UNIVERSITAT DE BARCELONA

Estudio de las propiedades conformacionales de las proteínas mediante el uso de modelos de baja resolución basados en la discretización de las coordenadas internas

Francisco Martin Bandera

PROGRAMA DE DOCTORADO DE BIOMEDICINA

VALL D'HEBRON INSTITUT DE RECERCA (VHIR)
GRUPO DE BIOMEDICINA TRASLACIONAL

Estudio de las propiedades conformacionales de las proteínas mediante el uso de modelos de baja resolución basados en la discretización de las coordenadas internas

TESIS DOCTORAL

Francisco Martin Bandera

Abril 2018

UNIVERSITAT DE BARCELONA

UNIVERSIDAD DE BARCELONA

PROGRAMA DE DOCTORADO DE BIOMEDICINA

VALL D'HEBRON INSTITUT DE RECERCA (VHIR)
GRUPO DE BIOMEDICINA TRASLACIONAL

TESIS DOCTORAL

Estudio de las propiedades conformacionales de las proteínas mediante el uso de modelos de baja resolución basados en la discretización de las coordenadas internas

---

Director
Xavier de la Cruz Montserrat

Tutor
Albert Tauler Girona

Doctorando
Francisco Martín Bandera

Abril 2018

La ciencia está hecha de datos, como una casa de piedras. Pero un montón de datos no es ciencia más de lo que un montón de piedras es una casa.

Henri Poincaré

RESUMEN

El campo que estudia las propiedades y comportamiento de las proteínas así como las interacciones entre ellas es tan complejo que a día de hoy resulta complicada su caracterización en detalle, sea utilizando herramientas experimentales o mediante análisis de modelización, como la homología o las técnicas 'ab initio'. Debido a su importancia en las funciones biológicas y el amplio rango de actividades que desempeñan las proteínas, resulta indispensable la creación y utilización de nuevos modelos/programas que sirvan como herramientas eficaces para su estudio. Esto ha hecho que se haya desarrollado una serie de modelos de baja resolución que permiten realizar simulaciones cuando no es necesario un gran detalle estructural pero se requieren tiempos de simulación elevados.

En el presente trabajo hemos desarrollado un modelo propio y hemos explorado su viabilidad en dos casos de relevancia biológica.

El objetivo principal de este trabajo ha sido el desarrollo de un programa versátil que pueda utilizarse en diferentes tipos de estudios; se basa en un modelo simplificado de la estructura de la proteína que utiliza únicamente los carbonos alfa de la proteína y discretiza su espacio conformacional.

Las aplicaciones constituyen el segundo objetivo de este trabajo, y van encaminadas a valorar la aplicabilidad del programa. En un primer estudio hemos desarrollado un protocolo específico para el estudio de las mutaciones de las regiones de baja complejidad en las proteínas (LCRs), y extraemos conclusiones biológicamente relevantes sobre la patogenicidad de sus mutaciones.

En el segundo estudio hemos obtenido un bloque de resultados donde hemos analizado el efecto del volumen excluido sobre las correlaciones entre residuos. Esto nos ha permitido identificar la existencia de un efecto no biológico en el que las correlaciones aumentan cuando mayor es la distancia geométrica entre los residuos.

AGRADECIMIENTOS

Hay mucha gente a la que me gustaría dar las gracias por su apoyo constante y por aguantarme en los malos momentos. Gracias a Ignacio donde empecé a tomar contacto con la ciencia, a la gente de los laboratorios de cristalografía y de administración del IBMB algunos de los cuales todavía conservo el contacto y la amistad. Gracias a Albert, Roeland, Joan, Dani, Marta, Esther, Leonor y a otros que seguro me olvido por el paso del tiempo, que los aleja de mi memoria. Gracias a Xavi que me animó a seguir con un trabajo que ya había abandonado y que pensaba que no iba a continuar. Y por supuesto, gracias a mi familia, especialmente a mi madre a la que estoy seguro le hubiera encantado ver como finalmente acababa mi trabajo y a mi hermana que increíblemente no perdió nunca la fe en que esto pudiera llegar a su fin.

ÍNDICE GENERAL

LISTA DE FIGURAS

LISTA DE TABLAS

ABREVIATURAS, ACRÓNIMOS Y SÍMBOLOS

Å	Amstrong
Cα	Carbono alpha
CRMSD	Coordinate root mean-square deviation
DCC	Correlación dinámica condicional
DRMSD	Distance root-mean square deviation
LCR	Low complexity region
MD	Molecular Dinamics
MSCL	Canales mecánico-sensitivos de gran conductancia
PDB	Protein Data Bank
RMN	Resonancia magnética nuclear
UNRES	Modelo UNited RESidue
VE	Volumen excluido

A mi familia

CAPÍTULO I: INTRODUCCIÓN GENERAL

1.- LA IMPORTANCIA Y DIFICULTAD DE LA COMPRENSIÓN ESTRUCTURAL

Las proteínas son macromoléculas biológicas hechas de cadenas lineales de aminoácidos que se pliegan en una estructura tridimensional, formada a su vez por diferentes elementos de estructura secundaria [1]. Juegan papeles esenciales en los sistemas biológicos [1], actuando como material de estructura, catalizadores, transportadoras y almacenadoras de otras moléculas (e.g. el oxígeno asociado a la respiración), proporcionan soporte y protección inmune, generan movimiento, transportan impulsos nerviosos, controlan el crecimiento, la diferenciación celular, etc.

Varias son las propiedades que permiten a las proteínas participar en este amplio rango de funciones. Una de ellas es su gran variedad estructural, sorprendente si consideramos el número limitado de bloques de monómeros -los aminoácidos-, y que se explica por el orden que adoptan estos aminoácidos en la secuencia de la proteína. De hecho la estructura de las proteínas está directamente relacionada con su función en los procesos biológicos. Otra propiedad que explica la variedad de roles que desempeñan las proteínas es su capacidad para interactuar con otras macromoléculas biológicas, formando ensamblajes complejos con nuevas propiedades, no observadas en los componentes individuales. Estos ensamblajes, o complejos, participan en la precisa replicación del ADN (fig. 1.1), la transmisión de señales dentro de las células y otros muchos procesos esenciales. Finalmente, otra característica destacada de las proteínas es la rigidez/flexibilidad de sus estructuras. Aquellas más rígidas pueden funcionar como elementos de la estructura de los citoesqueletos o en la conexión de tejidos [53]. Las proteínas más flexibles pueden actuar como bisagras, resortes y palancas, cruciales para ciertos procesos funcionales como pueden ser los procesos alostéricos, la transmisión de información entre células (fig. 1.2), etc.

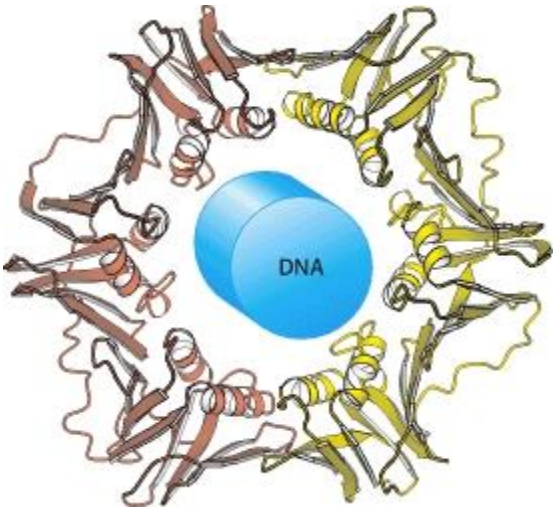


Fig. 1.1 Ilustración del complejo proteico que constituye la maquinaria de replicación del ADN, interaccionando con esta macromolécula.

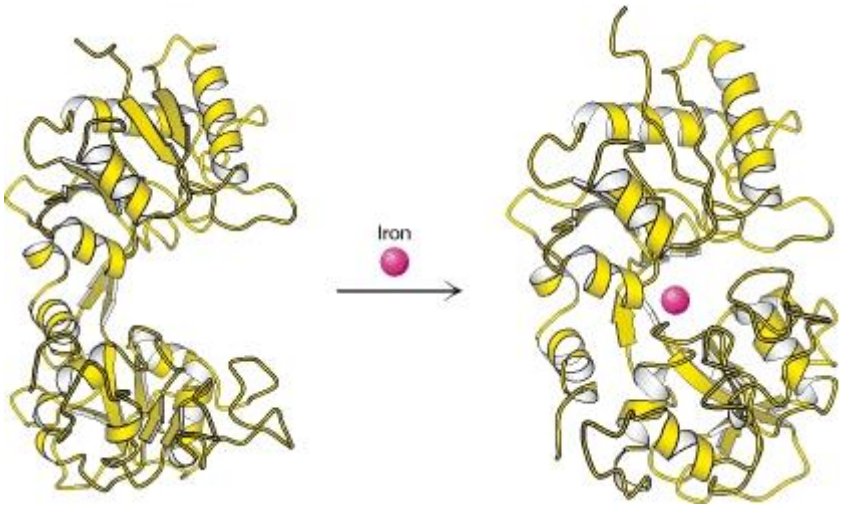


Fig. 1.2 Las proteínas detectan los cambios en el entorno. Aquí vemos cómo, en presencia de hierro, la proteína lactoferrina cambia su conformación estructural de forma que otras moléculas pueden distinguir entre las dos formas y participar en el transporte y regulación del hierro en las células [54]

Debido a la estrecha relación que une estructura y función en proteínas, se han hecho grandes esfuerzos para determinar con precisión la estructura de las proteínas. Los tres principales métodos para determinarla experimentalmente son: la cristalografía de rayos X, la espectroscopia de resonancia magnética nuclear (RMN) y la microscopia electrónica [55-57].

La cristalografía de rayos X usa el patrón de difracción que producen los rayos X al incidir en un cristal de la proteína de interés [94]. Este es un fenómeno físico complejo [58] en el que dicho patrón queda determinado por la densidad electrónica dentro del cristal irradiado. De forma más precisa, la difracción resulta de la interacción de los electrones con el haz de rayos X altamente energéticos; proceso en el que los electrones primero se activan, y al volver a su estado inicial emiten rayos X. Estos se superponen, amplificándose o cancelándose según si están en fase o no, lo que da lugar a un conjunto de interferencias, que constituyen del patrón de difracción. Este es recogido por un detector, y se utiliza para deducir la estructura tridimensional de la molécula, mediante diferentes técnicas numérico-computacionales. Las estructuras resultantes proporcionan una información muy rica de la proteína, a nivel atómico, mostrando detalles de ligandos, inhibidores, iones y otras moléculas presentes en el cristal. Sin embargo, esta técnica también tiene ciertas limitaciones [59]. Primero, el proceso de cristalización es difícil y no siempre es posible obtener cristales aptos para su uso. Segundo, la precisión de la estructura atómica obtenida depende de la calidad de los cristales, y esta puede variar bastante. Tercero, existen también limitaciones en la determinación estructural de las partes flexibles de las proteínas, ya que dichas partes no contribuyen al patrón de difracción obtenido.

La resonancia magnética nuclear también genera estructuras de alta resolución, pero utilizando una estrategia muy diferente [95]. En lugar de obtener la densidad electrónica de una molécula, RMN mide la distancia entre núcleos atómicos. La proteína es purificada y colocada en un campo magnético de alta intensidad, y luego se la somete a radio ondas. Los núcleos atómicos poseen una propiedad mecánico-cuántica denominada spin y es característico para cada tipo de núcleo. Si el valor es distinto de cero, entonces tendrá una diferencia en la distribución de la carga y por tanto un momento magnético nuclear que interaccionará con el campo magnético al que se vea sometido. Así, cuando el núcleo se somete a un campo magnético externo y se irradia con una radiofrecuencia, se produce un efecto de resonancia (desplazamiento de la orientación del spin respecto del equilibrio). Las resonancias resultantes son recogidas y analizadas, y dan lugar a una lista de distancias interatómicas que caracterizan la conformación local de los átomos enlazados o en contacto. Esta lista permite construir un modelo de la pro-

teína con la localización espacial de cada átomo. Una ventaja de esta técnica respecto a los rayos X es que utiliza proteínas en solución, en lugar de cristales; adicionalmente, las mediciones se realizan en rangos de tiempo muy cortos, lo que permite resolver la estructura de las partes flexibles de las proteínas. Por contra, esta técnica está normalmente limitada al estudio de proteínas pequeñas o medianas, ya que las proteínas grandes presentan problemas de solapamientos de picos en los espectros de RMN.

Finalmente, **la microscopia electrónica** utiliza haces de electrones para obtener imágenes directas, que después se integrarán para extraer la estructura 3D de la proteína [57]. Se suele utilizar para obtener la estructura de grandes complejos moleculares. Como normalmente las técnicas de microscopia electrónica no permiten obtener detalles alta resolución, se suelen combinar con técnicas de rayos X o RMN para generar detalle a nivel atómico.

Las dificultades técnicas de los tres procedimientos descritos anteriormente hacen que determinar una estructura molecular sea un proceso mucho más lento que el de secuenciación. De hecho, la aparición en los últimos años de técnicas de secuenciación de alto rendimiento [60] ha incrementado exponencialmente el número de secuencias de genes disponibles, creando una diferencia cada vez mayor entre la información estructural y de secuencia disponibles (fig. 1.3).

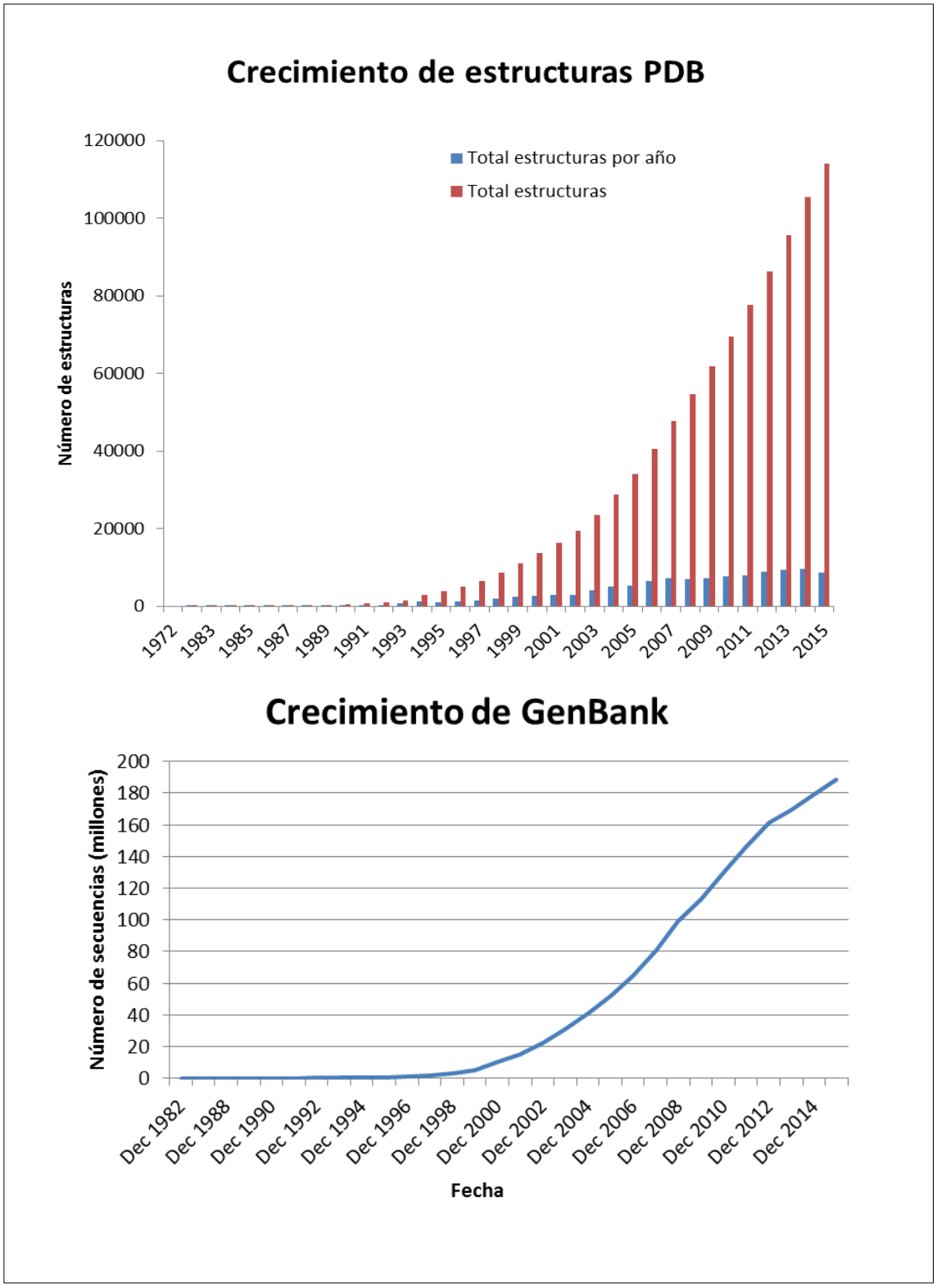


Fig. 1.3 El gráfico de arriba muestra el crecimiento en el número de estructuras obtenidas de forma experimental con las diferentes técnicas. El gráfico inferior describe el crecimiento de las secuencias depositadas en el Gene Data Bank.

Así, el número total de estructuras conocidas es mínimo, en relación al número de secuencias disponibles, lo que limita el uso de la información tridimensional a un reducido número de problemas biológicos. Ello ha hecho que los métodos de predicción de la estructura de las proteínas adquieran un papel relevante de cara a resolver este problema [61]. Estos métodos utilizan diferentes técnicas para modelar la estructura de las proteínas. Por ejemplo, pueden hacerlo utilizando información previamente disponible: es lo que conocemos como la modelización por homología [62]. O bien, pueden generar modelos para las proteínas sin homólogos conocidos [63], aquellas para las que aún no existe ninguna estructura tridimensional válida como plantilla. A continuación describo la esencia de estas técnicas.

