxica y en el uso de las manos para acciones conscientes, los cuales no comienzan a aparecer hasta los 6-18 meses de edad. Tales síntomas pueden derivar con el tiempo en autismo sindrómico, ataxia, microcefalia, entre otros, si bien posteriormente el estado de las pacientes se vuelve estable en la edad adulta.[8]

Realmente existen tres formas clínicas que se manifiestan en los pacientes que padecen dicha enfermedad: clásica o típica, variante o atípica, o con discapacidades de aprendizaje leve. La más frecuente es la forma clásica, que representa el 75%, y cuyo cuadro clínico es el descrito en el párrafo anterior, mientras que la forma con discapacidades de aprendizaje leve presenta un número de incidencia muy bajo.

Como ya se ha mencionado, esta enfermedad solo afecta a niñas (el número de casos en varones es prácticamente nulo, menos del 0,5%[9]). La razón de este hecho se encuentra en que las mujeres presentan dos cromosomas sexuales de tipo X, y los varones uno de tipo X y otro de tipo Y. El gen MeCP2 se encuentra en el cromosoma X, concretamente en el brazo largo de éste en la banda número 28 (Xq28), por tanto los varones, al presentar únicamente un cromosoma X, es imposible compensar fallos genéticos presente en MeCP2, y por tanto cuando tienen lugar mutaciones como estas, son incompatibles con la vida.[9]

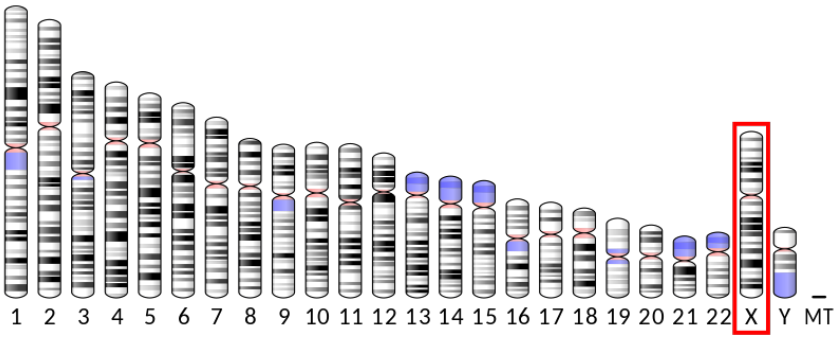


***Figura 1**: **Cariotipo representativo de un ser humano varón**. En éste se señala el cromosoma X. El gen MeCP2 que codifica la proteína se encuentra en el brazo largo del cromosoma en la banda número 28 (Xq28).*

La función concreta de la proteína codificada por el gen MeCP2 es la regulación de la expresión de otros genes distintos gracias a su unión con el DNA, además de ayudar a su empaquetamiento para dar lugar a la cromatina. El resultado de estas funciones es, por tanto, el silenciamiento o inhibición de los genes que se encuentran en contacto con la proteína según las necesidades de la célula en cada momento, ya que los factores de transcripción no son capaces de entrar en contacto con ellos por razones estéricas. Al sufrir una mutación, la proteína sufre cambios estructurales que derivan en su mal funcionamiento, teniendo como consecuencia un silenciamiento incompleto o ineficaz de los genes implicados. Dichos genes poseen funciones muy distintas, pero una de ellas es el desarrollo neuronal. Al no ser correctamente inhibidos o silenciados, se producen efectos en segundo plano que acaban derivando en el síndrome de Rett.[10] Esta idea demuestra la gran importancia de este gen, cuyo buen funcionamiento es fundamental para un buen desarrollo del individuo.

1.2. Aspectos estructurales de MeCP2

Este proyecto se centra en la proteína MeCP2, que presenta 6 dominios estructurales distintos, NTD, MBD, ID, TRD, CTDα y CTDβ, a lo largo de un total de 486 residuos. MeCP2 presenta aproximadamente un 40% de zonas estructuradas que se encuentran en los dominios MBD, TRD, CTDα y CTDβ, el resto son zonas intrínsecamente desestructuradas.[11]

***Figura 2**: **Imagen de los dominios que componen la proteína codificada por el gen MeCP2 representada por bandas**, destacando en rojo el subdominio MBD, que presenta el mayor número de mutaciones descritas como patogénicas en la proteína, que es la región de estudio de este proyecto (residuos del 90 al 163).*

La mayoría de las mutaciones que dan lugar al síndrome de Rett se encuentran en los dominios MBD (Methyl Binding Domain) y TRD (Transcriptional Repression Domain), aunque, de los dos, una mayor cantidad se encuentra en el dominio MBD.[11] Estas mutaciones en general afectan al plegamiento y estabilidad de MeCP2 y por ende a su comportamiento funcional.[11]

Las estructuras resueltas hasta día de hoy de la proteína codificada por el gen MeCP2, se indican en la Tabla 1. La estructura seleccionada para llevar a cabo las simulaciones en el presente estudio es la resuelta con el código 3C2I, que es íntegramente el dominio MBD de MeCP2.

Código PDB	Método de resolución	Resolución (Angstroms)
1QK9	NMR	
3C2I	Rayos X	2.5
5BT2	Rayos X	2.2
6C1Y	Rayos X	2.3
6OGJ	Rayos X	1.8
6OGK	Rayos X	1.65

***Tabla 1**: **Códigos PDB de todas las estructuras resueltas de la proteína**, donde se incluye también el método de resolución empleado. Información extraída de la página Uniprot. En el presente estudio, se eligió la estructura con código 3C2I para llevar a cabo las simulaciones de DM.*

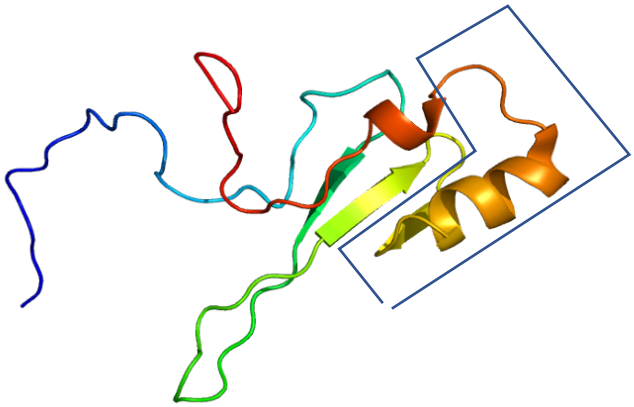


***Figura 3**: **Imagen del dominio MBD de la proteína codificada por el gen MeCP2**, extraída de una simulación.*

CAPÍTULO II: MATERIALES Y MÉTODOS

Para dar comienzo a la investigación, se escogieron las mutaciones a partir de las bases de datos de mutaciones asociadas al síndrome de Rett (ClinVar[12], HGMD[13] y RettBASE[14]), extrayendo 10 reportadas que se localizan en este subdominio y que de las cuales, 8 son predichas por programas como SIFT[15] y PolyPhen[16] como patogénicas, y se han reportado como causantes de casos reales de la enfermedad, y otras 2 como benignas, además de la proteína nativa (WT). Hemos supuesto que la patogenicidad de la mutación proviene de un efecto desestabilizante en la proteína, por lo que hemos decidido que todas ellas, situadas a lo largo de la hélice alfa del subdominio, son idóneas para evaluar sus estabilidades a partir de la aplicación de un enfoque basado en simulaciones largas de dinámica molecular en los que se analiza la relajación y los eventos conformacionales significativos que tienen lugar a lo largo del tiempo simulado (ver Tabla 2).