La modelización por homología [62] se basa en la suposición de que dos proteínas homólogas tienen una estructura similar. Debido a que la estructura de las proteínas está más conservada que sus secuencias de aminoácidos, podemos modelar la estructura de una secuencia nueva con una buena precisión, siempre que exista una similitud apreciable entre sus secuencias [64]. Una de las dificultades en esta técnica es la de encontrar alineamientos de secuencia correctos [3] y otra es la de encontrar candidatos adecuados, que posean una estructura conocida, para poder utilizarlos de plantilla.

En ausencia de estructuras conocidas que nos puedan servir de plantillas en el proceso de predicción estructural, los **métodos de predicción utilizados son los llamados 'ab initio'** [65]. Están basados en un modelo físico empírico de la energía libre del sistema proteína-medio y en unos potentes mecanismos algorítmicos adecuados para simular el proceso de plegamiento. En esta familia de métodos encontramos los cálculos de minimización de energía y los de dinámica molecular [66]. A pesar de su valor, todavía presentan grandes limitaciones debido al extraordinario tiempo de cálculo requerido para identificar la conformación nativa de la proteína con precisión experimental [52,2].

Las dificultades con las que se han encontrado los métodos de predicción estructural anteriores, han dado lugar paulatinamente a una nueva forma de abordar el problema, en la que se eliminan de la representación de la proteína aquellos aspectos no relacionados con nuestro objetivo científico. Los modelos resultantes son los que se denominan modelos simplificados o de baja resolución (en inglés se

suele utilizar la expresión "coarse-grained") [67]. Estos modelos constituyen una alternativa práctica a las simulaciones que incluyen todos los átomos. El interés por las representaciones de baja resolución de las proteínas, como fuente de simulaciones teóricas de la estructura y dinámica de las proteínas, ha aumentado rápidamente [4-6]. Ello es debido a que requieren mucho menos esfuerzo computacional, lo que facilita la rapidez de las simulaciones de la dinámica temporal, de las vías de plegamiento, etc. De hecho presentan una ganancia computacional de unos cuatro órdenes de magnitud, aproximadamente, en comparación con las simulaciones que incluyen todos los átomos y el solvente [7]. Ello permite alcanzar la escala de milisegundos al realizar simulaciones de procesos biomoleculares. Otra aplicación de los modelos simplificados, es la predicción de la estructura de la proteína a partir de su secuencia de aminoácidos, debido a la enorme reducción del espacio conformacional que conllevan.

Por los motivos citados, el desarrollo y la aplicación biológica de los modelos simplificados de las proteínas es todavía un problema sin resolver, y constituyen el objetivo principal de esta tesis. A continuación describo con mayor detalle la situación actual de dichos métodos.

2.- LOS MODELOS SIMPLIFICADOS

Muchas de las dinámicas e interacciones relevantes dentro de las células (acoplamiento entre proteínas, reorganización sobre uniones de ligandos o después de reacciones bioquímicas, plegamiento) ocurren en la escala de los microsegundos o milisegundos e incluyen grandes conglomerados de macromoléculas. En estos procesos, el número de grados de libertad y la escala de tiempo son muy superiores a lo que podemos simular a nivel atómico con la actual potencia de cálculo. Además, en algunos casos la simulación a nivel atómico puede no ser lo más apropiado, tal como muestran los estudios experimentales a baja o media resolución de conglomerados de biomoléculas [8,9]. Los modelos de baja resolución surgen así como alternativa práctica al análisis estructural. Se han propuesto varios tipos de modelos basados en la simplificación de diferentes aspectos estructurales de la proteína. En estos modelos no se consideran explícitamente todos los grados

de libertad del sistema. Teóricamente, ello se basa en las diferentes escalas temporales presentes en los sistemas de macromoléculas. Es decir, si nos referimos como 'variables' a las coordenadas de ciertos tipos de átomos (e.g. hidrógenos, carbonos, etc) o grupos de átomos (cadenas laterales, cadena principal de los amino ácidos, etc) de una proteína, se sabe que la dinámica de las variables 'lentas' regula el comportamiento del sistema sobre escalas de tiempo largas, mientras que las restantes variables, con fluctuaciones mucho más rápidas, se equilibran rápidamente a cada nuevo valor de las variables lentas. Generalmente, las variables rápidas corresponderán a átomos ligeros o a pequeños grupos de átomos, dependiendo de la escala del análisis. Habitualmente, los modelos simplificados se basan en representaciones de la proteína que excluyen las variables rápidas. Sin embargo, ello entraña un problema técnico importante: la visión físicamente realista de la proteína se ve afectada, y se requiere una nueva parametrización del campo de fuerzas que rige en el modelo simplificado de la proteína. Es decir, hay que modelar enlaces químicos entre átomos virtuales, o pseudo-enlaces entre amino ácidos que carecen de cadenas laterales, etc. En los últimos años, se han generado una amplia colección de modelos simplificados, que representan diferentes compromisos entre precisión y transferibilidad, tal como veremos a continuación.

2.1- Los diferentes tipos de modelos simplificados

2.1.1- Modelos de redes elásticas

En estos modelos se representa la proteína como una red cuyos nodos están conectados por enlaces modelados como muelles elásticos. Cada nodo representa un aminoácido, aunque también hay modelos de redes en los que cada nodo representa un átomo [10] (fig. 1.4c). Las longitudes de equilibrio de los enlaces/muelles corresponden a las distancias interatómicas (o interresiduo) observadas en la estructura experimental y las fuerzas suelen depender de la distancia de forma cuadrática. Hay otras aproximaciones más sofisticadas en las que las constantes de fuerza dependen de la distancia según una distribución Gaussiana [12] o se calculan en base a simulaciones de dinámica molecular [13]. En el extremo

opuesto, tenemos los modelos ultra-simplificados [11], en los que la fuerza es una función escalón igual a cero si la distancia excede 7 Å, y a uno, en caso contrario.

Generalmente, los modelos de redes elásticas generan correctamente la topología del sistema y son capaces de reproducir los patrones de los modos de movimiento principales de las proteínas. Entre sus aplicaciones fundamentales, se encuentran el estudio de movimientos globales de baja frecuencia, incluyendo fluctuaciones térmicas [11], movimientos de dominios [12], cambios conformacionales sobre estructuras [14] e interacciones entre proteínas. [15].

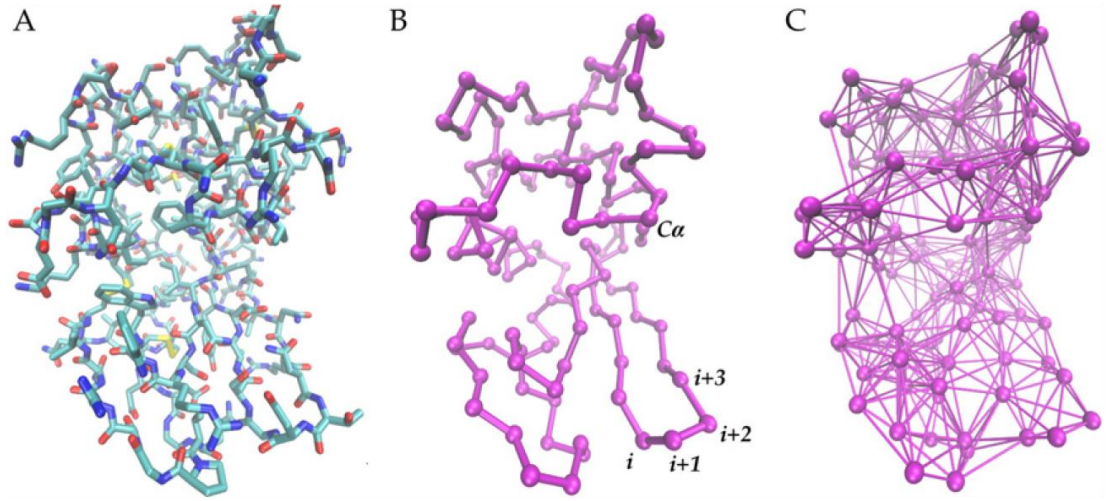


Fig. 1.4 Representación de modelos de la lisozima: **A**- Modelo de átomos pesados (C-cian, 0-rojo, N-azul, S-amarillo) **B**- Cadena de átomos C_α **C**- Átomos C_α (nodos) conectados por líneas (pseudo enlaces) representando de forma esquemática la interacciones entre nodos dentro de un rango menor de 8 Å.

2.1.2. – Modelos basados en la estructura

Los modelos basados en la estructura (también llamados modelos Gō, por ser Nobuhiro Gō el primero en proponerlos [96]) se basan en la asunción de que la topología de la estructura nativa de la proteína determina la mayor parte de las características de su cinética y sus caminos de plegamiento. Esto es debido a que el proceso de plegamiento ha evolucionado de forma general a satisfacer el principio de mínima frustración [16]. De esta forma la superficie de energía de la proteína puede ser descrita como un embudo débilmente rugoso, que apunta hacia la es-

tructura nativa (fig. 1.5). Esto implica que los contactos de la estructura nativa i.e. residuos que interactúan cuando la proteína está totalmente plegada juegan un importante papel en el proceso de plegamiento. Dicho de otro modo, las contribuciones energéticas de las interacciones de la estructura nativa actúan como las fuerzas motrices principales en el proceso de plegado.

En las primeras versiones de estos modelos [68], la proteína se representaba como una cadena de nodos cada uno de los cuales correspondía a un aminoácido. En estos planteamientos iniciales, las interacciones entre nodos reflejaban la estructura nativa final que se iba a alcanzar. Se observó, sin embargo, que dicho sesgo no permitía la reproducción de estados meta estables intermedios en el plegamiento. Para corregir este problema, se han desarrollado nuevos modelos, más sofisticados, en los que se añaden más términos de energía con el consiguiente aumento en la frustración de la energía, apareciendo así nuevos estados intermedios que aumentan la complejidad de la superficie de energía por la que se desplaza la proteína. Por ejemplo, la inclusión de una barrera de solvatación–desolvatación a las interacciones no-covalentes hace que aparezcan estados intermedios parcialmente desolvatados [17,18].

Los modelos basados en la estructura han sido útiles para investigar escenarios complejos de plegamiento [19] y movimientos funcionales de proteínas [20,21]. Así, hemos podido entender estos mecanismos como la resultante de unos pocos factores generales como son la simetría, frustración, competición entre plegamiento y ensamblaje, etc. Por ejemplo, estos modelos se han utilizado para desentrañar la función de las grandes máquinas moleculares [22.23], mostrando que efectos estéricos, frustración localizada, plegamiento parcial son los que cuentan para el funcionamiento de los grandes complejos de proteínas motoras.

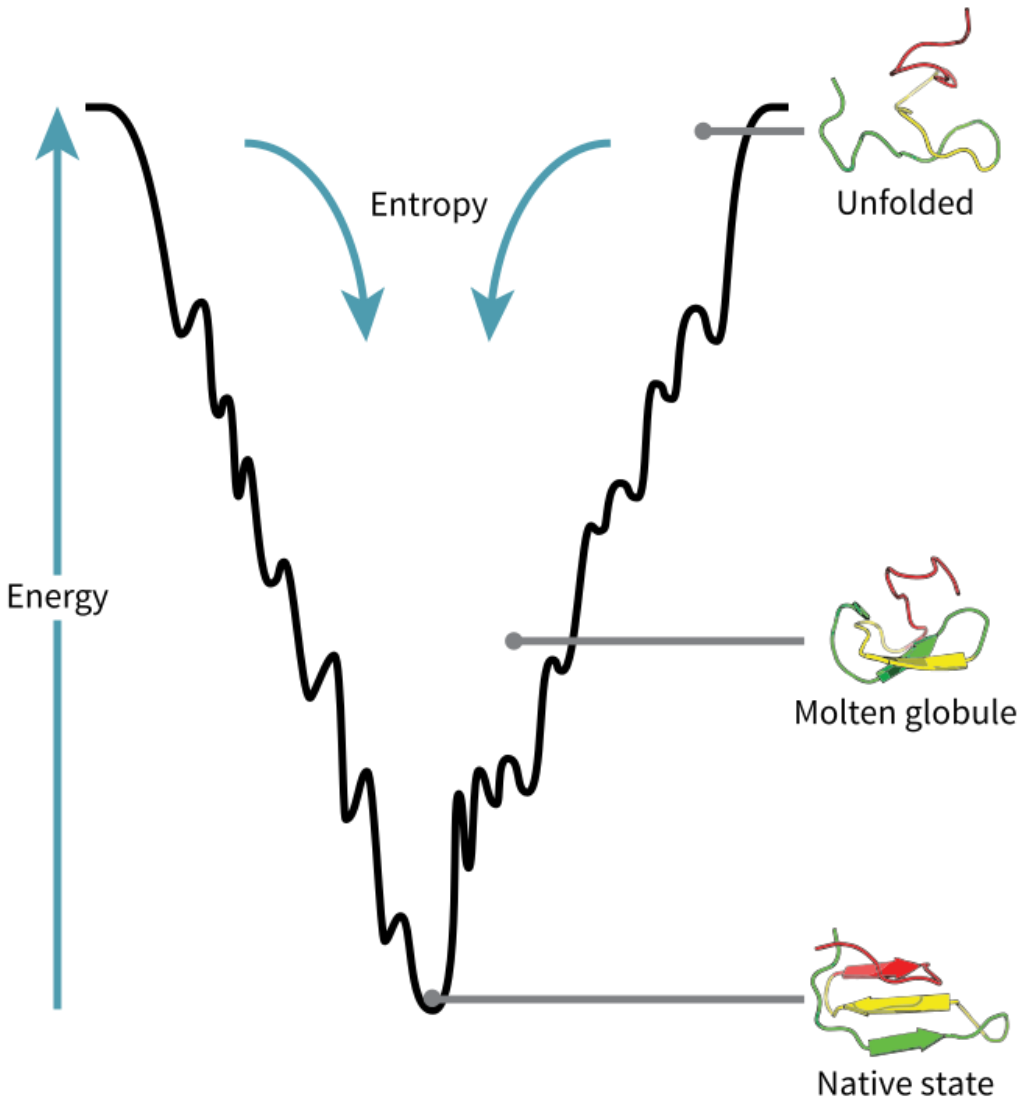


Fig. 1.5 Superficie de energía libre en el plegamiento de proteínas. El plegamiento ocurre a través del ensamblaje progresivo de elementos estructurales, siguiendo una superficie de energía con forma de embudo.

2.1.3.- Modelos basados en mecánica molecular

Como hemos visto anteriormente, los modelos de redes flexibles y los basados en estructuras utilizan una conformación de referencia que determina el campo de fuerzas, dando lugar a que los estudios predictivos resultantes tengan un valor únicamente académico, y restringido a la proteína considerada. La falta de descripciones de interacciones físico-químicas como la hidrofóbica, o los enlaces de hidrógeno que se pueden formar en el curso de la dinámica de la proteína está en los orígenes de esta falta de generalidad. Para superar esta limitación, se han

desarrollado otros modelos, basados en una descripción fina de la forma en que los átomos interaccionan entre sí, y excluyendo el sesgo hacia interacciones determinadas. Estos modelos, basados en una descripción de los mecanismos moleculares de interacción [69], fueron inicialmente desarrollados para el estudio del plegamiento de las proteínas, y son más fiables para simular su dinámica molecular [70].

Los primeros intentos para construir estos modelos datan de la década de los 70 con el trabajo de Levitt [24], que presentó una aproximación a la obtención de los parámetros energéticos de la proteína que inspiró muchos trabajos posteriores. La fiabilidad de estas representaciones reducidas depende de un fino equilibrio entre los diferentes términos de los campos de fuerza [97]. Estos campos de fuerza suelen combinar contribuciones de largo alcance (no-covalentes), que incluyen interacciones Van-der-Waals y electrostáticas, y de corto alcance (covalentes), que determinan la geometría y flexibilidad de la cadena polipeptídica [25.27]. Presentamos aquí tres tipos de modelos que se distinguen por el nivel de resolución de la descripción de la proteína.

2.1.3.1. Modelos de alta resolución

En este tipo de modelos se trabaja con una representación detallada de la cadena principal de la proteína y otra más simple para la cadena lateral. Los modelos más complejos utilizan cuatro unidades atómicas para la cadena principal, que representan: el nitrógeno y su hidrógeno, el C^{α}, el C carbonilo y el oxígeno [25,28]. En otras representaciones se agrupan el carbón carbonilo y su oxígeno en un sólo pseudo-átomo, reduciendo a tres el número de unidades [29,30]. En muchos casos, los átomos de la cadena lateral se agrupan en un pseudo-átomo único por aminoácido, aunque a veces se llegan a utilizar hasta 4 unidades por aminoácido [31].

Con estas descripciones, los parámetros que determinan la geometría de la cadena principal, y por lo tanto su ordenamiento en el espacio, deben definirse con cuidado. Por ejemplo, en las representaciones completas de la cadena principal se utilizan los ángulos torsionales convencionales:

$$\Phi_k \text{ -> } C_{k-1} - N_k - C_k^{\alpha} - C_k$$

$$\Psi_k \text{ -> } N_k - C_k^{\alpha} - C_k - N_{k+1}$$

Una vez definida la representación geométrica, se le asocia una parametrización del campo de fuerza deseado, i.e. una fórmula funcional con unos parámetros establecidos que nos permitan el cálculo de la energía potencial entre los átomos/pseudo-átomos definidos.