Mutación	Tipo	Número de casos
R133C	Desestabilizante	217
S134C	Desestabilizante	21
K135E	Desestabilizante	8
L138S	Desestabilizante	1
A140V	Desestabilizante	28
Y141C	Desestabilizante	5
K144R	No desestabilizante	1
D147N	No desestabilizante	1
P152R	Desestabilizante	71
F155S	Desestabilizante	2

***Tabla 2**: **Cuadro descriptivo de todas las mutaciones que se tratan en este proyecto**, indicando el tipo de efecto que ejercen sobre la proteína, así como su frecuencia. El número de casos se extrajo de RettBASE, y el tipo al que pertenece cada mutación de Clinvar.*

Como se ha descrito antes, el método seleccionado en el presente trabajo es el de simulación por dinámica molecular a nivel atómico. Para proceder a simular por DM un sistema biológico son necesarias etapas previas de preparación, de establecimiento de las condiciones y equilibrado, para finalmente proceder a la etapa de producción de las trayectorias. Para ello se ha recurrido al programa GROMACS[17] en su versión 5.1.4.

A partir de la estructura resuelta en 3C2I de la proteína nativa, se intercambió cada aminoácido por su correspondiente sustituto para generar los mutantes seleccionados (Tabla 2). Se eligieron los rotámeros más favorecidos energéticamente y con menos impedimentos estéricos con los residuos vecinos, utilizando el programa SwissPDBViewer[18]. Se procedió a preparar una réplica de simulación por mutante:

1) Preparación: Etapa que consta de cinco pasos, en los que se realiza: 1) selección del campo de fuerza y el modelo de agua explícita a utilizar (en este estudio CHARMM22/CMAP[19] y TIP3P[20]), 2) minimización de la proteína en vacío, 3) solvatación con el modelo de agua seleccionado (TIP3P), 4) neutralización del sistema con contraiones, y 5) minimización del sistema ya solvatado y neutralizado.

La minimización en vacío se llevó a cabo a través del método Steepest Descent, y el objetivo es aliviar tensiones e impedimentos estéricos. La fuerza máxima aplicada a la proteína fue de 1 kJ/mol durante un total de 5000 pasos de 2 fs.

La solvatación de la proteína se lleva a cabo en una celda dodecaédrica truncada, dejando una distancia mínima de 1 nm entre el borde de la celda y el átomo de la proteína más cercano a ésta, y se añadió la cantidad mínima de iones cloruro necesarios para neutralizar la proteína, lo cual puede variar según la mutación, viéndolo de forma general.

2) Calentamiento: en esta etapa se calentó el sistema hasta la temperatura de trabajo a través de un termostato Berendsen y el uso de una rampa de temperatura de 50 K por cada 50 ps (condiciones NVT en cada escalón de la rampa), en pasos de 2 fs.

Tres temperaturas de trabajo fueron seleccionadas inicialmente, 310, 410 y 510 K. Una cuarta temperatura fue añadida posteriormente, 360 K, como resultado del análisis de las simulaciones con las 3 temperaturas anteriores y la elección de una temperatura óptima para el análisis (ver Resultados y Discusión).

3) Equilibrado: esta etapa del proceso es esencial, y se basa en simular un breve periodo de tiempo las proteínas para asegurar que el sistema esté equilibrado bajo las condiciones elegidas. Dos subetapas consecutivas de equilibrado fueron lanzadas. En una primera etapa de 250 ps (125 000 pasos de 0,002 ps cada uno) se utilizó un termostato V-rescale y se aplica una presión de 1 Atm a través de un barostato Parrinello-Rhaman (condiciones NPT). A la vez se aplicaron fuerzas que obliguen a los átomos pesados de la proteína a no moverse (restricciones). En la segunda subetapa igualmente bajo condiciones NPT y utilizando el mismo termostato y barostato se simularon 250 ps de la proteína, ahora sin aplicar restricciones a los átomos pesados (proteína liberada). Al finalizar la etapa se analizó si la proteína había alcanzado un estado de equilibrio a través de análisis rutinarios de distancia promedio entre átomos (RMSD) a lo largo de las trayectorias (ver Figura 4).

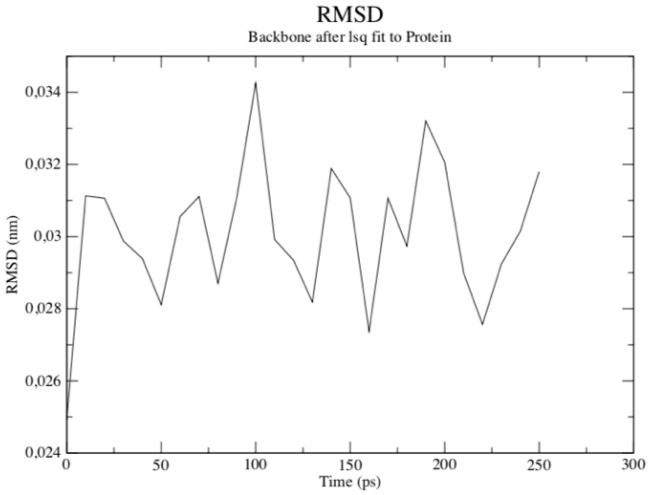


***Figura 4**: **Perfil de RMSD de la proteína nativa durante la etapa de equilibrado**. Se observan valores bajos, lo que indica un buen transcurso del proceso de equilibrado.*

4) Producción: etapa final en la que se ejecutó la simulación en sí de la proteína y se obtuvo una trayectoria a partir de la cual se extrajeron y analizaron los datos.

Inicialmente, se realizó una producción hasta alcanzar un tiempo de simulación de 300 ns, a partir de lo cual se eligió la temperatura óptima (que es uno de los objetivos del proyecto) para llevar a cabo el análisis previsto. Establecidas las condiciones, se extendió la simulación de las proteínas hasta alcanzar un tiempo de simulación de 1000 ns (1 μs). Para ello, se realizaron un total de 50 000 000 pasos de 0,002 ps cada uno, guardando todos los datos relevantes cada 100 ps.

Los datos obtenidos en la fase de producción se analizan a través de dos scripts (lenguaje de programación Python) elaborados *ad-hoc* en el grupo ProtMol que hacen uso a su vez de diferentes programas de análisis tanto del propio GROMACS como de otras de análisis de trayectorias como MDTraj[21]. El primero de los programas condensa el análisis de trayectorias a partir de obtener gráficas con los perfiles en el tiempo de simulación de los siguientes parámetros termodinámicos y estructurales:

- Energy and Pressure: energía y presión, nos informa sobre el correcto funcionamiento de la simulación.
- SASA: 'solvent-accesible surface area'. Superficie accesible al solvente.
- Gyration: radio de giro. Es una medida de volumen de la proteína.
- Alfa-Beta Structure: número de residuos en estructura alfa o beta.
- Coil Structure: expresa la estructura perfectamente desordenada.

- RMSF: ‘root mean square fluctuation’. Medida de movilidad interna de la proteína.
- RMSD: ‘root mean square deviation’. Medida de la distancia promedio entre los átomos (generalmente los del esqueleto o backbone) de las proteínas superpuestas. En el caso de trayectorias se comparan dichas distancias promedio de todos y cada uno de los *frames* de la misma con o el *frame* inicial o con otra estructura de referencia de la proteína.
- TM-Score: mide la similaridad entre las estructuras de dos proteínas.
- H-Bonds: número de puentes de hidrógeno formados entre los residuos de la proteína o con las moléculas de agua del entorno.
- Native contacts: contactos nativos de la proteína, nos indica cuanta interacción hay entre aminoácidos no consecutivos.

Este programa de análisis obtiene datos de todos estos parámetros para las trayectorias de forma global, y además analiza de forma local en torno al residuo mutado el número de puentes de hidrógeno, los contactos nativos y la superficie accesible al solvente, permitiendo así comparar el comportamiento de la proteína entera o por regiones. En definitiva, estas representaciones gráficas reflejan las variaciones en distintas propiedades de la estructura.