Las funciones de energía asociadas a los valores de los torsionales se incluyen en el campo de fuerza global de la simulación. Los parámetros correspondientes (esencialmente los valores de equilibrio de los ángulos y la energía que requiere su deformación) pueden ser extraídos de una colección de estructuras conocidas o mediante simulaciones de dinámica molecular. De hecho, incluso pueden ser recalibrados de forma que reproduzcan el mapa de energías de Ramachandran.

Una ventaja de estos modelos es que los dos pares de átomos agrupados de la cadena principal, NH y CO, permiten introducir de forma natural los enlaces de hidrógeno que estabilizan la estructura secundaria de la proteína, mediante el uso de interacciones atractivas [30] (fig. 1.6)

Normalmente, estas representaciones se han desarrollado para simular el proceso de plegamiento de las proteínas, usando dinámica molecular o dinámica Langevin [29.30], aunque en teoría también se pueden utilizar para el estudio de la dinámica conformacional de las proteínas alrededor de la estructura nativa.

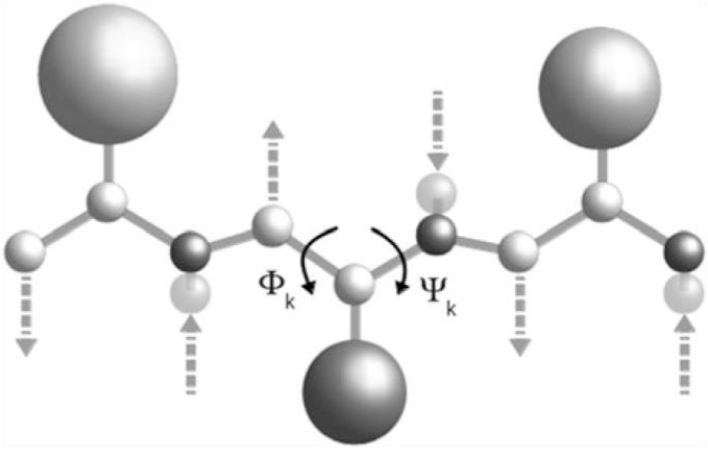


Fig. 1.6 Modelo de alta resolución. La cadena lateral de los aminoácidos se representa con un solo pseudo-átomo; las flechas indican los enlaces de hidrógeno que se pueden formar entre las partículas de la cadena principal [78]

2.1.3.2. Modelos de resolución intermedia

Una característica de estos modelos, es que simplifican la representación de la cadena principal utilizando un sólo pseudo-átomo (generalmente localizado en el C^{α}) por cada residuo. Una consecuencia importante de esta descripción, es que ya no pueden definirse con exactitud los ángulos torsionales que determinan la geometría de la proteína; deben redefinirse, utilizando en su lugar pseudo-ángulos basados en los C^{α} (fig. 1.7 y 1.8):

$$\Theta_k \text{ -> } C_{k-1}{}^{\alpha} - C_k{}^{\alpha} - C_{k+1}{}^{\alpha}$$

$$\tau_k \text{ -> } C_{k-1}{}^{\alpha} - C_k{}^{\alpha} - C_{k+1}{}^{\alpha} - C_{k+2}{}^{\alpha}$$

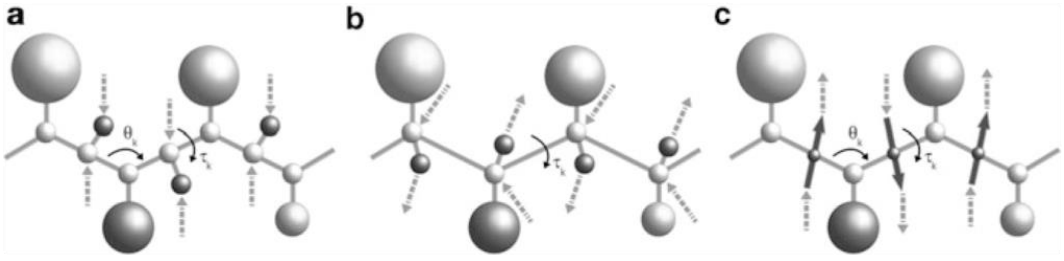


Fig. 1.7 Ejemplos de representación de la estructura en tres modelos a resolución intermedia: a) Levitt [24], b) Ha-Duong [34], c) Liwo [33] (tomado de [78])

Una vez dado este paso, la función de energía para estos pseudo-torsionales deja de tener un origen físico; en su lugar, viene derivada por un análisis estadístico de estructuras de proteínas conocidas [32] (fig. 1.8). Estas funciones de energía son obviamente insuficientes por si solas para estabilizar la cadena principal de la proteína, pero permiten definir los elementos de estructura secundaria propios de las proteínas tanto en su estado nativo, como durante su dinámica.

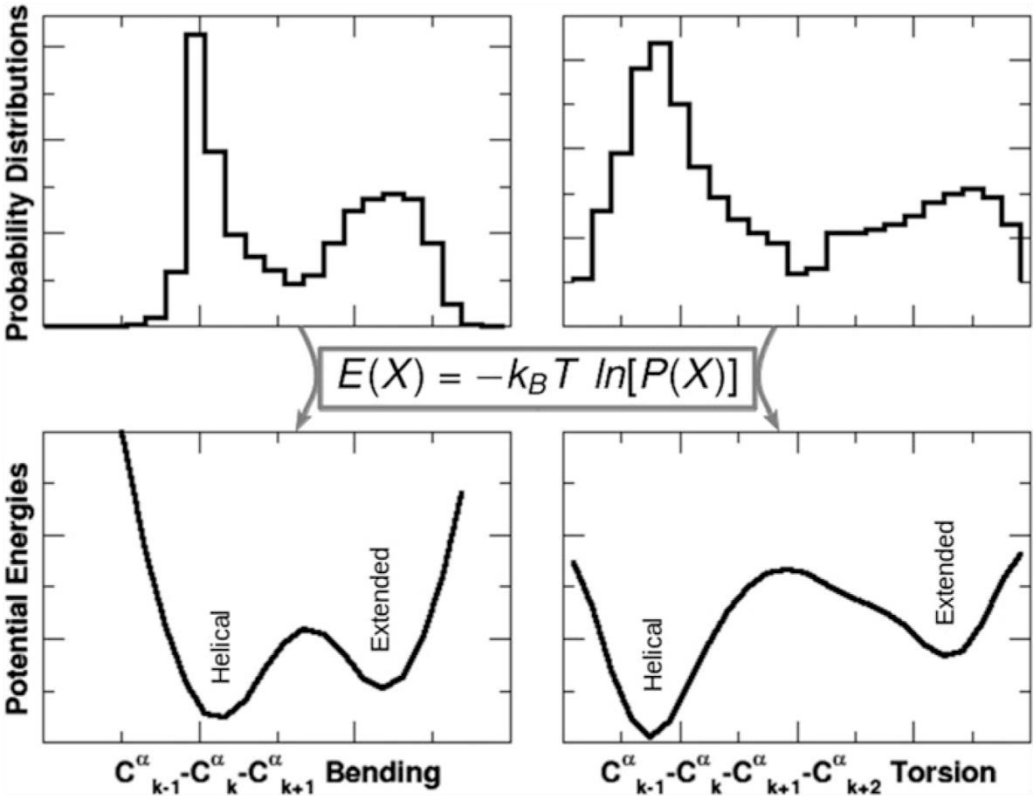


Fig. 1.8 En las figuras de la izquierda vemos la distribución de valores del ángulo θ en las estructuras almacenadas en el PDB (Protein DataBank [71]) y el perfil de energía correspondiente (figura inferior izquierda). A la derecha vemos lo mismo para el ángulo de torsión τ. Las leyendas 'helical' y 'extended' indican dos conformaciones secundarias preferenciales (hélices y hojas beta). El paso de frecuencia (gráficas superiores) a energía (gráficas inferiores) se realiza aplicando la transformación matemática que aparece en mitad de la figura (tomado de [78])

Algunos de estos modelos introducen uno o dos átomos virtuales en la representación de la cadena principal para simular puentes de hidrógeno que estabilicen las estructuras secundarias como hélices alfa y hojas beta. Esto se realiza mediante interacciones estilo dipolo-dipolo. Lo podemos ver, por ejemplo, en el tra-

bajo de Levitt [24] donde introduce dos átomos para tal fin, localizados uno de ellos (N'_k) entre los dos carbonos alfa $C_{k-1}^{\alpha} - C_k^{\alpha}$ y el otro (O'_i) desplazado 1 Å de N'_k y colocado perpendicularmente al plano que forman los carbonos alfa $C_{k-1}^{\alpha} - C_k^{\alpha} - C_{k+1}^{\alpha}$. A estos átomos se les asigna una carga parcial y se les hace interaccionar a través de la ley de Coulomb (fig. 7a).

La presencia de estas representaciones de los puentes de hidrógeno permite simular varios aspectos de la dinámica de las proteínas, como los procesos de plegamiento, las transiciones entre diferentes conformaciones y la diversidad y amplitud de las fluctuaciones conformacionales alrededor de la estructura nativa

2.1.3.3. Modelos 'One-Bead'

En esta categoría de modelos de baja resolución se incluyen aquellos que representan cada residuo mediante un único pseudo-átomo, o bien un pseudo-átomo para la cadena principal y otro para la cadena lateral. La conformación más usual es la que utiliza los dos pseudo-ángulos:

$$\Theta_k \rightarrow C_{k-1}^{\alpha} - C_k^{\alpha} - C_{k+1}^{\alpha}$$

$$\tau_k \rightarrow C_{k-1}^{\alpha} - C_k^{\alpha} - C_{k+1}^{\alpha} - C_{k+2}^{\alpha}$$

extraídos de análisis estadístico de estructuras de proteínas conocidas (fig. 1.8). Sin embargo, al carecer de interacciones de puentes de hidrógeno, estos modelos no son capaces de producir trayectorias realistas en las simulaciones de dinámica molecular. Por este motivo, se introducen variaciones en las funciones de energía que ayuden a mantener los elementos de estructura secundaria en su estado inicial. Por ejemplo, en el modelo utilizado por Tozzini et al [35] (fig. 1.9) los potenciales asociados al pseudo-ángulo Θ_k tienen dos mínimos, correspondientes a las dos conformaciones preferenciales observadas: las hélices alfa y las hojas beta. Sin embargo, los cambios del pseudo-ángulo torsional τ_k se describen mediante el uso de funciones armónicas. De esta manera los autores fueron capaces de generar largas trayectorias estables de la estructura nativa de la HIV-1 proteasa [35].

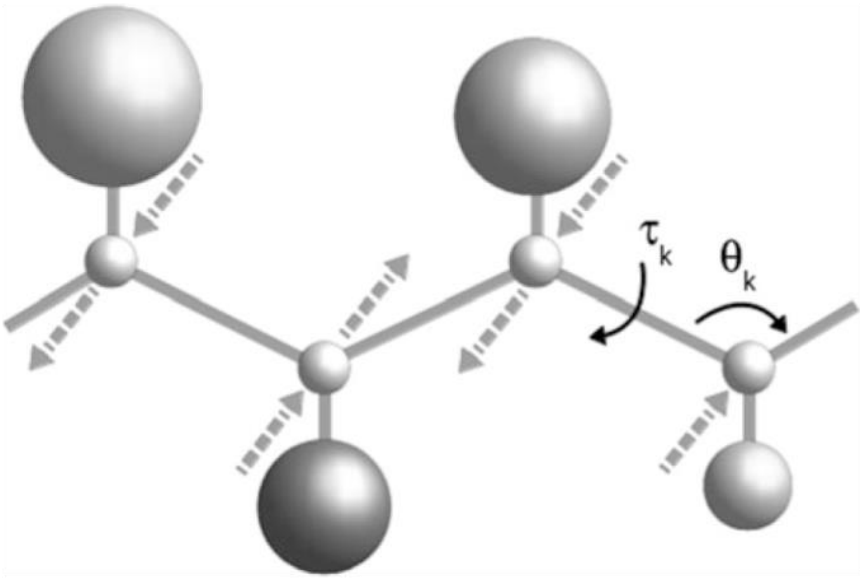


Fig. 1.9 Ejemplo de modelo 'One bead'. Las flechas indican las variaciones de los potenciales efectivos introducidos para mantener la estructura secundaria [78]

3. APLICACIONES DE LOS MODELOS SIMPLIFICADOS DE PROTEÍNAS

A lo largo de la sección anterior hemos mencionado las aplicaciones, a nivel molecular, más frecuentes de los diferentes modelos de la estructura de las proteínas. Concretamente, hemos visto que los modelos simplificados pueden utilizarse en un amplio rango de problemas, desde la predicción del plegamiento de las proteínas hasta el estudio de su dinámica y termodinámica en numerosos sistemas biológicos. En esta sección describimos en mayor detalle algunas de dichas aplicaciones, para proporcionar una idea más precisa del valor de estos modelos.

3.1 Plegamiento de proteínas

El principio fundamental en el que se basa la predicción de la estructura de proteínas es la hipótesis termodinámica de Anfinsen [36] según la cual, la estructura nativa de una proteína corresponde al mínimo global de energía libre del sistema proteína + solvente. Ello da lugar a que la predicción de la estructura de la proteína puede ser formulada como un problema de optimización global [37]. Desde un pun-

to de vista técnico, debido al gran número de grados de libertad involucrados, el problema es todavía intratable. Ello hace que a nivel de resolución difícilmente puedan incluirse todos los átomos del sistema, excepto para pequeñas proteínas, y aún y así, se requieren importantes recursos computacionales. Debido a esta importante limitación, los modelos simplificados han sido usados de forma frecuente para la predicción estructural. Por ejemplo, el modelo UNRES (UNited RESidue) desarrollado por Liwo et al. [38] y que modela la cadena principal a partir de dos pseudo-átomos, uno para el grupo peptídico localizado entre dos átomos C_α y otro para el C_α. La cadena lateral se modela con pseudo-átomos elipsoidales. Tanto el grupo peptídico como la cadena lateral actúan como puntos de interacción, mientras que el C_α solo sirve para definir la geometría de la cadena principal en el modelo y no interacciona (fig. 1.10).

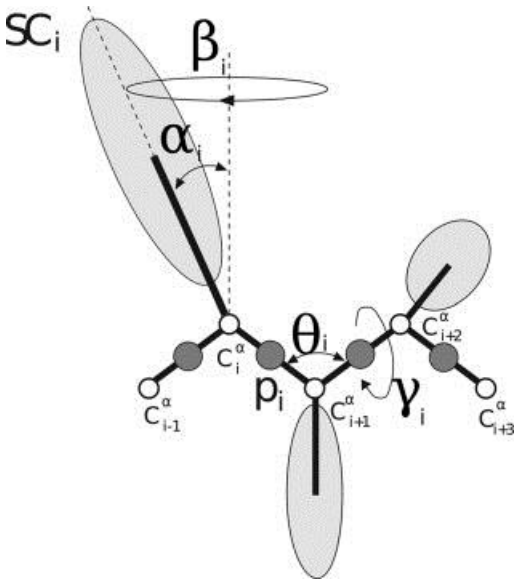


Fig. 1.10 El modelo UNRES de la cadena polipeptídica. Los grupos peptídicos están representados como círculos grises y la cadena lateral está representada como elipsoides grises de diferentes tamaños, dependientes del tipo de residuo. Los C_α son los círculos blancos. La geometría de la cadena se puede describir por los vectores de enlace virtuales dC_i y SC_i (representados por líneas de puntos) o en términos de longitudes de enlace virtuales, ángulos virtuales de la cadena principal (θ_i, γ_i) y ángulos que describen la localización de la cadena lateral respecto al marco de coordenadas definido (α ,β_i) [79].

El potencial de interacción entre pseudo-átomos utilizado en UNRES es una función de energía en la que todos los grados de libertad (incluida la interacción

con el solvente) participan y están promediados, dando lugar a potenciales efectivos. Las interacciones covalentes incluyen enlaces, ángulos y diedros para la cadena principal y un potencial rotacional que define el rotámero de la cadena lateral. Las interacciones no-covalentes (van der Waals y Coulomb) incluyen términos para los pares de pseudo-átomos cadena lateral/grupo peptídico y cadena lateral/cadena lateral. Los términos no-covalentes se derivan a partir de modelos físicos, o de cálculos semi-empíricos realizados en pequeños sistemas modelo, o de potenciales de fuerza media extraídos de simulaciones de dinámica molecular de pares de cadenas laterales análogos. La función de potencial final de UNRES también incorpora términos dependientes de la temperatura.

El modelo UNRES ha ido desarrollándose durante veinte años y ha sido utilizado durante este tiempo para estudios de plegamiento de proteínas [39], predicción de estructuras [40], ensamblajes entre proteínas [41] y mecanismos de fibrilación de proteínas [42]. En la fig. 1.11a vemos dos predicciones realizadas con UNRES: la proteína diana T0215 con tres hélices predicha con un crmsd C_α de 3.5 Å respecto la estructura nativa, y la T0281, una proteína α/β predicha con un crmsd C_α de 5,5 Å respecto la estructura nativa [43]. UNRES también ha descrito la vía de plegamiento de varias proteínas, constituidas por una o varias cadenas. Por ejemplo, en la fig. 11b vemos el proceso de plegamiento del dominio de 48 residuos Lysm [44[. Dicho proceso se inicia partiendo de una conformación inicial arbitrariamente establecida en forma de hélice α; se observa un desplegamiento posterior y a continuación un replegamiento de las regiones N-terminal y C-terminal hasta que alcanzan la estructura nativa en forma de hoja β antiparalela. Otro ejemplo es el plegamiento *ab initio* de la proteína 1C6U en la fig. 11c donde las dos cadenas pliegan independientemente a su estructura nativa y posteriormente se ensamblan en una estructura predicha, que está a 2.4 Å de crmsd C_α respecto a su estructura nativa [45].