El segundo de los programas efectúa un análisis a través de la técnica de agrupamiento (clustering) basado en RMSD-2D, que permite comparar dos trayectorias de simulación de manera conjunta. Esto se basa en calcular el RMSD de cada *frame* con el resto de *frames* pertenecientes a la misma y los de una segunda. Se obtiene así una matriz donde en la posición (m, n) se calcula el RMSD de los *frames* m y n.

El concepto de cluster, centrándonos en el tema del proyecto, es un conjunto de estructuras muy parecidas entre sí, pudiendo asignar a cada *frame* de la trayectoria un cluster concreto. Utilizando esta técnica podemos obtener información sobre cuánto se diferencian dos estructuras a lo largo de sus respectivas simulaciones. Normalmente, podría decirse que la mayor utilidad del programa es al aplicarlos sobre dos trayectorias de simulación distintas pero con un mismo esqueleto (backbone), como puede ser el caso de una proteína nativa y su mutante. Cuando dos proteínas poseen una conformación distinta -según una distancia de corte previamente establecida por el usuario- el programa utilizado lo traduce como un cambio de cluster, ya que sus estructuras ya no se consideran similares. Por tanto, utilizando este tipo de análisis se podrá establecer de una manera más visual y fundamentada si se observan cambios estructurales significativos en la proteína tras mutarla en comparación con la proteína nativa, por ejemplo. En nuestro programa, el algoritmo de clustering utilizado es Agglomerative Clustering, utilizando el RMSD-2D como matriz de distancias.

En las gráficas obtenidas tras ejecutar el programa se representan el número de cluster (asignados de forma arbitraria) frente al tiempo (hasta 1000 ns), observando dos perfiles, cada una perteneciente a una de las dos trayectorias de las proteínas analizadas (la línea azul perteneciente a la proteína salvaje, y la naranja al mutante). Si en la gráfica se observa que las proteínas no presentan estructuras similares (se encuentran en clusters distintos, distinto comportamiento a lo largo del tiempo, entre otros criterios), se considera que la mutación es de tipo desestabilizante. A la hora de realizar el análisis, hay que asignar un valor que sirve como límite entre la conformación de un cluster y otro. Cuando la diferencia estructural de una proteína a lo largo de una trayectoria excede dicho valor, se produce un salto en la gráfica. Este valor es llamado distancia de corte (umbral o threshold), y se va ajustando según las preferencias y objetivos del usuario (en este caso, en la mayoría de los casos se asignó un valor de threshold de 0,25 nm).

CAPÍTULO III: ANÁLISIS Y DISCUSIÓN DE RESULTADOS

3.1. Temperatura

Centrándonos en la estabilidad de la proteína a lo largo del proceso de creación de la simulación, se ha comentado antes un cambio en la elección de la temperatura de trabajo, pasando de trabajar a 310K, 410K y 510K, a una temperatura intermedia próxima a su temperatura de desnaturalización media (360K). Esto se debe a que a las temperaturas de 410K y 510K, la proteína posee demasiada energía vibratoria y cinética cedida por el calor impuesto en la celda, que la obligó a desplegarse de forma irreversible, separándose de forma drástica de su disposición inicial. La temperatura que se elige debe de ser idónea para acelerar el muestro conformacional, sin llegar a provocar un desplegamiento brusco o la inestabilidad del sistema. Por tanto, los resultados obtenidos a 410K y 510K no fueron válidos, ya que no se podían interpretar correctamente. Se realizó un análisis para corroborar nuestras observaciones a través del análisis RMSD (root-mean-square deviation) de las trayectorias. Si la evolución de la gráfica es demasiado elevada, es un indicio de que la proteína no esta en condiciones estables. En la figura 5 se puede observar la gráfica correspondiente junto con un *frame* de la proteína desplegada a 510K.

Tanto por la trayectoria de la proteína (donde se observa que ya no hay aminoácidos que formen estructuras secundarias) como por los altos valores RMSD de la misma (hasta 1,5 nm), nos obligó a cambiar la temperatura a una intermedia, la cual ya permite trabajar en condiciones óptimas. Estos hechos ocurrieron tanto en la proteína salvaje como en todas las mutantes a las temperaturas de 410K y 510K. La elección de la temperatura idónea es crucial para que los resultados sean fácilmente

interpretables. La energía suministrada al sistema a la temperatura seleccionada para acelerar el muestreo conformacional no debe de cambiar significativamente el efecto intrínseco desestabilizante de la mutación en relación con la proteína nativa. Para poder usar de manera fiable este método de trabajo es importante poder contar con mutaciones cuyo efecto haya sido evaluado previamente (al menos en el fenotipo), así las proteínas cuyas mutaciones que son a priori de tipo desestabilizante deberán desplegarse (al menos parcialmente), mientras que la proteína nativa y las proteínas con mutaciones no desestabilizantes deberán mantenerse en sus estados plegados.

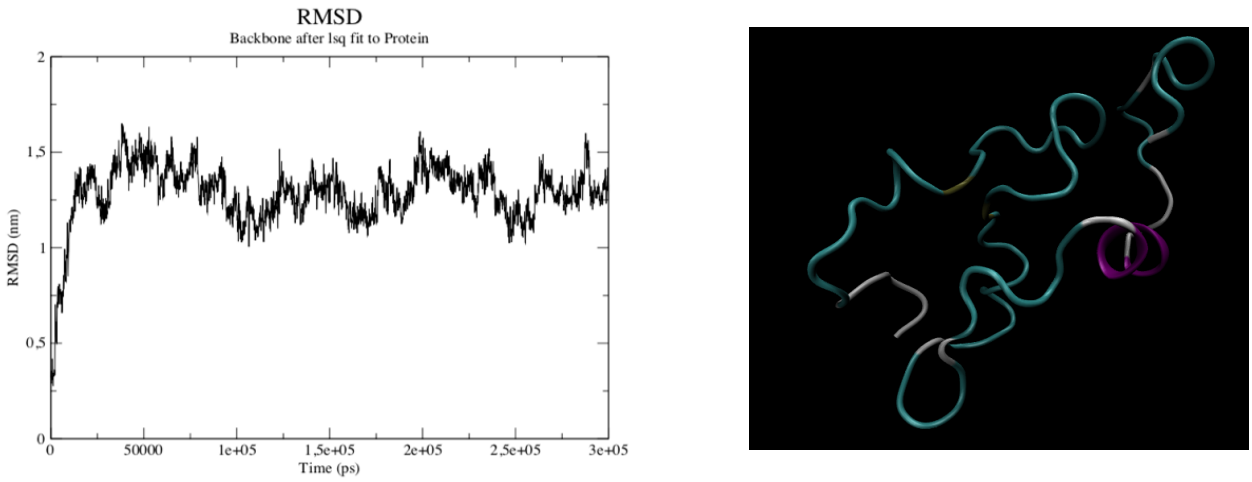
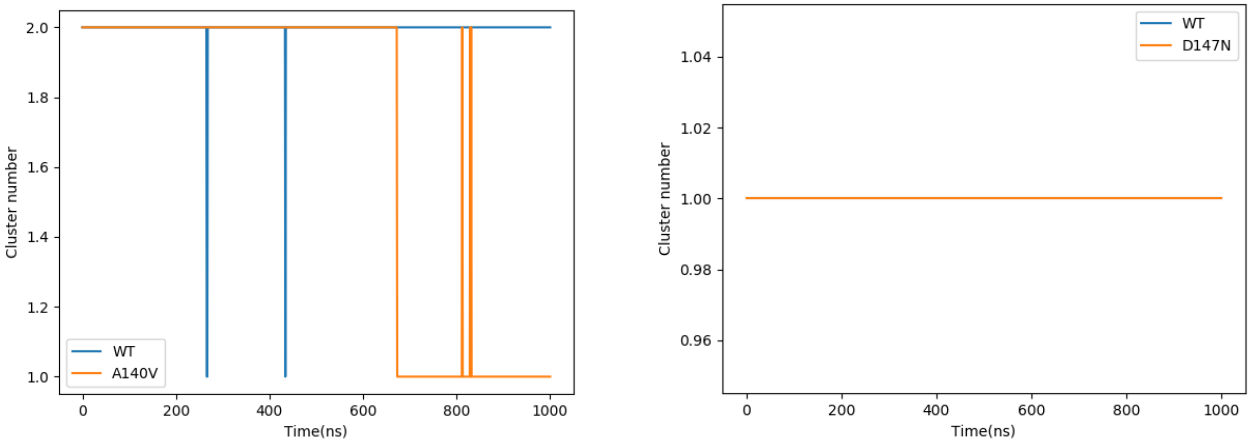