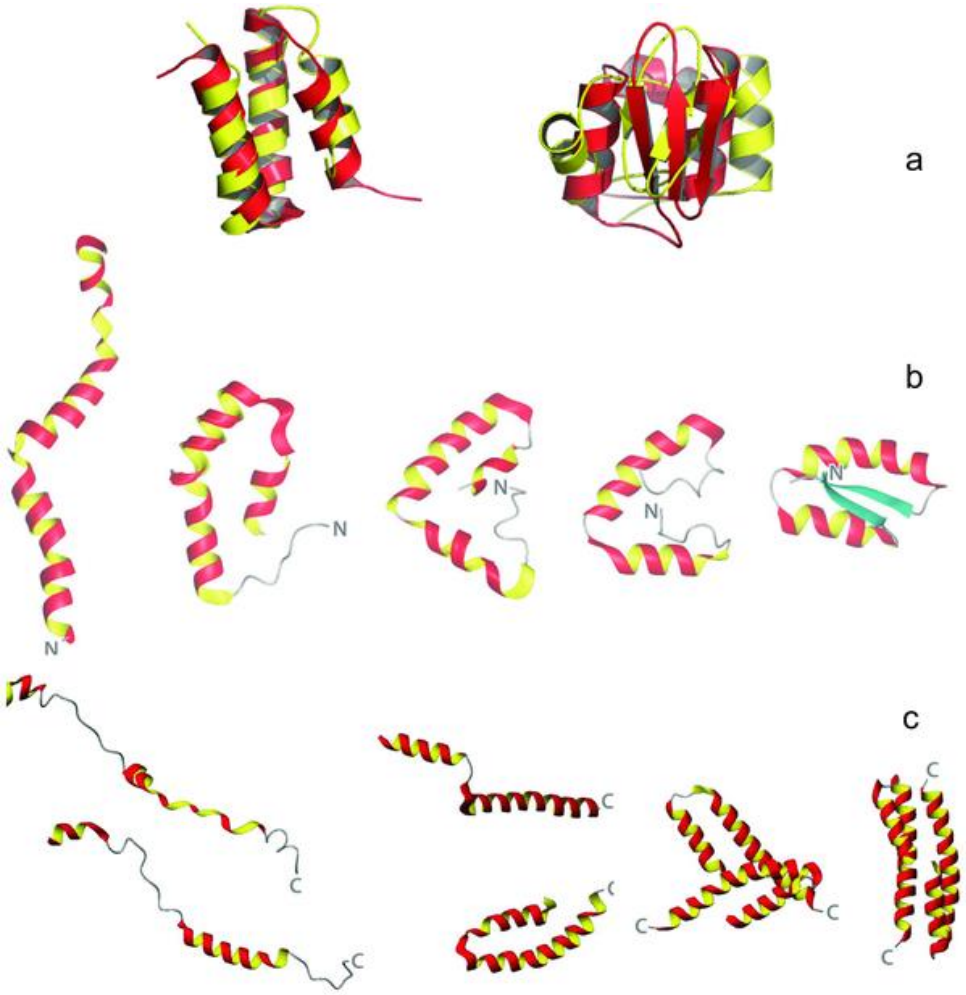


Fig. 1.11 Ejemplos de predicciones de plegamiento de proteínas con el modelo UNRES a) Proteínas T0215 (izquierda) y T0281 (derecha). La estructura nativa se muestra en rojo y la estructura predicha de color amarillo. b) Instantáneas predichas del proceso de plegamiento del dominio de 48 residuos Lysm. c) Proceso de plegamiento de dos cadenas de 48 residuos de la proteína 1C6U [80]

3.2 Modelado de los canales mecánico-sensitivos de apertura / cierre

Los canales mecánico-sensitivos de gran conductancia (Mscl) son un componente de la envoltura de la célula bacteriana que prevé ajustes rápidos de la presión de turgencia en respuesta a reducciones osmóticas. Cuando la tensión de la membrana se acerca al punto de ruptura (límite lítico), el Mscl forma un gran poro no selectivo que libera los osmolitos sobrantes, actuando de esta forma como una válvula [46]. La comprensión y descripción de este mecanismo mediante simulaciones computacionales es particularmente costosa debido al enorme número de átomos involucrados. Por ello, y a pesar de su interés, el proceso de apertura / cie-

rre de Mscl no había podido ser reproducido a escala atómica sin utilizar potenciales fuertemente sesgados. Recientemente este proceso ha podido ser simulado de forma no sesgada utilizando el modelo simplificado Martini [47,48].

El modelo Martini usa un mapeo de 4:1 en su representación atomística, esto quiere decir que en promedio cuatro átomos más sus hidrógenos asociados se representan por un simple centro de interacción o pseudo-átomo. A estos pseudo-átomos se les asigna un tipo específico con un carácter más o menos polar, dependiendo de la naturaleza de su estructura química subyacente. El modelo de Martini tiene cuatro tipos de pseudo-átomos: polar (P), no polar (N), apolar (C) y cargado (Q). Dentro de cada uno de ellos, se distinguen varios subtipos indicando su capacidad para formar puentes de hidrógeno o el grado de polaridad (fig. 1.12). Las interacciones no covalentes como la de van der Vaals y la electrostática se describen mediante potenciales de Lennard-Jones (LJ) [49], basando su parametrización en datos experimentales de origen termodinámico. Por ejemplo, cada par de partículas i y j a una distancia r_{ij} interactúan por según el potencial LJ

$$V_{LJ}(r_{ij}) = 4\, e_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right]$$

La fuerza de cada interacción entre pares viene dada por el valor del potencial LJ y depende de los tipos de partículas. Los valores del parámetro e_{ij} varían según la interacción, e.g. e_{ij}=5.6 kJ mol^{-1} para interacciones grupos fuertemente polares a valores de e_{ij}=2.0 kJ mol^{-1} para interacciones entre grupos polares y apolares (imitando en efecto hidrofóbico). El parámetro σ_{ij} representa la distancia más cercana de aproximación entre dos partículas, siendo su valor de 4.7 Å para las partículas normales.

Para las interacciones covalentes, los enlaces y ángulos se describen mediante potenciales armónicos cuyos parámetros dependen de la conformación de estructura secundaria del residuo (fig. 1.13).

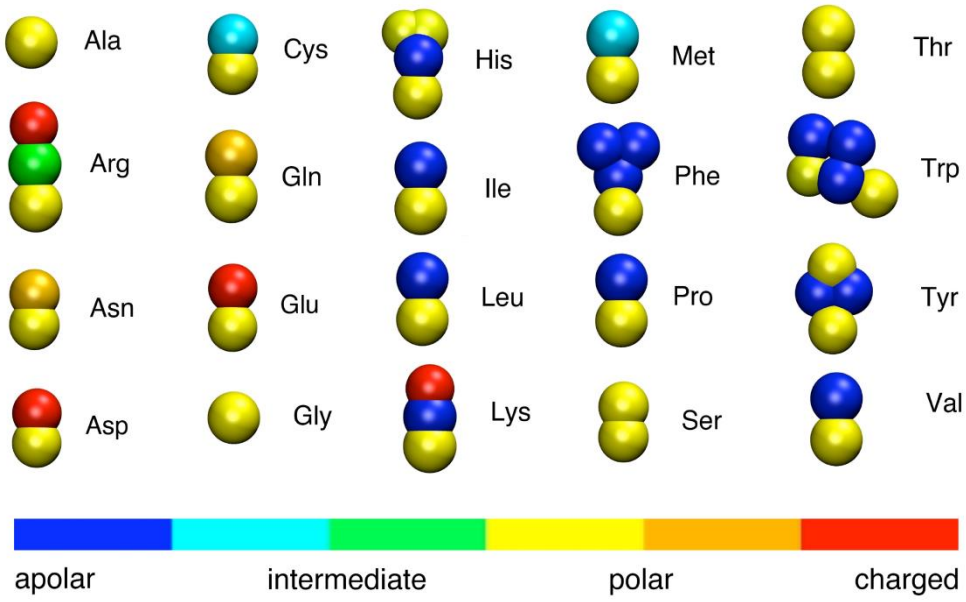


Fig. 1.12 Representación en el modelo de Martini [72] de los residuos de una proteína. Cada residuo está formado por un único pseudo-átomo para la cadena principal, y de 0 a 4 para la cadena lateral, dependiendo del residuo. A cada partícula se le asigna un tipo, representado en la figura por el color, basado en su carácter polar.

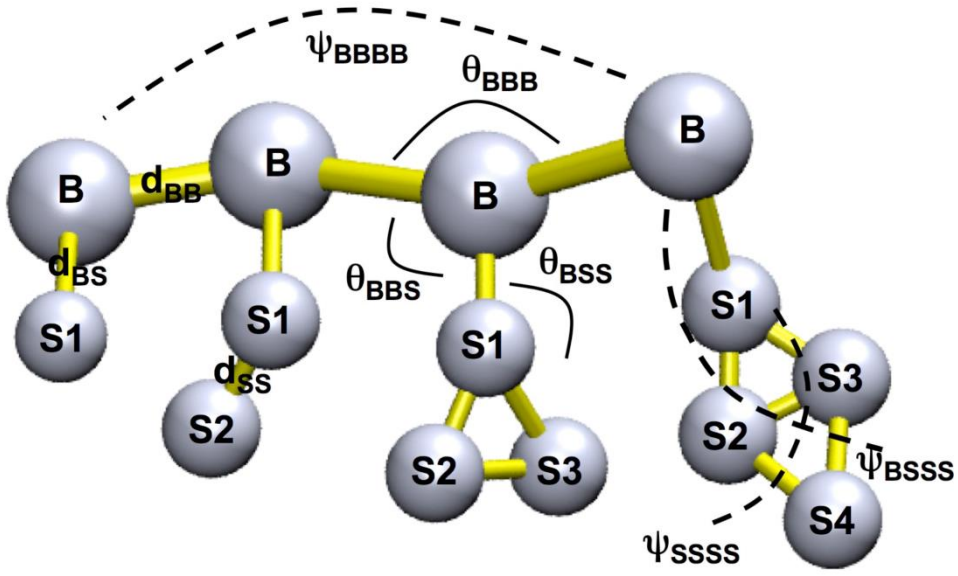


Fig. 1.13 Ejemplo de representación de la cadena principal de una proteína (partículas B) y la cadena lateral (partículas S) en el modelo de Martini [72]. La estructura queda determinada por los enlaces, ángulos de enlace (θ) y ángulos de torsión (ψ). Los ángulos de la cadena principal, θ_{BBb} y ψ_{bbb}, dependen de la estructura secundaria.

Los trabajos realizados sobre Mscl utilizando el modelo Martini confirman que el mecanismo de apertura del canal es parecido al mecanismo del iris, y proporcionan información valiosa sobre cómo los cambios en la forma de las proteínas influyen en su equilibrio conformacional preferido. Para llegar a este resultado, se siguió un protocolo de simulación específico: primero, se equilibró el canal en un entorno de solvente dividido en dos capas durante unos pocos microsegundos; posteriormente, se le aplicó rápidamente una tensión lateral. En los 10 – 100 nanosegundos posteriores, se observó cómo las hélices transmembrana del Mscl se inclinaron, extendiendo de esta forma la cavidad extracelular del canal. La puerta del canal hidrofóbico tarda unos 0.2-2 microsegundos adicionales antes de extenderse y abrir así el canal (fig. 1.14).

Para esta simulación se necesitó una computadora con 12 CPUs y se tardaron unos 5-10 días en completarla, lo que nos da una idea de la utilidad de estos modelos, ya que un modelo que incluyese todos los átomos habría tardado años en completar la simulación, utilizando el mismo sistema computacional.

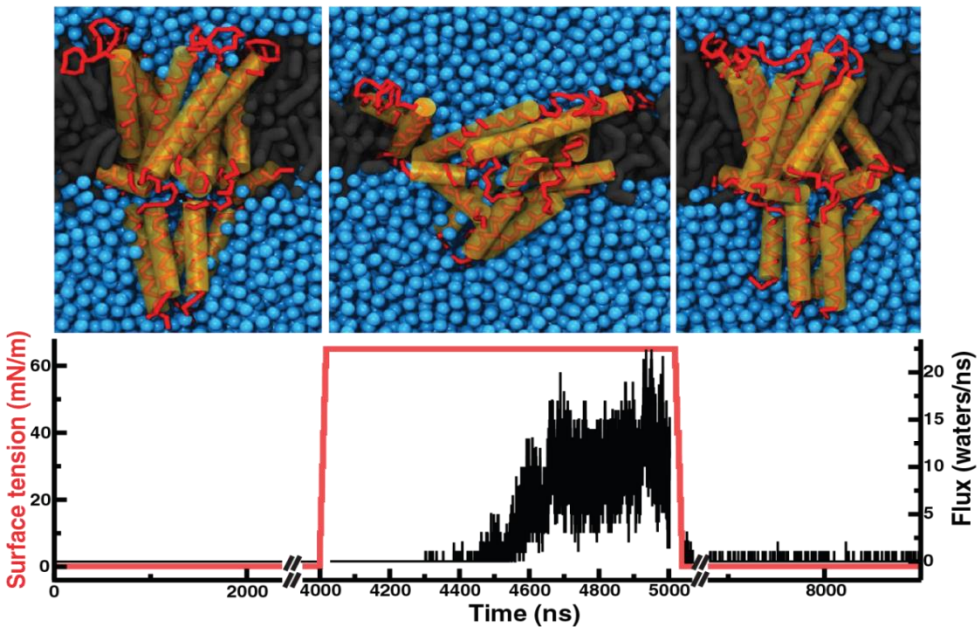


Fig. 1.14 Mecanismo reversible de apertura / cierre de Mscl utilizando el modelo simplificado Martini [47,48]. En la parte superior-izquierda de la figura vemos el Mscl equilibrado en un solvente entre dos capas. En la figura superior-central se ve como el canal se abre de forma permeable al agua, tras aplicar una tensión a las capas. Finalmente, cuando se quita la tensión, el canal recupera su posición original, cerrando el paso al agua. (Figura superior-derecha). En la parte inferior de la figura se muestra la tensión superficial aplicada (rojo) y el flujo de agua (negro) en función del tiempo

3.3 La proteína de membrana auto-ensamblada

Las membranas biológicas tienen una organización supramolecular compleja y dinámica que juega un papel importante en muchos procesos fundamentales [73]. La naturaleza transitoria de estos procesos ha hecho que su estudio mediante métodos convencionales, tanto experimentales como computacionales sea un gran desafío. Las simulaciones con el modelo de Martini, cuya simplicidad se refleja en un menor coste computacional, se han utilizado para intentar esclarecer las fuerzas involucradas a nivel molecular. Un conjunto reciente de estudios [81-85,90-93] han aprovechado esta visión a nivel casi atomístico para revelar algunos papeles significativos de la interrelación entre proteínas y lípidos. Por ejemplo, en la predicción de modos de unión de las proteínas a las membranas, así como la adaptación de la membrana alrededor de las proteínas [81-85].

Podemos ver en la fig. 1.15 un sistema típico utilizado en estos estudios. Consiste en 64 receptores visuales de rodopsina dentro de una membrana bicapa de dioleoyl-fosfatidilcolina (DOPC) en una proporción 1/100 molar de proteína / lípido.

- La membrana lipídica responde a la presencia de la proteína con una deformación anisotrópica que permite una coincidencia entre la superficie hidrofóbica de la proteína y la parte lipídica de la bicapa membranosa [90].
- El grado de deformación de la membrana determina la propensión de la proteína a auto-organizarse [90].
- Las propiedades de la superficie de las proteínas determinan sitios específicos de unión a la membrana lipídica [91], induciendo además la formación de interfaces entre proteínas que a su vez desembocan en la formación de complejos altamente organizados [92] donde las proteínas se ordenan mediante las propiedades de los lípidos en parches de membrana multidominio [93].