***Figura 5**: **Izquierda, perfil de RMSD de la proteína nativa durante la trayectoria de producción de 300 ns a la temperatura de 510 K**, mostrando un relativamente alto valor de RMSD con respecto a la estructura resuelta por rayos X. **Derecha, representación en cartoon de la proteína salvaje de la misma trayectoria**, mostrando el estado totalmente desplegado de la proteína.*

3.2. Resultados del análisis

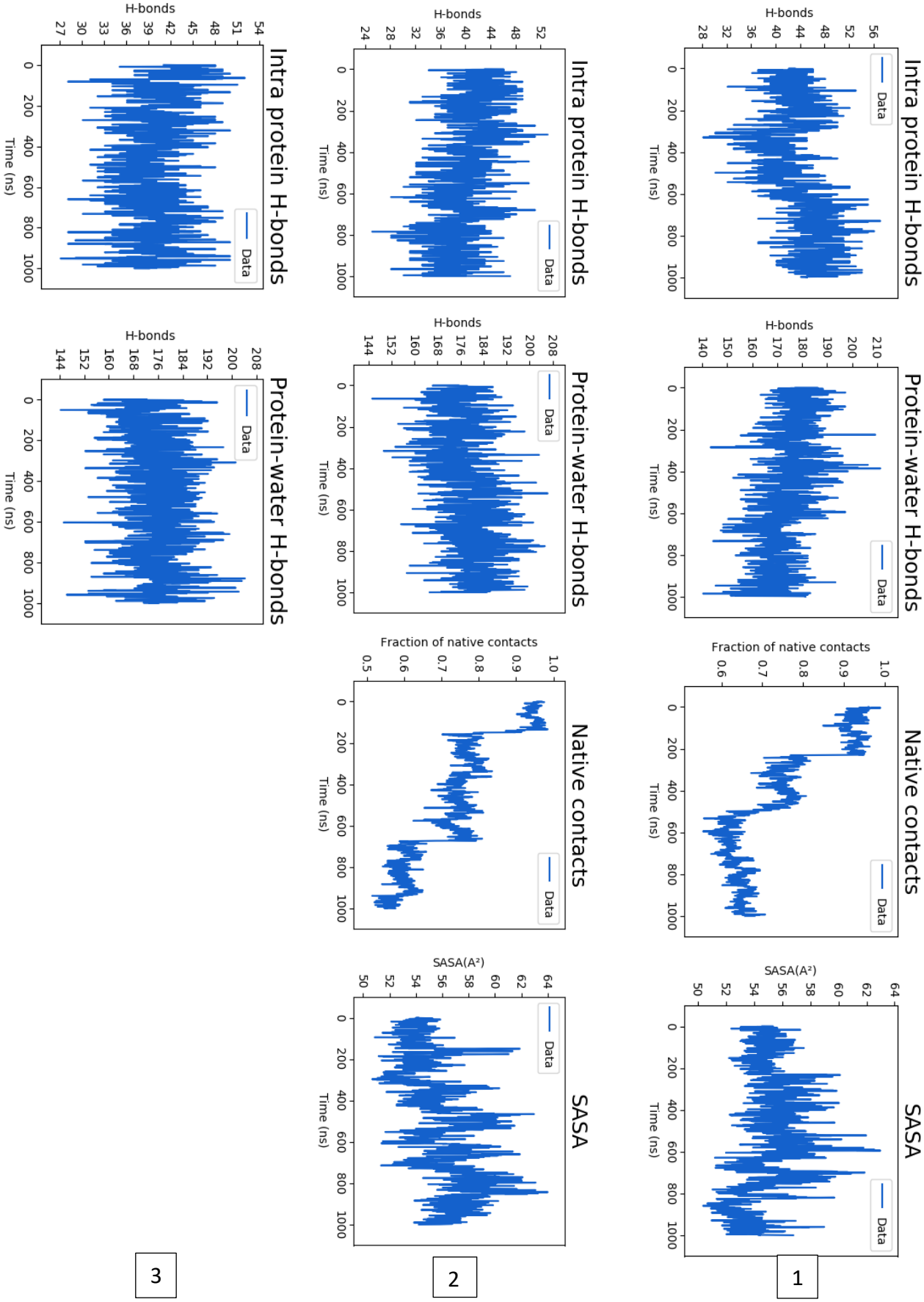
Para explicar brevemente las pautas a seguir para el análisis, en esta primera parte a modo de ejemplo se llevará a cabo un análisis comparativo de los resultados obtenidos para la proteína salvaje (WT), un mutante desestabilizante (en este caso A140V) y uno no desestabilizante (D147N) a la temperatura 360K, los cuales muestran los resultados más claros, pudiendo así detectar las diferencias en cuanto a efectos provocados en la proteína de forma más clara.



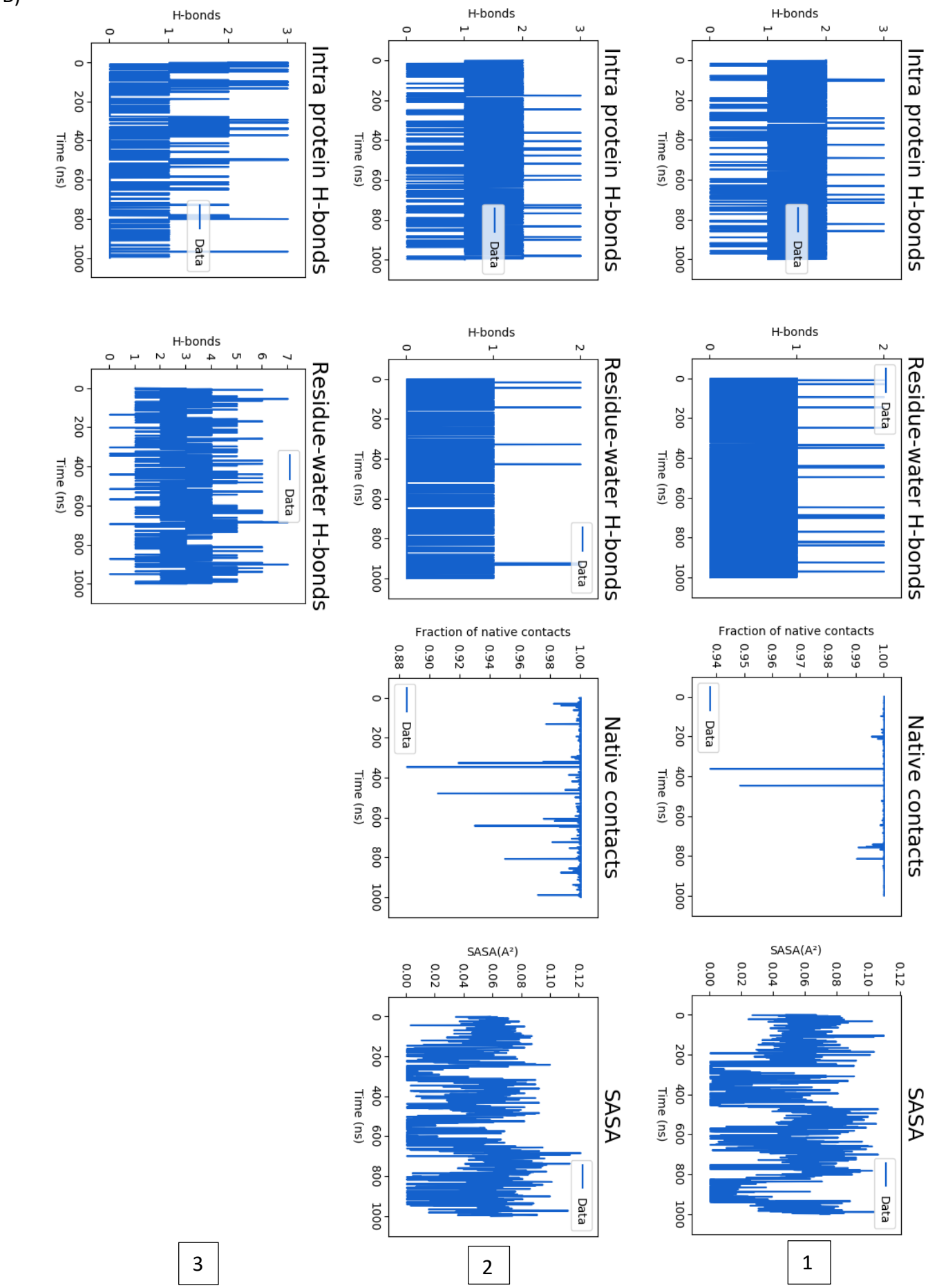
***Figuras 6*: Dos gráficas de clustering, siendo el de la izquierda la de la mutante A140V, y el de la derecha el D147N**. *Estos sirven de punto de partida para diferenciar el efecto de una mutación que provoca efectos desestabilizaciones en la proteína o no provocan ningún cambio en ésta. El valor umbral utilizado es de 0.25nm.*

Una clara diferencia se obtuvo entre la proteína mutante de tipo desestabilizante y la nativa a partir de la gráfica de la figura 6 (imagen de la izquierda). El mutante sufre cambios estructurales al cambiar un aminoácido por otro, provocando un desplegamiento que se traduce en saltos entre cluster distintos, ya que se ha sobrepasado la distancia de corte. Esto implica que la proteína mutante ha adoptado una conformación espacial distinta a la proteína nativa, por lo que la proteína tras el cambio de aminoácido no podrá realizar su función correctamente en el organismo. Como se puede observar en la imagen de la izquierda de la figura 6, a medida que va avanzando la simulación, las líneas pertenecientes a la proteína nativa y mutante siguen el mismo recorrido, lo que implica que su estructura es similar, pero al llegar a los 700 ns de simulación, la proteína mutante pasa del cluster número 2 al cluster número 1, por tanto su estructura ya es distinta a la original, continuando en este estado hasta el final de la simulación. Fijándonos a continuación en la gráfica de la derecha de la misma figura, perteneciente a la proteína mutante de tipo benigno, se observa que, durante toda la simulación, las trayectorias de ambas proteínas son similares, permaneciendo en el mismo cluster durante los 1000 ns, por lo que la mutación inducida en la proteína no es dañina para ésta, y por tanto no afecta a su estructura y función.

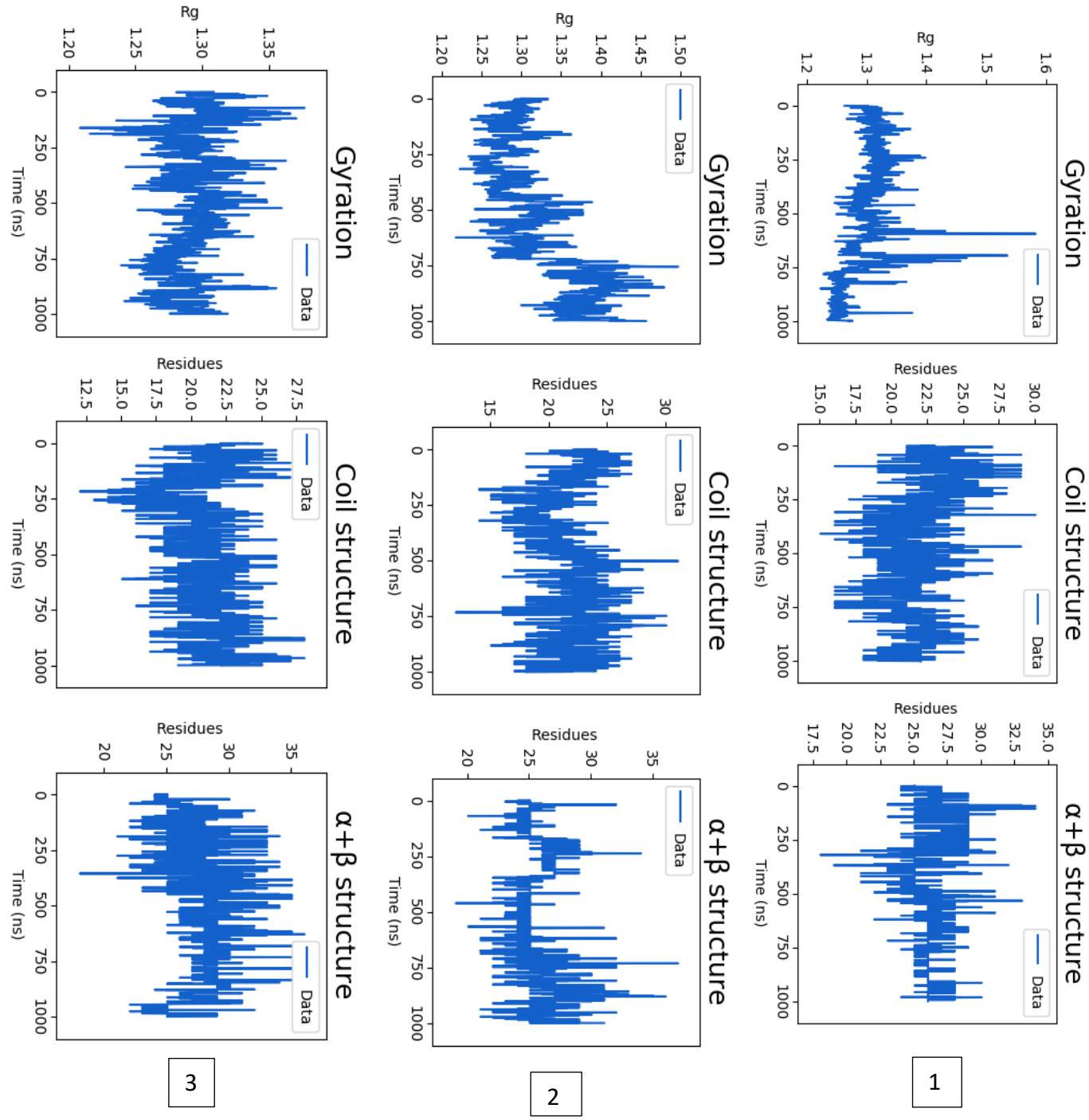
A)


Intra protein H-bonds
Protein-water H-bonds
Native contacts
SASA
Intra protein H-bonds
Protein-water H-bonds
Native contacts
SASA
Intra protein H-bonds
Protein-water H-bonds
H-bonds
Fraction of native contacts
SASA(A²)
Time (ns)
Data
1
2
3

B)