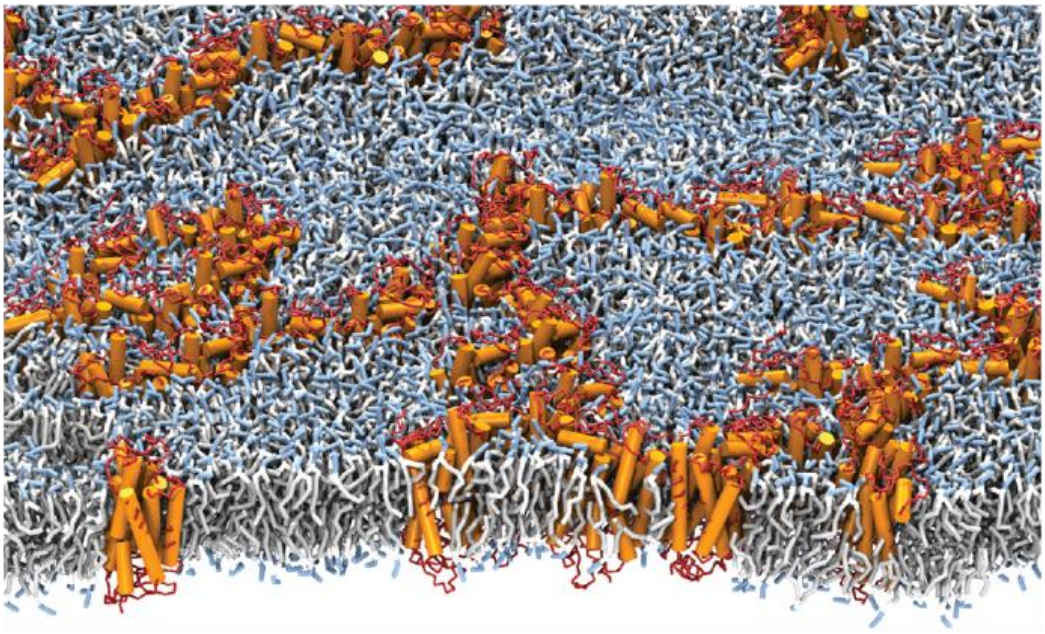


Fig. 1.15 Proteína de membrana auto-ensamblada. Los receptores se pusieron inicialmente en una rejilla 8x8 y se les dejó libres para que se auto-ensamblaran en un periodo de 100 μs. Las hélices del receptor transmembrana se presentan como tubos de color naranja. Los grupos de cabeza de lípidos se muestran en azul claro, los grupos de glicerol en blanco, y las colas en gris

4. APLICACIONES DE LOS MODELOS DE BAJA RESOLUCIÓN REALIZADAS EN ESTE ESTUDIO: DE LA BIOLOGÍA FUNDAMENTAL A LOS ANÁLISIS TÉCNICOS DE LAS SIMULACIONES

En los apartados anteriores hemos descrito diferentes tipos de modelos simplificados, y hemos comprobado cómo pueden aplicarse con provecho al estudio de procesos biológicos a nivel molecular. Todos estos modelos se basan en una reducción de los grados de libertad de las estructuras macromoleculares, para aliviar el coste computacional de las simulaciones, pero manteniendo intactos aquellos aspectos relacionados con el problema biológico estudiado. En esta tesis hemos querido ahondar en este tipo de aplicaciones, y para ello implementamos un modelo reducido original, que por su uso de las coordenadas internas tiene la virtud de permitir una exploración muy rápida de vastas zonas del espacio conformacional, particularmente de sub-estructuras determinadas de la proteína, como son los giros/loops de varios amino ácidos. Esta propiedad nos ha permitido aplicarlo

al estudio del espacio conformacional de las denominadas "low-complexity regions" o LCRs. Por otra parte, nuestro modelo conserva uno de los aspectos fundamentales para la existencia de la estructura de las proteínas: el efecto de volumen excluido, o de la no interpenetrabilidad atómica. Esta característica nos ha permitido estudiar un tema de interés en las simulaciones masivas destinadas a identificar correlaciones entre diferentes sub-estructuras proteicas, con la finalidad de desvelar mecanismos alostéricos o relacionados. A continuación, se proporciona una breve descripción de ambos temas, explicación que será ampliada en el capítulo correspondiente.

4.1 Las "low-complexity regions" o LCR

Las regiones de baja complejidad en las proteínas son fragmentos de la secuencia de las proteínas que muestra una diversidad reducida en su composición aminoacídica [74,75]. Este fenómeno se manifiesta de diferentes maneras; a veces las regiones están compuestas por unos pocos aminoácidos, y otras veces están compuestas por solo uno (fig. 1.16). Las posiciones de los aminoácidos en estas regiones pueden estar dispuestas de diferentes maneras, ya que la baja complejidad es un fenómeno composicional, no secuencial:

- Agrupadas sin un orden definido

PDB	Posiciones Seq: 290-308
2r0y	LAIGGGGPAAAGALAISAL

- Dispuestas de forma periódica

PDB	Posiciones Seq: 123-130
1i7w	DQDQDYDY

- Dispuestas de forma irregular

PDB	Posiciones Seq: 133-142
2it2	SNIKSISNKK

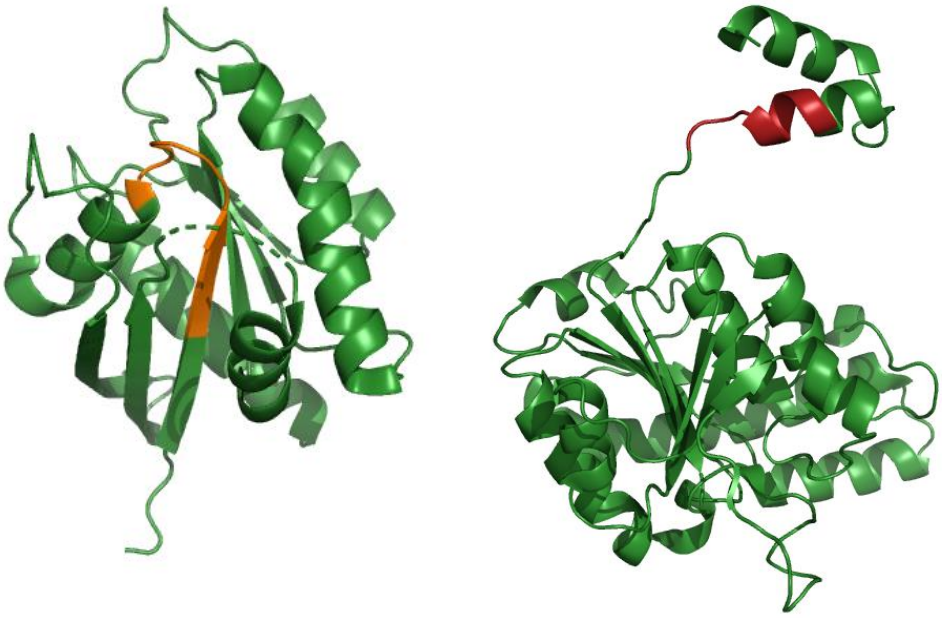


Fig. 1.16 Dos ejemplos de LCR. A la izquierda tenemos la cadena A de la proteína 2ERY y en naranja la región de secuencia repetitiva compuesta por dos tipos de aminoácidos (Valina y Glicina): VVVGGGGVG. En la figura de la derecha, la cadena B de la proteína 2OGX cuyo LCR (en rojo) está formada por un solo tipo de aminoácido (Alaninas): AAAAAAA.

Las LCRs son comunes en las secuencias de las proteínas aunque es difícil saber su abundancia real debido a que no hay una sola forma de definirlas, y cada programa crea su propia definición de LCR. A pesar de ello sabemos que son numerosas, pero que debido a su estado generalmente desordenado, no se observan en el experimento de difracción de rayos X.

Se cree que las LCRs juegan un importante papel en un amplio rango de funciones biológicas [50], mediante mecanismos que han sido ampliamente documentados, aunque los modelos funcionales propuestos permanecen sin verificar [51]. Como ejemplos de posibles funciones tenemos la mediación entre proteínas [86], la unión entre dominios de proteína [87] y la adaptación genética [88]. Aunque sus mutaciones han sido asociadas a diferentes enfermedades, todavía descono-

cemos los mecanismos moleculares de estos procesos. Nuestro objetivo será analizar qué ocurre a nivel estructural cuando las LCR mutan.

4.2 El uso de la correlación en el estudio de las trayectorias de dinámica molecular

La dinámica molecular es un método que nos ayuda a simular el comportamiento temporal de las proteínas (y otras macromoléculas) a lo largo del tiempo, basándose en la mecánica de Newton. Su descripción de las proteínas es detallada, a nivel atómico, y por ello nos permite obtener una visión muy fina de los sistemas moleculares. Aunque proporciona abundante información sobre los movimientos atómicos, esta información debe de sintetizarse en un número reducido de parámetros que permitan su interpretación. Uno de los parámetros más utilizados, es la correlación interatómica, también conocida como correlación dinámica condicional (DCC),

$$DCC(i,j) = \frac{\langle \Delta \mathbf{r}_i(t) \cdot \Delta \mathbf{r}_j(t) \rangle_t}{\sqrt{\langle \|\Delta \mathbf{r}_i(t)\|^2 \rangle_t} \sqrt{\langle \|\Delta \mathbf{r}_j(t)\|^2 \rangle_t}},$$

que es un método popular para estudiar las simulaciones de dinámica molecular. Dicho método permite realizar análisis sobre el grado de coordinación que hay en los movimientos entre las diferentes partes de los sistemas moleculares (fig. 1.17). Esta información en forma de correlación puede utilizarse para comprender ciertos procesos biológicos que ocurren a nivel molecular como son la señal alostérica [76] o el transporte mecánico entre regiones de las proteínas [77]. Sin embargo, ciertos factores triviales, y no relacionados con el problema de estudio, pueden dar lugar a la existencia de correlaciones espurias sin valor interpretativo. Nuestro objetivo será identificar el rol que juega el volumen excluido como fuente de ruido en el cálculo de las correlaciones interatómicas.

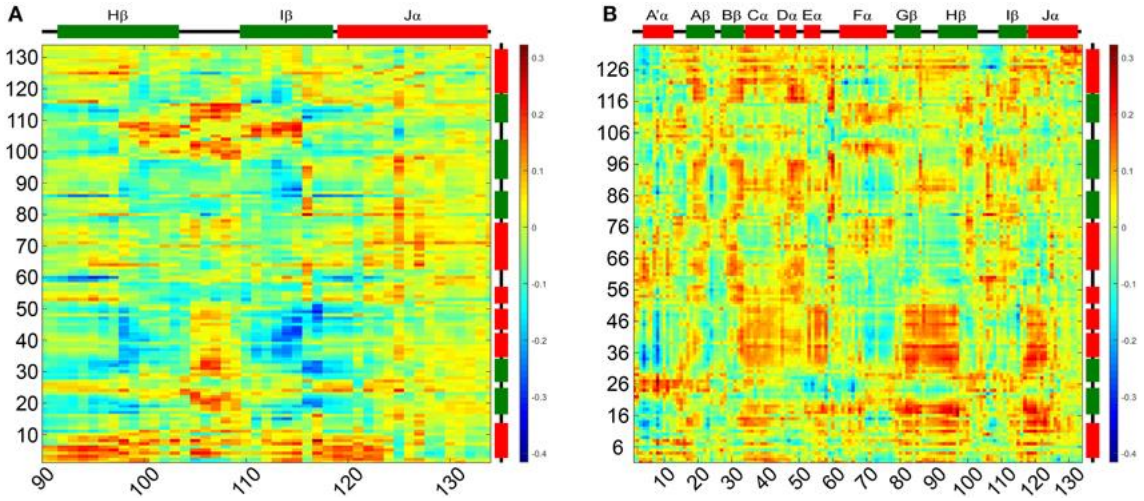


Fig. 1.17 Dos ejemplos de matrices DCC (DCCM) representando las correlaciones entre pares de residuos e ilustradas según códigos de colores. Colores rojos corresponden a correlaciones positivas y colores azules a correlaciones negativas. Estudio realizado para la estructura del dímero PpSB1-LOV (PDB-ID: 3SW1) sensible a la luz y corresponde a la transición de estados 'oscuridad' – 'luz'. En la figura de la izquierda vemos la DCCM entre los residuos de la secuencia 90-134 y la cadena A. En la figura B vemos la DCCM entre las cadena A y B del dímero [89].

5. OBJETIVOS

Los objetivos de la presente tesis son los siguientes:

- Diseñar y desarrollar un programa de simulación estructural basado en la representación de Ca de la proteína.

- Estudiar la variabilidad de secuencia y estructural de las LCR utilizando la capacidad del programa para la simulación local de cambios estructurales.

- Estudiar el impacto del volumen excluído sobre las correlaciones entre residuos, utilizando la capacidad del programa para la simulación global de la estructura.

CAPITULO 2: DESCRIPCION DEL PROGRAMA DE SIMULACIÓN

1.- INTRODUCCIÓN

Bajo circunstancias apropiadas una proteína se pliega espontáneamente desde un estado desnaturalizado a una estructura tridimensional definida. Este proceso tiene un elevado valor biológico, ya que en la estructura final radica la funcionalidad de la proteína. Por lo tanto, los estudios experimentales de dicha estructura y de su plegamiento son valiosos para comprender el rol biológico de la proteína. Desgraciadamente, no siempre es fácil hacer dicho estudio; en tales circunstancias debemos recurrir a las simulaciones, que nos proporcionan una primera solución para el problema estructural. Normalmente, se utilizan cálculos energéticos detallados, que nos permiten aumentar nuestro entendimiento acerca de este proceso [104], de la estabilidad intrínseca de la proteína en su estado nativo y de cómo todo ello se relaciona con su flexibilidad global y local [101]. Sin embargo, una mirada precisa a nivel atómico sobre los cambios conformacionales a gran escala de las proteínas todavía nos resulta computacionalmente inaccesible. Las razones de esta dificultad residen en: (i) el enorme número de conformaciones viables para cada cadena polipeptídica, (ii) la cantidad de mínimos locales de energía que dificultan la identificación del mínimo absoluto y (iii) los pequeños pasos requeridos para atravesar energías de superficies detalladas (1-2 fsec) en comparación con los tiempos utilizados en la construcción de las cadenas (> 1msec).

Por todas estas razones, se ha hecho necesario el desarrollo de modelos simplificados para el entendimiento y predicción de la estructura de proteínas. Un claro ejemplo es el trabajo de Ramachandran [98] donde se utilizaron modelos de esferas rígidas para enumerar las conformaciones admitidas/preferentes de un

dipéptido. Este estudio, que ahora nos parece sencillo, reveló la estabilidad local intrínseca de las estructuras secundarias fundamentales, la 'hélice alfa' y la 'lamina beta'. El estudio de Ramachandran muestra como unos modelos simplificados bien escogidos en relación al problema que queremos abordar permiten sacar lecciones valiosas, sin recurrir a grandes esfuerzos computacionales. Como hemos adelantado en la Introducción, la simplificación de la estructura de proteínas se puede abordar mediante modelos de redes o mediante la discretización de ángulos diédricos. A continuación los describo con mayor detalle.

Una forma sencilla de reducir el espacio conformacional consiste en su discretización. En esta aproximación sólo permitimos un pequeño número de estados por cada residuo, de tal manera que el espacio conformacional pasa de ser continuo a ser numerable. Una estrategia seguida habitualmente consiste en construir un modelo de redes (lattice), donde cada residuo se modela como un simple punto en el espacio. Cada punto puede ser de un tipo determinado, por ejemplo, se pueden diferenciar por tipos de aminoácidos o según su carácter hidrofóbico o hidrofílico; la posición de estos puntos está restringida a aquellas posiciones que definen una red determinada (usualmente una trama cúbica). Para garantizar la conectividad de la cadena de la proteína, los residuos adyacentes de la cadena deben ocupar vértices adyacentes de la red y se impone la condición de que dos o más residuos no pueden ocupar el mismo vértice (correspondiente al concepto de volumen excluido). Las interacciones energéticas entre residuos se modelan mediante una función que reproduce la interacción energética entre los vecinos. Esta función simula, de forma aproximada, las interacciones entre residuos observadas en proteínas reales, tales como efectos hidrofóbicos y enlaces de hidrógeno. El efecto estérico no se incluye en dicha función ya que se trata mediante una condición de volumen excluido, explícitamente programada en el software de simulación. Uno de los modelos de red más populares son los HP (fig. 2.1) donde solo hay dos tipos de 'puntos': hidrofóbico (H) y polar (P) y se simula la interacción hidrofóbica asignando valores negativos (atracción) a la energía de interacción entre dos 'puntos H.

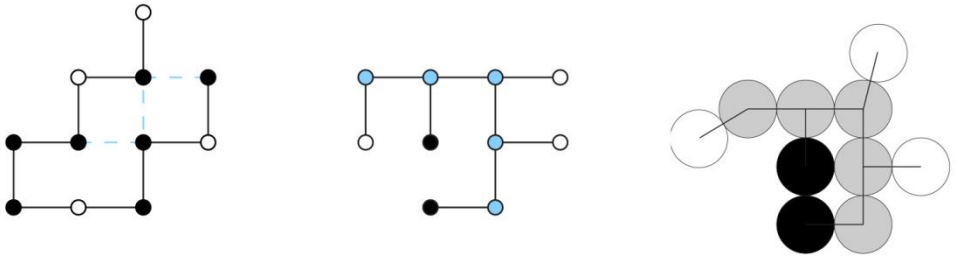


Fig. 2.1: Ilustraciones de conformaciones para distintos modelos de red: a) modelo standard HP; b) modelo HP con cadenas laterales; c) modelo HP de esferas tangentes con cadenas laterales. Los colores negros indican residuos hidrofóbicos, los blancos hidrofílicos y los azules son elementos de la cadena principal [108-110].

Otra estrategia habitual para simplificar la representación de la estructura proteica, consiste en discretizar los ángulos diédricos de la proteína [105,106] principalmente los de la cadena principal, ϕ y ψ. Estos ángulos y sus valores permitidos, así como la correlación que existe entre ellos, fueron perfectamente descritos por Ramachandran y Sasisekharan en su influyente estudio de 1963 [107]. En el proceso habitual de discretización, lo que se hace es describir las diferentes regiones de la distribución (ϕ, ψ) mediante un reducido número de puntos representativos. El número de puntos utilizados determinará la calidad del modelo, tal como muestran Levitt & Park [100] en un trabajo donde comparan la precisión de las estructuras reconstruidas utilizando diferentes modelos de estados discretos (destaca por su calidad el modelo propuesto por los autores ϕ y ψ) (Tabla 2.1).