Intra protein H-bonds
H-bonds
Time (ns)
Data
Intra protein H-bonds
H-bonds
Time (ns)
Data
Intra protein H-bonds
H-bonds
Time (ns)
Data
Residue-water H-bonds
H-bonds
Time (ns)
Data
Residue-water H-bonds
H-bonds
Time (ns)
Data
Residue-water H-bonds
H-bonds
Time (ns)
Data
Native contacts
Fraction of native contacts
Time (ns)
Data
Native contacts
Fraction of native contacts
Time (ns)
Data
SASA
SASA(A²)
Time (ns)
Data
SASA
SASA(A²)
Time (ns)
Data
1
2
3

C)


Gyration
Coil structure
α+β structure
Rg
Residues
Time (ns)
Data
1
2
3

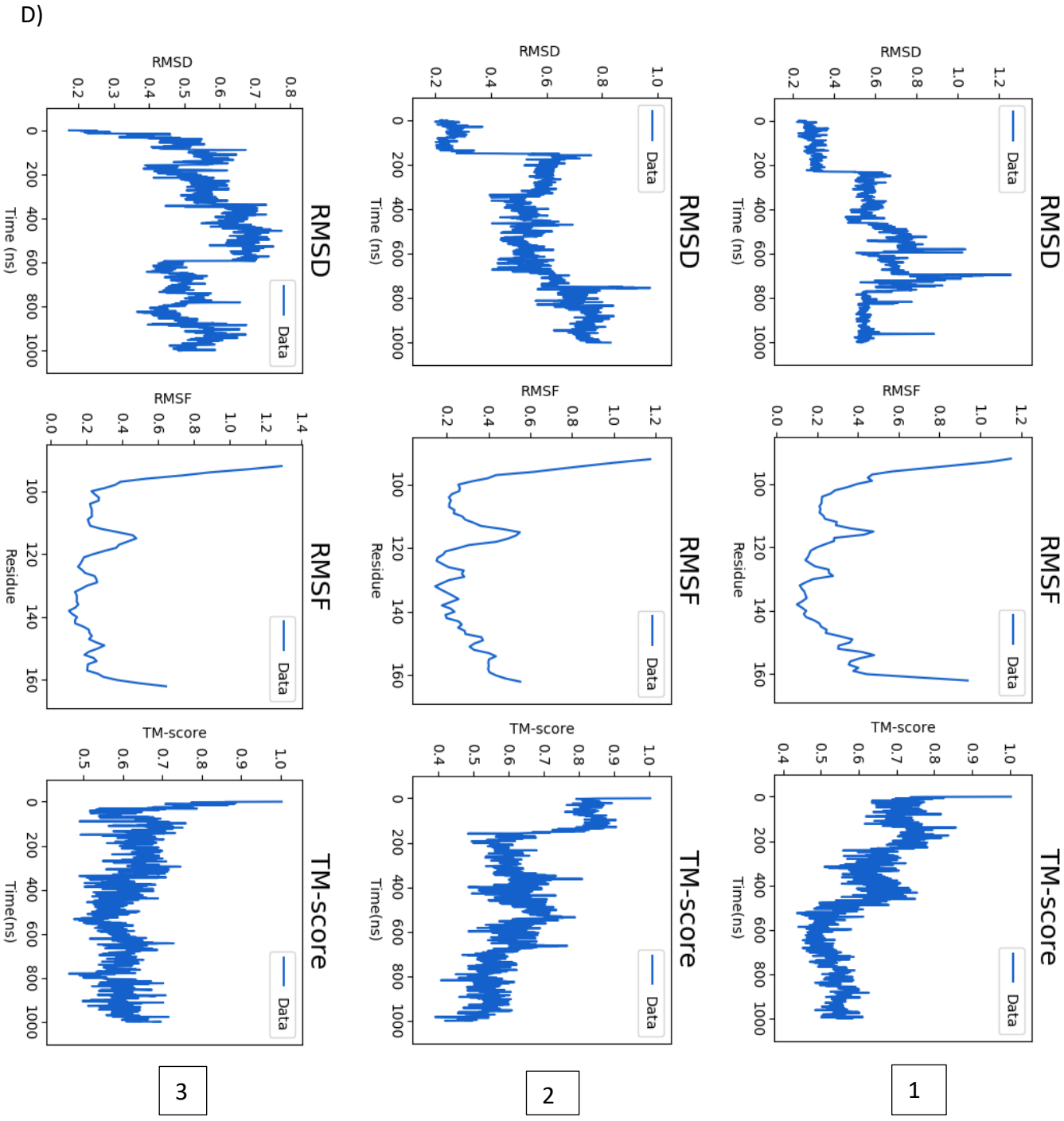


***Figura 7**: **Conjunto de gráficas de análisis de la estabilidad relativa de la proteína. A) Global interactions. B) Local Interactions. C) Structure. D) Variance**. En cada apartado, las gráficas de cada proteína están divididas por filas, siendo 1) WT, 2) A140V (desestabilizante), 3) D147N (no desestabilizante)*

Todas las observaciones anteriores se pueden complementar con las gráficas de la figura 7. En el caso de la A140V, se observa una pérdida de los contactos nativos (figura 7A, asignado con el número 2) en comparación con la proteína nativa (asignado con el número 1) o un mayor contacto con el solvente (figura 7A), lo que corrobora que la proteína mutante se ha desplegado, confirmando la predicción que se hizo antes del análisis. También se pueden comentar otros parámetros como el radio de giro, mayor en el caso de la mutante que de la nativa (figura 7C), lo que indica un aumento de volumen de la mutante; unos valores de RMSF altos en la zona de 140 (residuo donde se ha realizado el intercambio entre un aminoácido y otro), lo que indica una mayor movilidad en esa zona de la proteína (figura 7D); o una ligera disminución del número de puentes de hidrógeno entre los aminoácidos de la proteína, a la vez que un ligero aumento en los puentes con las moléculas de agua que se encuentran en su entorno (figura 7A y B). En el caso de la proteína mutante de tipo benigno también se pueden confirmar lo esperado, ya que las gráficas que se han obtenido son similares a las de la nativa, alcanzando valores del mismo rango que nos indican que la proteína no se ha visto afectada por el cambio de aminoácido, siendo una mutación no dañina que no dificulta la realización de la función de la biomolécula.

Por tanto, el método seguido para sacar conclusiones sobre el efecto de una mutación sobre la proteína es, primeramente observar las gráficas de clustering para conocer el comportamiento de ambas proteínas, para luego confirmar lo esperado con el resto de gráficas, comparando los distintos parámetros entre las dos, y así saber con certeza cómo afecta el cambio a la estructura de ésta.

La realización de este método de análisis con cada una de las mutantes propuestas, ofreció los siguientes resultados, resumidos en la tabla 3.

Mutación	Predicción	Número de casos	Resultado
R133C	Desestabilizante	217	Desestabilizante
S134C	Desestabilizante	21	No desestabilizante
K135E	Desestabilizante	8	Desestabilizante
L138S	Desestabilizante	1	Desestabilizante
A140V	Desestabilizante	28	Desestabilizante
Y141C	Desestabilizante	5	Desestabilizante
K144R	No desestabilizante	1	Dudoso
D147N	No desestabilizante	1	No desestabilizante
P152R	Desestabilizante	71	Dudoso
F155S	Desestabilizante	2	Dudoso

***Tabla 3: Versión completa de la tabla 2, incluyendo los resultados tras el análisis**, en la cual las mutaciones se han ordenado por orden de aparición en la secuencia lineal de la estructura. Solo hay 1 caso en los que no se cumple la predicción propuesta al principio del estudio.*

Como visión general, se cumplen 6 de las predicciones propuestas previamente al análisis de la proteína, dejando otros 3 casos en duda, y otro donde se comprueba que el efecto es contrario al esperado. Se procede a comentar los casos especiales, dando una posible explicación de por qué han ocurrido estos hechos.

Las mutantes F155S y P152R, se encuentran entre los casos especiales de la investigación. Éstas presentan un comportamiento distinto a los demás casos, ya que se encuentran en el extremo de la estructura cristalizada y resuelta del PDB sobre el que se realiza el estudio (se recuerda que está comprendido entre el residuo 90 y el 163). Se muestran las gráficas de clustering de ambas mutaciones en la figura 10.

En el caso del F155S (izquierda), se observa una alta movilidad de ambas proteínas (tanto de la nativa como de la mutante), dado que hay numerosos saltos entre clusters, es decir, que adoptan distintas conformaciones espaciales a lo largo de la simulación, acabando con una conformación desigual entre ambas y con respecto a la estructura inicial, lo que se traduce en un número de cluster distinto. El hecho de que la proteína nativa tenga tanta movilidad (aspecto comprobado con su correspondiente RMSD), no ofrece confianza en el resultado ya que nuestro objetivo es que ésta quede estable durante los 1000 ns de simulación, mientras que la proteína que posee la mutación sí que presente cambios estructurales. Por tanto, si ambas proteínas presentan igual movilidad, no se puede tener certeza de que el efecto predicho sea el correcto. En el caso de la P152R (derecha), se observa un comportamiento similar entre las trayectorias de las proteínas a lo largo del tiempo, lo que nos hace sospechar que los resultados son dudosos e inconclusos, ya que el número de casos de síndrome de Rett que se han dado con esta mutación son varios, además de poseer también una alta movilidad como en el anterior caso. Tanto en un mutante como en otro, se llega a la conclusión de que los resultados no son fiables, y por tanto no se puede tener certeza de que el cambio de aminoácido pueda provocar verdaderamente efectos desestabilizantes en la estructura y cambios en el comportamiento de la proteína.

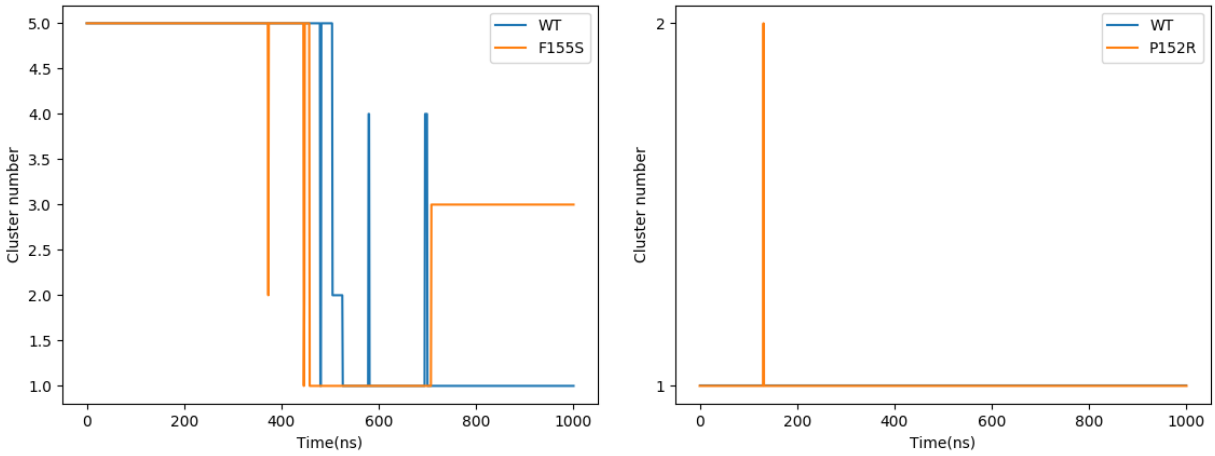


***Figura 8**: **Gráficas de clustering pertenecientes a las proteínas con mutaciones de tipo desestabilizante, siendo el de la derecha F155S y el de la izquierda P152R**. Se utilizó un valor umbral de 0.2nm.*

Todo lo anteriormente comentado se puede justificar con el hecho de que ambas mutaciones se hallan en el extremo del subdominio elegido para el estudio. Todos los extremos en las estructuras PDB suelen presentar una alta movilidad, además de que se desconoce los contactos entre aminoácidos no pertenecientes al subdominio, por tanto no se pueden sacar conclusiones de calidad con estos impedimentos. Se tendría que estudiar más en profundidad las estructuras vecinas, y utilizar métodos alternativos al utilizado que sean aplicables a residuos localizados en el extremo del subdominio para poder obtener resultados más fiables.

La mutación S134C también se cataloga como caso especial, ya que es el único entre todos los propuestos en el que la predicción propuesta ha sido errónea, ya que las gráficas nos indican que la proteína no se ve afectada por el cambio de un aminoácido por otro distinto en esa posición. En la figura 9 se observa que las trayectorias son similares, lo que indica que la estructura no se ha visto afectada. Se comprobó dicha suposición con el resto de las gráficas del otro método de análisis, las cuales confirman lo esperado ya que sus valores de RMSD son incluso más pequeños que los de la proteína nativa, además de mantener el número de puentes de hidrógeno intra-proteína prácticamente constantes y una movilidad (RMSF) similar. Por tanto, esta mutante no genera síndrome de Rett por desestabilización, sino que puede ser causado por otros motivos como un cambio en la interacción.

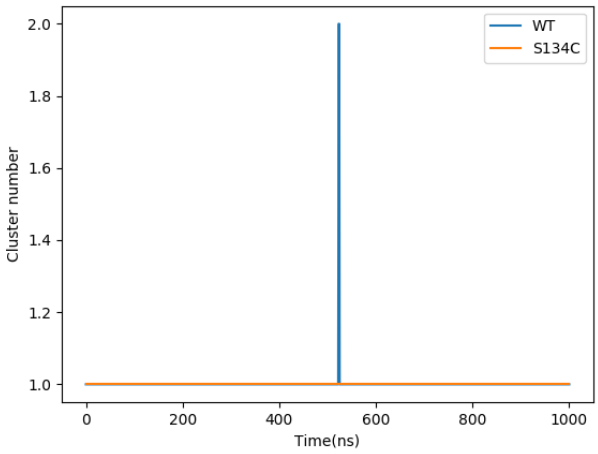


***Figura 9**: **Gráfica de clustering de la mutante S134C**, donde se observa que las trayectorias de la proteína mutada y la proteína original son similares exceptuando un instante a la altura de 550 ns. Se utilizó un valor umbral de 0.25 nm.*

La K144R es el último caso especial, la cual es supuestamente de tipo benigno, pero tras realizar las pruebas de análisis, no se pudo comprobar su verdadero efecto sobre la proteína. El número de casos vistos sobre esta mutación es solo de uno, es decir, que está muy poco estudiado. La otra mutación de tipo benigno, concretamente la D147N, también se ha visto en un solo caso, pero el origen de cada uno es completamente distinto. En el caso de la D147N, ésta proviene de condiciones hemicigóticas (es decir, que la mutación se presenta en un cromosoma que no tiene homólogo, como es en el caso de los hombres, que poseen un cromosoma sexual X y otro Y)[22], aspecto que no cumple la K144R. Además, en el estudio donde se realizó la investigación sobre dicha mutación se indica que solo se ha encontrado en 1 cromosoma entre 1380 afectado con dicho cambio[23]. En conclusión, aunque ambas mutaciones, supuestamente benignas, solo se hayan visto en un caso, la D147N se reporta como un caso de hemicigosis que si fuera patogénica, sería incompatible con la vida, pero en el caso de la K144R, el conocimiento sobre la procedencia de ésta es prácticamente nulo, y por tanto los resultados de los datos reales que hemos usado como punto de partida no son tan fiables como considerábamos inicialmente.