Model	(τ, α) or (ϕ, ψ) values	Number states	c.r.m.s. (Å)	d.r.m.s. (Å)	% Native Contacts	% Native α	% Native β
Models of increasing complexity							
Tetrahedral/2	α = 109.5, τ = 60,60,180[a]	√3	5.39	4.66	60	NA[b]	NA
Tetrahedral	α = 109.5, τ = −60,60,180	3	3.63	2.99	78	68	80
Half cubic	α = 90, τ = −90,0,90,180; α = 180[a]	√5	4.81	4.18	81	NA	NA
Simple-4 A	α = 120, τ = −90,0,90,180	4	3.07	2.51	75	0	79
Simple-4 B	α = 120, τ = −135,−45,45,135	4	3.02	2.46	72	68	73
(ϕ, ψ) 4-state	(ϕ, ψ) = (−90,−90),(90,−90),(−90,90),(90,90)	4	3.22	2.67	77	29	89
Cubic	α = 90, τ = −90,0,90,180;α = 180	5	2.84	2.34	78	0	59
Simple-5 A	α = 120, τ = −144,−72,0,72,144	5	2.37	1.94	82	0	76
Simple-5 B	α = 120, τ = −108,−36,36,108,180	5	3.01	2.47	77	62	68
Simple-6 A	α = 120, τ = −120,−60,0,60,120,180	6	2.69	2.21	76	56	75
BCC	Body-centered cubic lattice[b]	7	2.59	2.14	87	50	74
FCC	Face-centered cubic lattice[b]	11	1.78	1.46	88	66	82
S + FCC	Simple + face-centered cubic[c]	17	1.60	1.31	88	68	80
18-state	α = ⟨mean⟩, τ = 160,−140, . . . 0,20, . . . 180[d]	18	1.24	1.02	92	90	88
KW	Knight's walk lattice (2,1,0)[e]	23	1.24	1.02	93	67	89
36-state	α⟨mean⟩, τ = −170,−160, . . . 0,10, . . . 180[d]	36	0.97	0.80	94	91	91
XFCC	Extended face-centered cubic lattice[f]	41	1.15	0.94	89	79	83
XKW	Extended knight's walk[e]	55	0.90	0.73	96	79	88
Optimized models							
A	(ϕ, ψ) = (−64,−40),(−123,134),(111,−46),(117,105)	4	2.43	1.99	85	86	80
B	(ϕ, ψ) = (−66,−40),(−119,114),(−36,124),(132,−40)	4	2.31	1.89	86	86	74
C	(ϕ, ψ) = (−63,−63),(−132,115),(−42,−41),(−44,127)	4	2.22	1.82	86	75	75
D	(ϕ, ψ) = (−58,−31),(−127,126),(−97,−24),(109,108)	4	2.28	1.87	85	91	72
E	(ϕ, ψ) = (−71,−57),(−131,122),(−42,−36),(107,−25)	4	2.52	2.08	85	75	72
F	(ϕ, ψ) = (−58,−51),(−133,135),(−33,174),(114,−40)	4	2.42	1.97	84	68	76
G	(ϕ, ψ) = (−56,−48),(−129,128),(−108,35),(−31,−109)	4	2.48	2.03	84	61	76
H	(ϕ, ψ) = (−74,−31),(−131,125),(−101,179),(105,−40)	4	2.37	1.94	85	94	69
6-state	(ϕ, ψ) = (−57,−47),(−139,135),(−119,113), (−49,−26),(−106,48),(−101,−127)	6	1.90	1.55	87	71	80
Rooman *et al.*[g]	(ϕ, ψ) = (−65,−42),(−123,139),(−70,138), (−87,−47),(77,22),(107,−174)	6	1.74	1.42	89	80	81

Tabla 2.1: Comparativa de diferentes modelos de tipo' lattice' y 'off-latice' de estados discretos con diferente complejidad. Según el modelo, se utilizan diferentes sistemas de coordenadas internas, e.g. el par (α,τ) (fig. 3) o el par (ϕ y ψ) (fig.2). La precisión se representa por los parámetros crmsd y drmsd. (Apartado 3.). La tabla se ha construido utilizando un set de 149 proteínas de entre 36 y 753 residuos.

El aspecto global de las proteínas puede ser representado razonablemente bien mediante los modelos anteriores. Sin embargo, en algunos casos, y buscando una mayor velocidad de exploración del espacio conformacional, se han buscado representaciones más simplificadas, aunque inspiradas en la geometría real de las proteínas. Estas representaciones [101] se benefician de la planaridad del enlace peptídico C'=N y utilizan únicamente las posiciones de los carbonos alfa (fig. 2.2)

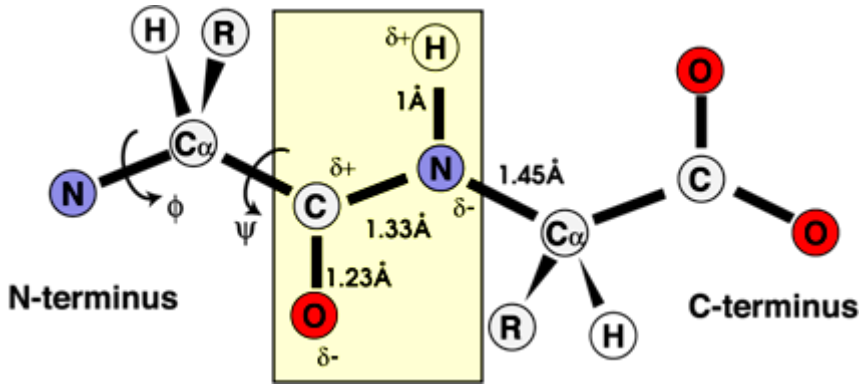


Fig. 2.2: Los péptidos solo pueden cambiar de conformación mediante el giro en torno a los enlaces φ y ψ, adyacentes a los Cα: φ es la rotación en torno al enlace N - C_α y el ángulo ψ es la rotación en torno al enlace C' – C_α. Los enlaces peptídicos (cuadro amarillo) son planares.

Esta aproximación fue introducida inicialmente por Levitt [99] que propuso una representación de la cadena principal basada en el pseudodiedro τ. Dicho ángulo, que corresponde al torsional entre cuatro C_α consecutivos, se obtiene mediante la siguiente fórmula:

$$\tau_i(\phi_i, \psi_i, \phi_{i+1}, \psi_{i+1}) = 180^\circ + \phi_{i+1}, \psi_i + 20^\circ\ [\sin(\phi_i) + \sin(\psi_{i+1})]$$

Esta representación permite reconstruir y estudiar la traza (la estructura tridimensional definida por los átomos C_α) de la proteína. También nos permite una gran reducción en el número de grados de libertad, requisito necesario para las simulaciones estructurales en las que se exploran grandes regiones de espacio conformacional de la proteína. En la presente tesis utilizamos esta representación de la proteína basada en los ángulos pseudodiédricos. Para obtener una descripción completa de la cadena polipeptídica, además del ángulo torsional τ, se necesitan también dos pseudo-ángulos de enlace, θ_1 y θ_2, definidos tal como se describe en la fig. 2.3.

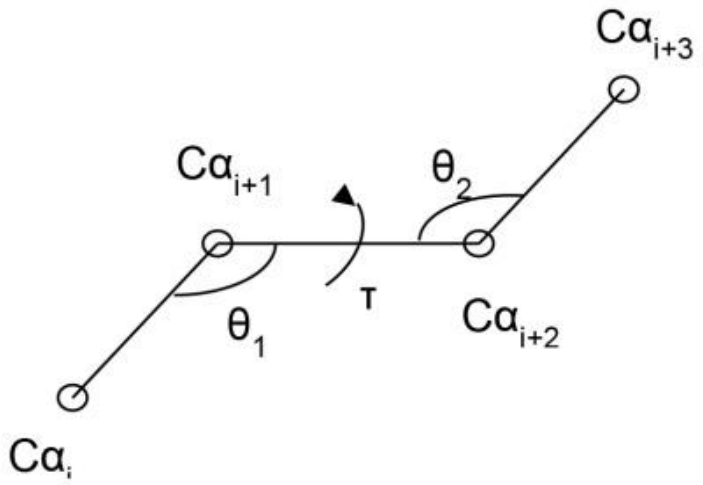


Fig. 2.3 Definición de los ángulos θ_1, θ_2 y τ. Los segmentos de líneas conectan 4 átomos Cα consecutivos.

Un aspecto fundamental que nos llevó a escoger esta descripción es que las estructuras secundarias principales, hélice alfa y lámina beta, se expresan naturalmente mediante el uso de cuatro residuos, o equivalentemente, de la tripleta de ángulos θ_1 - τ - θ_2. Efectivamente, cálculos preliminares (Figura 2 en de la Cruz et al. [101]) muestran que los valores de esta tripleta se distribuyen de acuerdo con la estructura secundaria que representan. Una vez escogida esta representación, y comprobado que efectivamente reproduce los aspectos esenciales de la estructura de las proteínas, consideramos diferentes particiones del espacio (θ_1, τ, θ_2) que corresponden a diferentes niveles de complejidad en la familia de modelos basados en la traza. Para esta parte nos guiamos por los trabajos previos de Park B & Levitt M. [100] y de De la Cruz et al. [101]

En la tabla 2.2 podemos ver los valores que corresponden al modelo utilizado en esta tesis.

Código estado (DS)	θ_1(min,rango)	θ_2(min,rango)	τ(min,rango)	tipo estructura secundaria
A	93(80, 30)	93(80, 30)	51(20, 60)	alpha
g	93(85, 20)	121(100, 30)	99(90, 20)	a-b link 1*/
h	93(85, 20)	121(130, 30)	115(105, 20)	a-b link 2*/
d	93(90, 10)	93(90, 10)	106(80, 50)	3
B	121(100, 40)	121(100, 40)	200(170, 60)	big beta
b	121(100, 40)	121(100, 30)	248(230, 50)	small beta
i	121(100, 25)	93(90, 10)	190(180, 20)	b-a link 1*/
j	121(100, 25)	93(90, 10)	216(205, 20)	b-a link 2*/
f	93(90, 10)	93(90, 10)	246(230, 30)	left h.hel
1	121(110, 20)	93(90, 10)	15(5, 25)	turn 2
2	93(90, 10)	121(110, 20)	43(40, 10)	turn 7
3	93(90, 10)	93(90, 10)	308(300, 20)	turn 1'
4	121(110, 20)	93(90, 10)	344(340, 10)	turn 2'
5	121(110, 20)	93(90, 10)	4(0, 10)	turn 6a1
6	121(110, 20)	93(90, 10)	346(340, 10)	turn 6a2
7	121(110, 20)	93(90, 10)	30(26, 10)	turn 6b
C	117(75, 85)	117(75, 85)	180(0, 360)	coil

Tabla. 2.2 Tabla 'ds.table' utilizada en nuestro programa que contiene un listado de diferentes estados discretos que representan el espacio (θ_1, τ, θ_2). Cada estado corresponde a un tipo de estructura secundaria de los habitualmente conocidos [101].

En este capítulo se presenta un programa de simulación de la estructura de las proteínas y de su plegamiento, basado en la representación (θ_1, τ, θ_2) de la cadena Cα de la proteína. Se describe la estructura lógica del programa e implementación en lenguaje C de todas las subrutinas necesarias para resolver el problema de simulación.

2.- ASPECTOS GENERALES DEL PROGRAMA DE SIMULACIÓN

2.1 Diseño y subrutinas del programa de modelización

El programa presentado en esta tesis fue diseñado para muestrear, al nivel de complejidad escogido, el espacio conformacional de una proteína. Se ha concebido, para hacerlo alrededor de una estructura de referencia, aunque también es posible utilizarlo sin esta, con funciones de energía estadísticas. Los aspectos principales del programa son:

- Modelo de la proteína basado en la simplificación del espacio conformacional mediante el uso de tripletas de torsionales (θ_1, τ, θ_2) para la cadena principal Cα.

- Utilización de funciones de energía simplificadas que reflejan diferentes propiedades estructurales, basadas en tomar una estructura de referencia determinada; por ejemplo, crmsd, drmsd, etc.

Además de estas dos funcionalidades principales, también se han implementado herramientas de manipulación manual de ciertos aspectos estructurales. Finalmente, el programa dispone de subrutinas encargadas de los controles de entrada / salida, para facilitar su manejo. Todo ello se describe más abajo, con más detalle.

El diseño del programa (Tabla 2.3 y fig. 2.4) es modular para hacer más inteligible su lectura y facilitar el desarrollo de versiones posteriores, en las que se mejoren subrutinas ya escritas y/o programen nuevas funcionalidades. El lenguaje escogido para su desarrollo es el lenguaje C, ya que se ha querido primar la velocidad de cálculo y la portabilidad del código entre computadores que pueden tener capacidades de cálculo muy heterogéneas, desde estaciones de trabajo sencillas hasta clústeres masivos.

SUBRUTINAS DE CONTROL PRINCIPAL	
main.c	Programa principal que gestiona la ejecución global e invoca las subrutinas necesarias
ci.c	Intérprete de comandos. Se utiliza para controlar internamente la ejecución
man.c	Subrutinas que fijan aspectos principales de la ejecución y que dependen de parámetros proporcionados por el usuario, están relacionadas con dpar.c
dpar.c	Subrutinas para el establecimiento de los parámetros que definen la dinámica
SUBRUTINAS DE ENTRADA / SALIDA DE INFORMACIÓN	
io.c	Subrutinas que regulan la entrada / salida de las estructuras que se están manejando y simulando
enq.c	Subrutinas de información que permiten al usuario saber qué está ocurriendo en cierto punto de la ejecución
SUBRUTINAS DE EJECUCIÓN DE LA SIMULACIÓN	
dyn.c	Subrutinas de dinámica de Monte Carlo. Controla los diferentes aspectos de la evolución en la simulación temporal. Realiza los cálculos de viabilidad estructural y energética
energy.c	Subrutinas de cálculo de energía en cada conformación. Se utilizan para aceptar / rechazar estructuras generadas
kab.c	Subrutinas para la superposición de moléculas y cálculo del crms
ds.c	Subrutinas relacionadas con la gestión de la representación de la proteína en coordenadas internas
move.c	Subrutinas para los diferentes tipos de movimientos de la cadena principal
SUBRUTINAS DE CALCULO	
vec.c	Subrutinas encargadas de realizar cálculos vectoriales rutinarios
misc.c	Subrutinas de cálculos matemáticos y gestión de entrada / salida
geom.c	Subrutinas del cálculos de ciertos aspectos geométricos de la estructura

Tabla. 2.3 Las principales subrutinas del programa de modelado, clasificadas por tipo de función

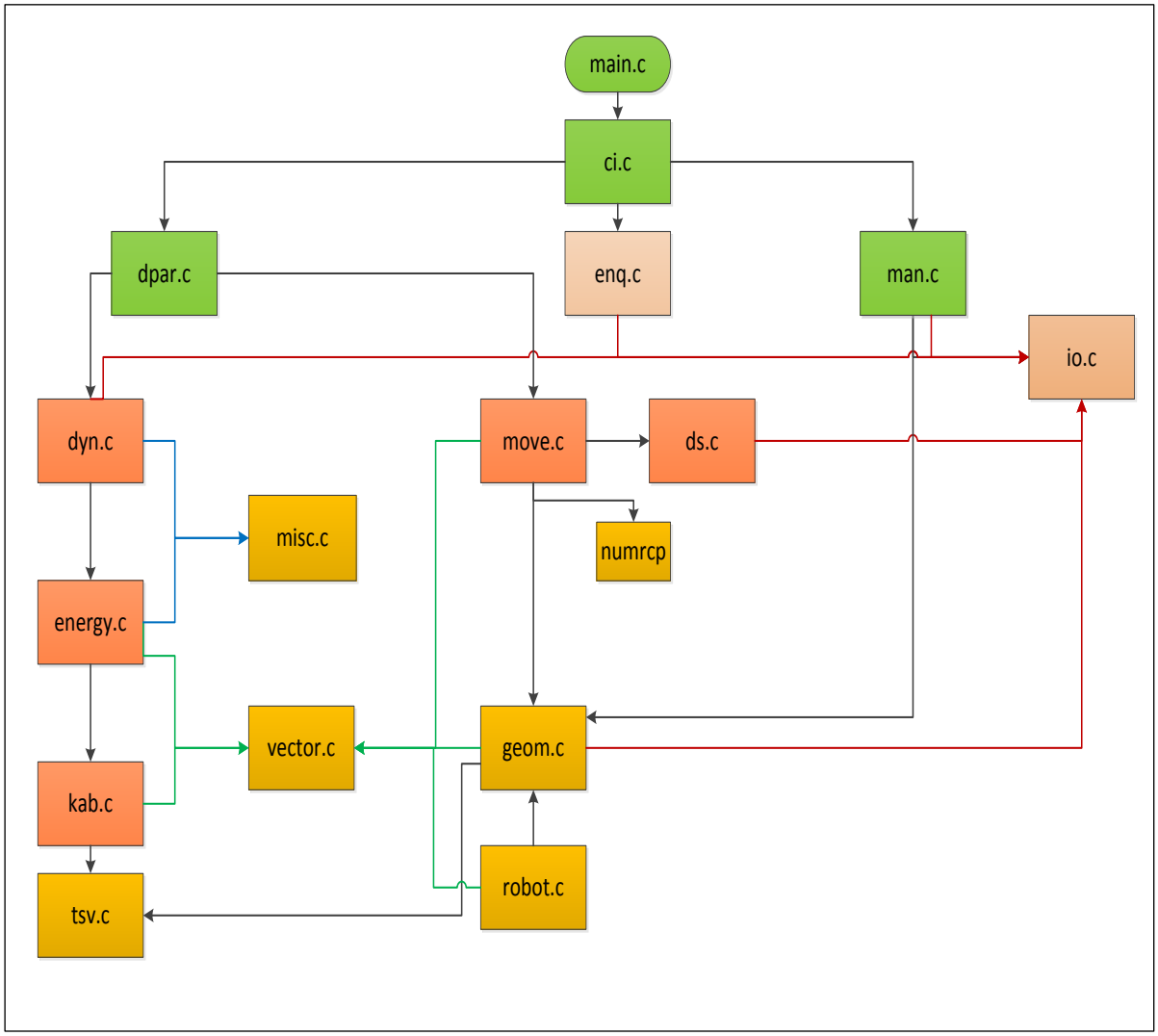


Fig. 2.4 Diagrama de la estructura del programa con las subrutinas y su interrelación. Los colores corresponden al tipo de función, y coinciden con los utilizados en la tabla 2.3

2.2. Generación de conformaciones

2.2.1- Algoritmo de metrópolis: Descripción

El método utilizado para la obtención y evaluación de nuevas conformaciones de las cadenas de proteínas está basado en el algoritmo de Metrópolis. Este popular algoritmo se ha utilizado para obtener valores de las propiedades de equilibrio en un amplio rango de sistemas clásicos. Desde su publicación por Metrópolis, Rosenbluth, Teller [102], el aumento en potencia de cálculo de los computadores ha hecho cada vez más popular y exitosa la utilización de este algoritmo.

El algoritmo de Metrópolis consiste en muestrear de forma aleatoria la configuración de un sistema partiendo de una conformación dada y repitiendo un gran número de pasos. En su versión general, cada paso consiste en intentos de transición hacia una nueva conformación, escogiendo entre una serie de movimientos permitidos y aceptando la nueva conformación únicamente si cumple una determinada condición. En nuestro caso, el muestreo se obtiene perturbando de forma aleatoria una configuración inicial de la estructura tridimensional de la proteína (X_o), generando una configuración de prueba (X_o'). Esta nueva configuración se aceptará, pasando a formar parte de la trayectoria ($X_1 = X'_o$) si su energía es más pequeña o igual que la de la configuración inicial ($U(X_o') \leq U(X_o)$). En caso contrario, la probabilidad de aceptación dependería de la función de probabilidad de Boltzman para una temperatura dada:

$$P_{acceptacion} = \min\left[1, \frac{\exp((-U(X') - U(X))}{K_o T}\right]$$

donde U es la energía potencial y K_oT es la temperatura absoluta en unidades de constante de Boltzman. Cuando el intento de nueva configuración no es aceptado, la nueva configuración se considera igual a la previa ($X_1 = X_o$). Este proceso se repite millones de veces para garantizar un muestreo adecuado para todos los grados de libertad. En la fig. 2.5 se puede ver claramente el proceso seguido.

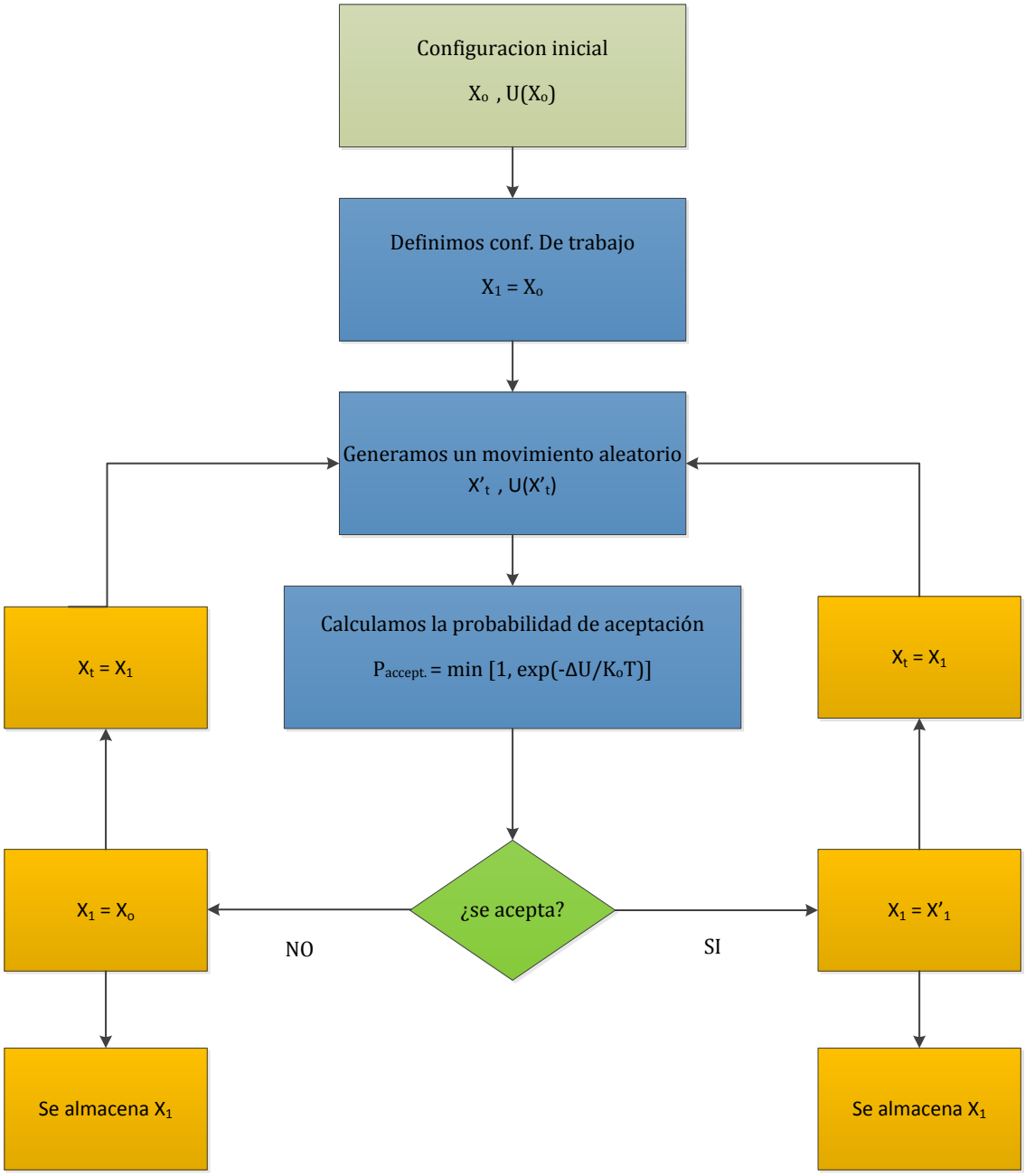


Fig. 2.5 Diagrama de flujo del proceso de elección de conformaciones en una simulación de Montecarlo, cuando la energía de la nueva conformación es mayor que la de la conformación original.

2.2.2. Elección de conformaciones: Términos de energía

Para los términos de energía consideramos varios, relacionados con la estructura tridimensional de la proteína en relación a la estructura de referencia. A continuación exponemos los más utilizados, que nos permiten realizar simulaciones controladas, en las que la función minimizada corresponde a la desviación respecto a la estructura de referencia de la proteína.

- crmsd

'Coordinate root mean-square deviation' (crmsd) es una métrica que se utiliza para cuantificar la similaridad de dos conformaciones [111,112]. El crmsd compara dos conjuntos de coordenadas cartesianas correspondientes a los átomos de dos conformaciones de una molécula y calculando su desviación promedio, o rmsd (root mean-square deviation). El resultado final requiere encontrar la superposición óptima (eliminando la rotación/translación entre las dos moléculas) de las conformaciones comparadas.

Para exponerlo mejor veamos el siguiente ejemplo: consideramos los siguientes parámetros: molécula M con n átomos r_1,...., r_n, supongamos dos conformaciones t y t' de M, y sean $r_i(t)$ y $r_i(t')$ las posiciones del átomo 'i' en las conformaciones t y t', respectivamente. Definimos el rmsd entre dos conformaciones como:

$$\mathrm{rmsd}(t, t') = \sqrt{\frac{1}{n}\sum_{i=1}^{n} ||r_i(t) - r_i(t')||^2}$$

crmsd es el rmsd mínimo sobre las transformaciones rígidas T entre t y t'.

$$\mathrm{crmsd}(t, t') = min_T \sqrt{\frac{1}{n}\sum_{i=1}^{n} ||r_i(t) - T[r_i(t')]||^2}$$

En la Fig. 2.6 se observa mejor como influye la orientación relativa entre las dos conformaciones para el cálculo del rmsd

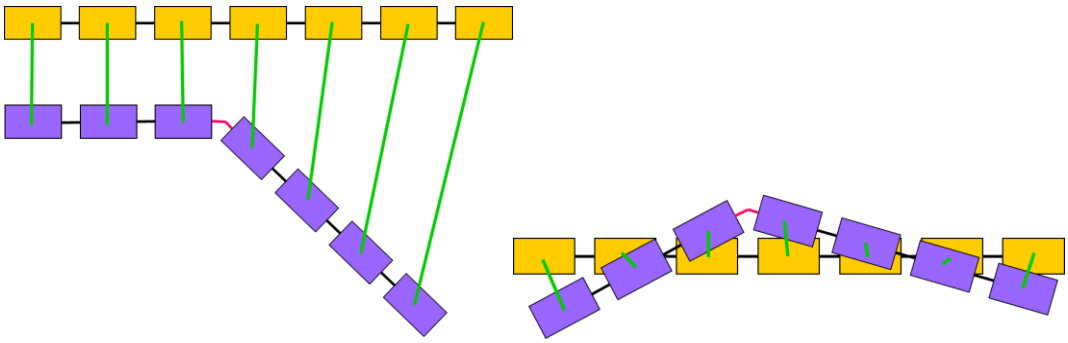


Fig. 2.6 Representación esquemática de dos conformaciones de una estructura. El cálculo de rmsd depende de la orientación de estas dos conformaciones. En la imagen de la derecha se ha realizado una traslación y rotación de una de las conformaciones para minimizar el rmsd y obtener así el crmsd.

Para la obtención de la superposición se utilizó el algoritmo Kabsch [117] que emplea un método de diagonalización para calcular la matriz de rotación óptima que minimiza el rmsd entre dos series de puntos.

Una de las desventajas del crmsd es que no considera explícitamente las distancias entre pares de átomos en la misma conformación. Esto hace que el crmsd no esté necesariamente relacionado con la energía, la cual es una función de las distancias interatómicas [113]. Otra desventaja del crmsd reside en que el resultado depende de un proceso de optimización que es más costoso en tiempo computacional que el propio cálculo del rmsd.

- drmsd

Otra métrica basada en la geometría frecuentemente utilizada es la 'distance root-mean square deviation' (drmsd), basada en las distancias entre los átomos de las dos conformaciones [114-116]. Se define a continuación: sea la molécula M y sus conformaciones t y t', y sean $[d_{ij}(t)]$ y $[d_{ij}(t')]$ las matrices n x n de las distancias entre los átomos de la molécula M en las conformaciones t y t', respectivamente, definimos drmsd(t, t') como:

$$\text{drmsd}(t, t') = \sqrt{\frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} (d_{ij}(t) - d_{ij}(t')^2}$$

Esta métrica no necesita un alineamiento estructural entre las dos conformaciones y, por lo tanto, es menos sensible a los átomos que presentan grandes desviaciones estructurales entre t y t'. Adicionalmente, muestra una más alta relación con la energía [113]. En su contra, hay que señalar que el resultado del drmsd suele estar bastante influenciado por el peso de las distancias interatómicas más grandes. Para reducir este efecto se suele considerar solo las distancias más pequeñas que un cierto valor [115].

Una desventaja del drmsd respecto al crmsd es que utiliza más memoria para el almacenamiento de las matrices de distancia.

El programa desarrollado permite la utilización de estas dos métricas (crmsd y drmsd) como función de energía asociada al proceso de Montecarlo, para el cálculo de nuevas conformaciones. También permite el uso de variantes de dichas métricas, como por ejemplo el cálculo del drmsd utilizando solo los átomos vecinos. De la misma forma se pueden combinar estás métricas añadiéndoles un peso determinado a cada una de ellas, para el cómputo total de energía.

2.2.3 Construcción de conformaciones: Método de 'loop clousure'

Uno de los aspectos a tener en cuenta en las simulaciones de plegamiento de proteínas, es la necesidad de refinar localmente las conformaciones generadas, para minimizar mejor su energía. Para este proceso local es necesario aplicar algún método que no sea computacionalmente muy costoso y permita explorar el espacio conformacional de partes concretas de la proteína, dejando inalterado el resto. Para esto el programa se basa en el algoritmo SPC [103] (Stochastic Partial Closure), que definimos a continuación.

El algoritmo se describe en la fig. 2.7 y se explica de la siguiente manera.

Supongamos 5 átomos consecutivos de la cadena. Si al mover el quinto átomo, su distancia de enlace con el cuarto átomo queda fuera del rango tenemos dos posibilidades para reparar la cadena:

a) Si las esferas centradas en los átomos 3 y 5 con radio –d- tienen intersecciones, simplemente basta colocar el átomo 5 en una de las intersecciones (fig. 2.7a)

b) Si no se produce dicha intersección se procede de la siguiente manera (fig. 2.7b):

   1- Se encuentra el punto de intersección (IP) entre la línea que une los átomos 4 , 5 y la esfera de radio –d- centrada en el átomo 5 (S)

   2- En el punto de intersección IP se construye un plano tangente a la esfera S (S^T)

   3- Se escoge un punto aleatorio P^T de la superficie tangencia S^T usando una distribución normal bidimensional de media centrada en el punto IP y desviación estándar σ_{spc}. La nueva posición del átomo 4 se obtiene proyectando P^T en la esfera S. Este procedimiento se realiza de forma recursiva en una cadena de átomos hasta que llegamos a una zona donde las dos esferas tienen intersecciones (fig. 2.7c)

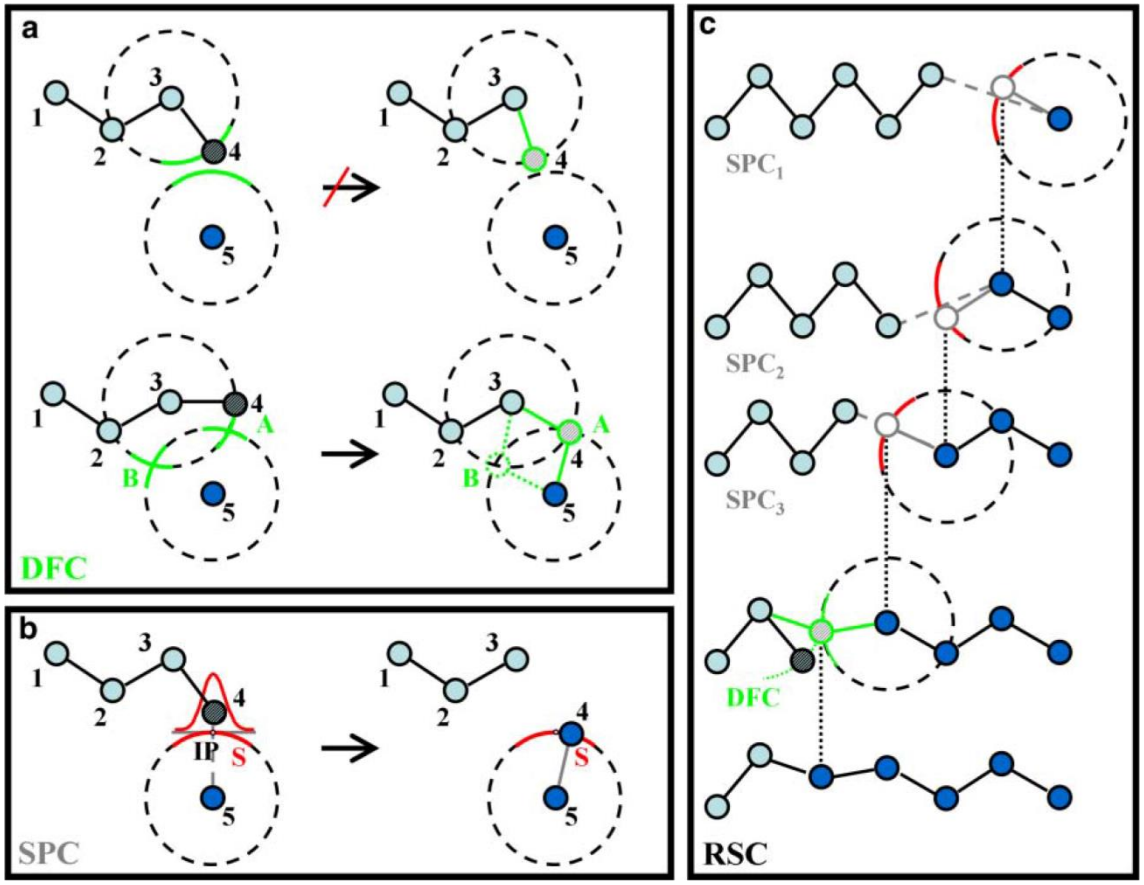


Fig. 2.7 Ilustración bidimensional del método utilizado para los movimientos locales de la cadena polipeptídica. En la figura superior izquierda vemos la aplicación del DFC (Deterministic Full Closure) cuando las esferas de prueba tienen intersecciones. En la figura inferior derecha se muestra el método SPC (Stochastic Partial Closure) y en la figura de la derecha se aplica el RPC (Recursive Stochastic Closure)

3.- ILUSTRACIÓN DEL FUNCIONAMIENTO

Como se ha visto en la sección anterior, el programa realizado ofrece al usuario diferentes opciones para las simulaciones. Para ilustrar las tres más relevantes consideraremos una estructura de referencia como ejemplo: la cadena A de la proteína 1SPH, a la que aplicaremos nuestro programa y veremos los diferentes resultados que podemos obtener.

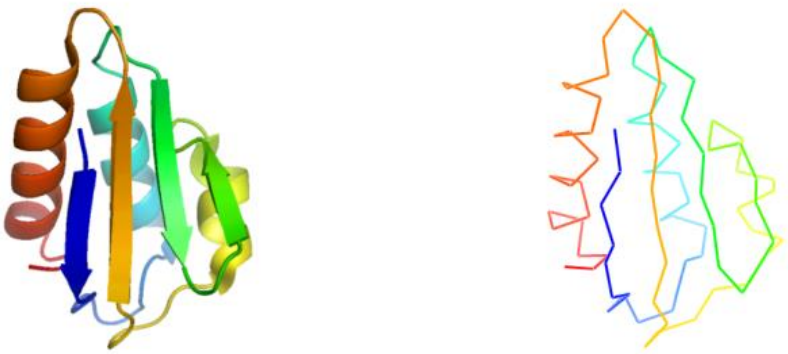


Fig. 2.8 En la imagen de la izquierda representamos la estructura tridimensional de la cadena A de la proteína 1SPH resaltando su estructura secundaria. En la imagen de la derecha representamos solo los átomos Cα de la cadena principal de la misma proteína.