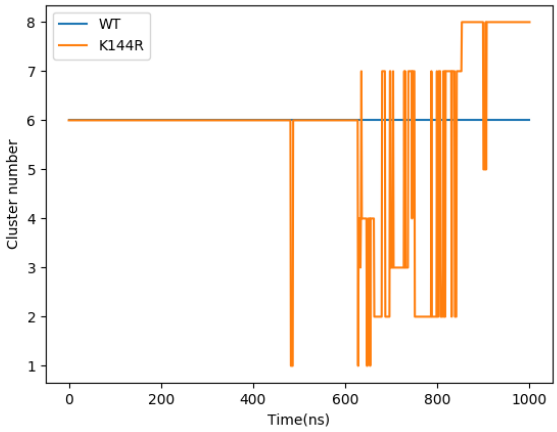


***Figura 10**: **Diagrama de clustering de la proteína mutante K144R**, de tipo supuestamente benigno. Se observan multitud de saltos, lo que indica que la proteína está adoptando conformaciones distintas a la de la proteína nativa, haciéndonos dudar sobre si el efecto provocado por la mutación sobre la proteína es benigno o desestabilizante. Por el casi nulo conocimiento sobre esta mutación, no se pueden sacar conclusiones fiables. El valor umbral elegido fue de 0.2nm.*

A manera de resumen, del análisis realizado de las 10 mutaciones de MeCP2 seleccionadas, en el caso de las 8 clasificadas a priori como desestabilizantes, nuestro método basado en simulaciones por DM permite validar tal clasificación en 5 de ellos (R133C, K135E, L138S, A140V y Y141C). Por otro lado, en el caso de las 2 clasificadas como no desestabilizantes, en 1 de los casos nuestro método coincidió con la clasificación previa (D147N). Es importante también destacar, que en los 3 casos en los que los resultados no son fiables, una explicación razonable ha sido encontrada. Entre los 7 casos de mutantes cuyos resultados no han sido dudosos, 6 de las predicciones que se hicieron sobre ellas fueron correctas, y solamente una errónea, por lo que hay un 85% de acierto. Entre los 10 casos, el 30% simplemente no ofrecía suficiente fiabilidad al estudio, dejándolos como casos en duda. Por tanto, el método de clustering no es el más adecuado para conocer la estabilidad de proteínas cuyas mutaciones se encuentran en el extremo de la estructura, dejando esta incógnita para futuras investigaciones.

CAPÍTULO IV: CONCLUSIONES

Las conclusiones más importantes para destacar sobre este proyecto son las siguientes:

- La temperatura idónea para simular el dominio MBD de la proteína MeCP2 y estudiar la estabilidad de sus mutaciones es la de 360K.

- Simulaciones de 1000ns (1 μs) de duración parecen ser suficientes a la temperatura seleccionada para evaluar la estabilidad intrínseca del dominio MBD.

- El porcentaje de aciertos en la predicción de la patogenia de las mutaciones en MeCP2 con este método es de un 85%, dejando en duda 3 casos en particular.

- Solo 1 caso entre los 10 fue realmente una predicción errónea, siendo patogénica (y por tanto, según nuestras suposiciones, desestabilizante), y tras el análisis, se comprobó ser de tipo no desestabilizante.

- Se necesitan técnicas de análisis alternativas a las utilizadas en este proyecto en proteínas cuyas mutaciones se encuentren en el extremo de la estructura.

- El método utilizado se puede complementar con otros métodos de interacción de la proteína, para casos en los que la patogenia pueda estar causada por una alteración de la interacción sin perjudicar su estabilidad.

BIBLIOGRAFÍA

[1] Valastyan, J. S., & Lindquist, S. (2014). Mechanisms of protein-folding diseases at a glance. *Disease Models & Mechanisms*, 9-14.

[2] Aleman, C., & Muñoz-Guerra, S. (2003). Aplicaciones de los métodos computacionales al estudio de la estructura y propiedades de polímeros. *Polímeros*, 250-264.

[3] Espinosa Angarica, V., Sancho, J., & Orozco, M. (2016). Exploring the complete mutational space of the LDL receptor LA5 domain using molecular dynamics: linking SNPs with disease phenotypes in familial hypercholesterolemia. *Human Molecular Genetics*, 1233-1246.

[4] V. López-Ferrando, A. Gazzo, X. de la Cruz, M. Orozco, J.L. Gelpí;
PMut: a web-based tool for the annotation of pathological variants on proteins, 2017 update. Nucleic Acids Research 2017.

[5] Improving the Assessment of the Outcome of Nonsynonymous SNVs with a Consensus Deleteriousness Score, Condel (2011). Abel González-Pérez and Nuria López-Bigas, American Journal of Human Genetics 10.1016/j.ajhg.2011.03.004.

[6] Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. Nat Methods 7(4):248-249 (2010).

[7] Bird, A. (2008). The methyl-CpG-binding protein MeCP2 and neurological disease. *Biochemical Society Transactions*, 575-583.

[8] Amir, R. E., Van Den Veyver, I., Wan, M., Tran, C., Francke, U., & Zoghbi, H. (1999). Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. *Nature Genetics*, 185-188.

[9] Trappe, R., Laccone, F., Cobilanschi, J., Meins, M., Huppke, P., Hanefeld, F., & Engel, W. (2001). MeCP2 mutations in sporadic cases of Rett syndrome are almost exclusively of parernal origin. *The American Journal of Human Genetics*, 1093-1101.

[10] Ehrhart, F., Coort, S., Cirillo, E., Smeets, E., Evelo, C., & Curfs, L. (2016). Rett syndrome – biological pathways leading from MECP2 to disorder phenotypes. *Orphanet Journal of Rare Diseases*, 11:158.

[11] Adkins, N., & George, P. (2011). MeCP2: structure and function. *Biochemistry and Cell Biology*, 1-11.

[12] Human Mutation Landrum MJ, Kattman BL. ClinVar at five years: Delivering on the promise. Hum Mutat. 2018 Nov;39(11):1623-1630. doi: 10.1002/humu.23641. [PubMed PMID:30311387].

[13] Stenson et al (2003), The Human Gene Mutation Database (HGMD®): 2003 Update. Hum Mutat (2003) 21:577-581.

[14] Krishnaraj R, Ho G, Christodoulou J. 2017. RettBASE: Rett syndrome database update. Hum Mutat 2017;00:1-10. Human mutation Pubmed: 28544139.

[15] SIFT web server: predicting effects of amino acid substitutions on proteins.Nucleic Acids Research, 2012 Jul; 40 (Web Server Issue): W542-7

[16] Vasily Ramensky, Peer Bork, and Shamil Sunyaev Nucleic Acids Res (2002) 30(17): 3894-900. European Molecular Biology Laboratory, Meyerhofstrasse 1, D-69117 Heidelberg, Germany.

[17] Berendsen, et al. (1995) Comp. Phys. Comm. 91: 43-56.

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

[19] Subramanian A, et al. A Next Generation Connectivity Map: L1000 Platform And The First 1,000,000 Profiles. Cell. 2017/12/1. 171(6):1437–1452.

[20] MacKerell, Bashford, Bellott, Dunbrack, Evanseck, Field, Fischer, Gao, Guo, Ha, et al, J Phys Chem, 102, 3586 (1998).

[21] Klepeis J.L., Lindorff-Larsen K., Shaw D.E. et al. Long-timescale molecular dynamics simulations of protein structure and function. Curr. Opin. Struct. Biol. 2009; 19: 120-127

[22] Schollen, E., Smeets, E., Fryns, J., & Matthijs, G. (2003). Gross Rearrangements in the MECP2 Gene in Three Patients With Rett Syndrome: Implications for Routine Diagnosis of Rett Syndrome. *Human Mutation*, 116-120.

[23] Maortua, H., Martinez-Bouzas, C., Garcia-Ribes, A., Martinez, M. J., Guillen, E., Domingo, M. R., Lopez-Arizteg. (2013). MECP2 Gene Study in a Large Cohort. Testing of 240 Female Patients and 861 Healthy Controls (519 Females and 342 Males). The Journal of Molecular DIagnostics, 723-728.