En las opciones escogidas, se muestra cómo el programa permite explorar el espacio conformacional de las proteínas bajo diferentes restricciones geométricas, utilizando la representación simplificada de los residuos mediante el único uso de los Cα. Veremos la estrategia de muestreo del espacio conformacional, basada en modificar al azar una o varias de las coordenadas internas de la proteína, recalcular la estructura resultante, evaluar su viabilidad de acuerdo con los criterios del usuario, y si es aceptada, repetir el ciclo. El primer ejemplo consiste en generar variaciones estructurales en torno a la estructura cristalográfica de la proteína; el segundo elimina esta restricción, permitiendo variaciones libres de la cadena polipeptídica; finalmente, el tercero se centra en las variaciones estructurales de una sola parte de la proteína, un loop exterior (o sea, expuesto mayoritariamente al solvente). Esta última parte involucra una aproximación ligeramente diferente a la generación de conformaciones, ya que corresponde a un movimiento estrictamente local.

3.1 Variaciones alrededor de la estructura nativa de la proteína

En este caso se utiliza como restricción el crmsd global, y se permiten fluctuaciones en los valores de los torsionales de cada átomo C_α, pero sin abandonar la zona de estructura secundaria que ocupa en la estructura nativa. En el script de la

fig. 2.9 se muestran los grandes bloques de comandos mediante los cuales se construye y lanza la simulación, y cómo finalmente se salvan los resultados.

```
io; rs ./1SPHA.seq; rr ./1SPHA.pdb; sstr -a ./1SPHA.ds;q;   → E/S de ficheros
dpar; seed h20t10; nstep 500; nspb 10;                      → generación de semilla aleatoria
    R 1.0; T 2.5 0.005; bump 2; enmps 1;                    → Definición de diversos parámetros
    wet  0 0 0 0 0 0 0 0 0 0                                → Pesos relativos de términos energéticos
    mxenr 9 3                                               → limitación del crmsd respecto a la ref.
    trj -n 10 1SPHA.trj                                     → fichero de salida de la trayectoria
    mtf 0. 0. 1. 0. 0. 0. 0. 0                              → Movimientos sin alterar estructura sec.
io; wcc 1SPHAf1.pdb; wic 1SPHAf1.int; wds 1SPHAf1.ds        → Salida de resultados
```

Fig. 2.9 Script ejemplo utilizado por el programa para la generación de modelos donde se ilustran los diferentes parámetros y su función

Los 'dscode' corresponden a las coordenadas internas asociadas a cada residuo, que el programa obtiene de la tabla 'ds.table' (Tabla 2.2.), la cual almacena todos los estados discretos admitidos por el programa. Cada estado tiene asociada su frecuencia, tal como se ha calculado (de la Cruz et al. [101]) a partir de las estructuras almacenadas en la base de datos PDB.

```
# 1SPHA.seq - Reference structure
# rcode, seq.no., dscode, ba(deg.), ta(deg.)

 A      1          C     117       180
 Q      2          C     143       182
 K      3          B     132       217
.....
```

Fig. 2.10 Primeras líneas del fichero '1SPH.ds' donde se definen los estados discretos según la tabla 2.2

En la fig. 2.11 podemos comprobar cómo los modelos generados se distribuyen en torno a la estructura original, sin exceder el crmsd especificado por el usuario en el script.

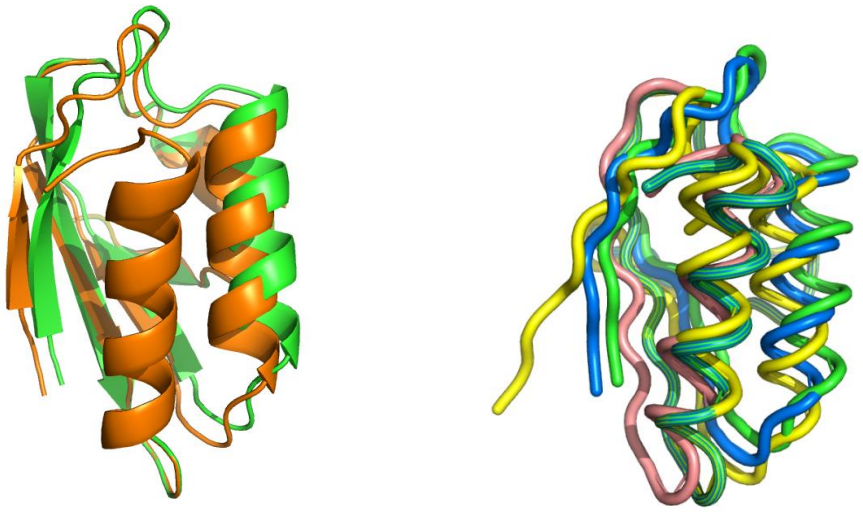


Fig. 2.11 En la imagen de la izquierda se muestra un movimiento sencillo, que da lugar a la generación de una nueva estructura en la simulación, con un crmsd de 3 Å respecto a la nativa. En la imagen de la derecha se muestran cuatro estructuras generadas durante la simulación, cumpliendo todas ellas la restricción de 3 Å de crmsd.

3.2 Variaciones estructurales libres, sin restricciones conformacionales

En el capítulo IV de los resultados, realizamos un estudio que requiere un muestreo del espacio conformacional de la proteína libre de restricciones conformacionales. Es decir, se permitirá a la proteína variar su estructura sin necesidad de que mantenga un determinado grado de similitud con la estructura experimental. Para ello eliminaremos las restricciones conformacionales impuestas por el crmsd (o drmsd), preservando únicamente la integridad covalente de la proteína (es decir, no se romperá la cadena polipeptídica). Las estructuras generadas constituirán una muestra al azar del espacio conformacional de la proteína. Respecto a la sección anterior, aquí se utilizan también variaciones en el muestreo de las coordenadas internas, pero se permitirá que los residuos de la proteína cambien su estado de estructura secundaria. Para realizar este tipo de simulación libre, basta con modificar los siguientes comandos en la simulación anterior.

~~*mxenr 9 3*~~	→ se elimina la limitación por crmsd
mtf 0. 0. 0. 0. 0. 0. 0. 0	→ mismo peso relativo de tipos de movimiento

Fig. 2.12 Representación gráfica en la que vemos la estructura de referencia en color verde y dos modelos generados sin restricciones de crmsd. Esta aproximación se utilizó en el capítulo IV.

3.3 Modelización de un loop

En el capítulo III describimos el uso de nuestro programa para el muestreo del espacio conformacional de las regiones altamente repetitivas de las proteínas, con la finalidad de averiguar la sensibilidad de este espacio a las mutaciones potencialmente dañinas. En este caso, el muestreo del espacio conformacional se limitó a los residuos que pertenecían a dicha región. Para ello se implementó una solución sencilla al conocido problema del loop-closure anteriormente explicado. Debajo mostramos los comandos que sería necesario modificar en nuestro ejemplo.

dpar; loop 65 69 5 m → Definición del loop

mtf 0. 0. 1. 0. 0. 0. 0. 0 → Generación de movimientos tipo loop

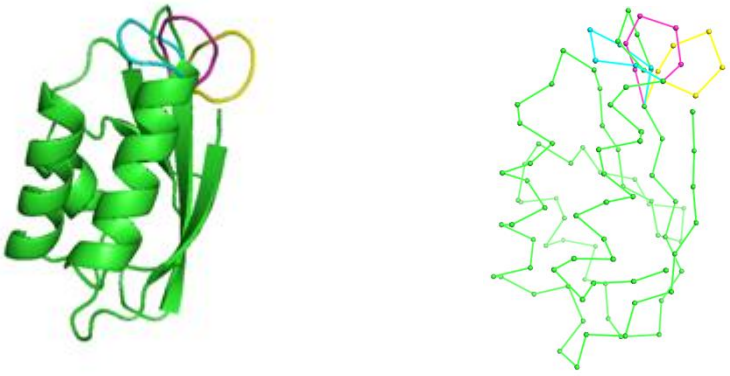


Fig. 2.13 Representación gráfica en la que vemos la estructura de referencia en color verde y 3 loops generados (la representación de la derecha muestra los CA). Esta aproximación se utilizó en el segundo capítulo de resultados simulando diferentes conformaciones de LCRs (Low Complexity Regions)

CAPITULO 3: LAS REGIONES DE BAJA COMPLEJIDAD (LCRs)

En este capítulo se presenta una primera aplicación del programa de simulación presentado. Nos centramos en un problema de claro interés biológico como son las regiones de baja complejidad de las proteínas y las variaciones estructurales que presentan sus mutaciones. La aplicación del programa nos permite caracterizar de forma rápida el espacio conformacional de las LCRs y el de sus posibles mutaciones.

1.- INTRODUCCION

Las LCR (Low Complexity Regions) son secuencias de aminoácidos que contienen repeticiones de un solo aminoácido o de pequeños motivos de aminoácidos [118]. Este tipo de secuencias son muy abundantes en las proteínas de los eucariotas [119]. De hecho, en muchas especies eucariotas la mayoría de sus proteínas muestran una tendencia a la repetición mayor de lo esperado, dada su composición de aminoácidos [120].

1.1. LCRs: origen

Desde el punto de vista de secuencia, un aspecto interesante de las LCRs es que son altamente variables, es decir que su secuencia puede ser diferente entre individuos de la misma especie. Ello se explica a nivel de ADN por la acción combinada del "replication slippage" y de la recombinación [121]. Una consecuencia de esta inestabilidad de secuencia es que en algunos casos, la expansión incontrolada de motivos de secuencia corta se ha visto asociada a ciertas enfermedades huma-

nas [122]. Dado este potencial dañino que tienen las LCRs puede parecer paradójica su abundancia. Se postula [12] que dicha abundancia sería consecuencia de su capacidad para incrementar la variación fenotípica dentro de las poblaciones, lo que favorecería su adaptación. A nivel molecular, ello sería consecuencia de la capacidad de las LCRs para modular las interacciones entre proteínas [124] y/o su localización celular [125]. Otra hipótesis alternativa sobre la abundancia de LCRs es que facilitan la formación de nuevas funciones [119]: los motivos cortos tienen mayor probabilidad de expandirse lo cual favorece la extensión de motivos "semillas" para formar motivos repetitivos mayores. Este mecanismo, unido a la posterior acumulación de mutaciones en las subsecuencias repetidas, podría dar lugar a la aparición de nuevas funciones.

1.2. LCRs y estructura de proteínas

En algunos casos estas regiones pueden ser usadas como señalizadores, ayudando a mediar en las interacciones entre proteínas. En este caso, las LCR pueden estar más estructuradas al producirse la unión a un ligando específico o al formar un complejo con otra proteína [126]. En la misma línea, se ha observado también que las LCRs pueden ser usadas para unir diferentes dominios dentro de una proteína [127].

Tal como se ha mencionado anteriormente, son muchas las proteínas que presentan LCRs; pero entre ellas destaca un caso particularmente interesante: las proteínas del parasito de la malaria humana *Plasmodium falciparum.* En estas proteínas, las LCRs tienen un tamaño inusual y son muy abundantes [128]. Este hecho ha sido relacionado con una de las características más dañinas de este parásito, su capacidad de adaptación a los fármacos y al sistema inmune humano [129]. Son varios los trabajos [130,131] que han destacado la significancia adaptativa de las LCRs en relación a la respuesta inmune que suscita el parasito. Es decir, se detecta en *P. falciparum* una variabilidad substancial de las secuencias de las LCRs entre diferentes individuos y que dificultaría la acción efectiva del sistema inmune. A nivel biomédico, esta variabilidad en las LCR hace muy difícil el desarrollo de vacunas efectivas [140].

Como acabamos de ver, parte del impacto biológico de la variabilidad de secuencia de las LCR es de origen estrictamente bioquímico. Es debido a que la substitución de un aminoácido por otro cambia la composición atómica de forma local y, por consiguiente, cambia la capacidad de la molécula afectada de interaccionar con su entorno. Pero es indudable también, que asociado a la variación de secuencia se producirá un cambio de las propiedades estructurales de la molécula afectada: tanto a nivel de conformación atómica, como a nivel de la dinámica de esta conformación. A pesar de su posible impacto funcional, no hay ningún estudio conocido que trate este aspecto de las LCRs y de sus variantes. En este capítulo abordamos este problema, y nos planteamos dar una primera respuesta a la pregunta siguiente: ¿cuál es el impacto de la variabilidad de secuencia de las LCRs sobre sus propiedades estructurales? Para responderla, nos hemos centrado en el caso de las LCRs de estructura desordenada, cuyo estudio es difícil de plantear experimentalmente, pero es abordable mediante el tipo de simulación computacional extensa obtenida con el programa desarrollado en esta tesis.

2. – MATERIALES Y MÉTODOS

Los métodos presentados a continuación tienen como finalidad la descripción, para un conjunto representativo de LCRs, de la influencia de los cambios de secuencia (substituciones, inserciones y deleciones) en el espacio conformacional de las LCRs. Con el fin de realizar nuestro estudio, utilizamos cinco LCRs que pudiesen ser representativas de los diferentes comportamientos, y para las que su estudio pudiese ejecutarse con los recursos computacionales disponibles. La principal característica de las LCRs escogidas era su falta de estructura definida, total o parcial, en el experimento de difracción de rayos X de la proteína que las contenía.

A las LCRs escogidas les aplicamos un protocolo de selección y modelización basado en el esquema de la fig. 3.1 que procedemos a explicar en los siguientes apartados

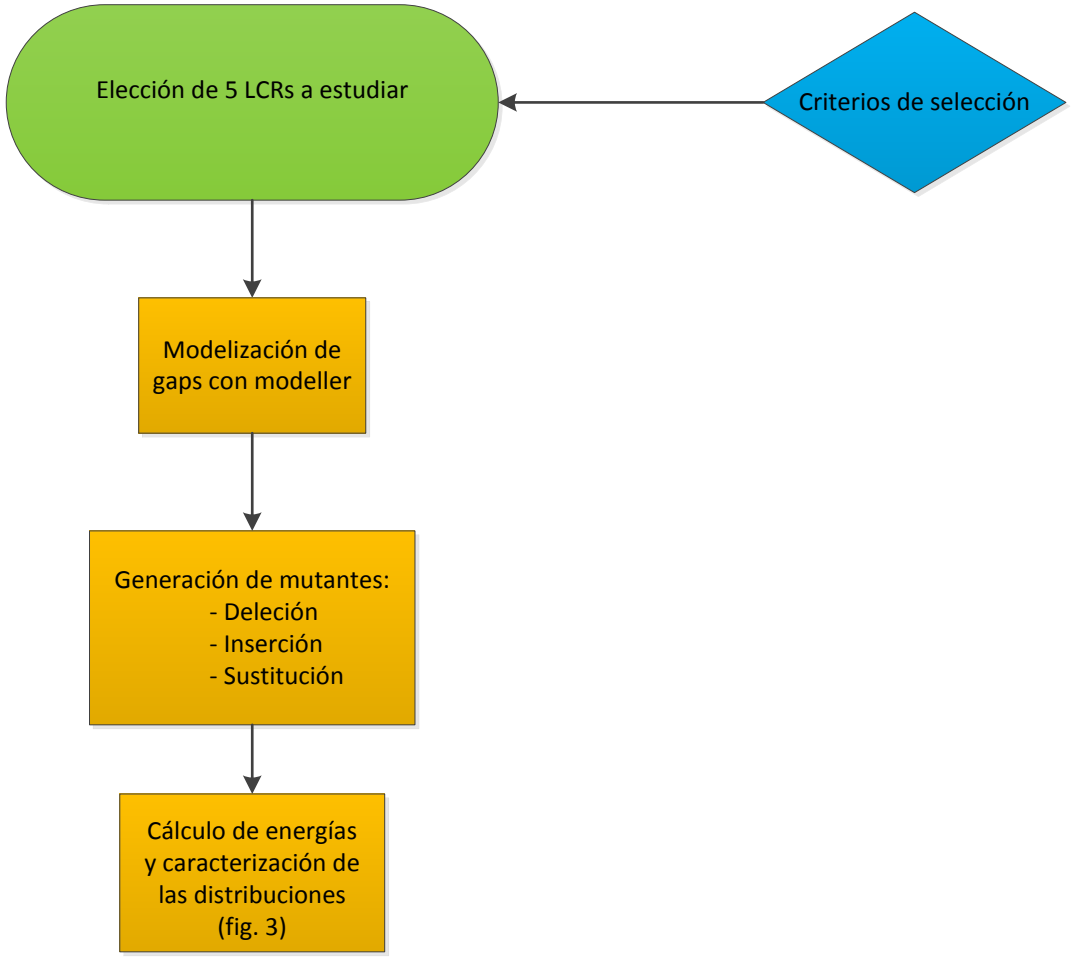


Fig. 3.1 Representación esquemática del proceso seguido para la caracterización de las distribuciones de las LCRs escogidas.

2.1 Criterios de Selección

Primero obtuvimos una lista de posibles candidatos utilizando como fuente la base de datos del Protein Data Bank que contiene todas aquellas proteínas con estructura elucidada experimentalmente. Procedimos a un primer filtrado utilizando las secuencias en formato FASTA y el programa PSEG [132]. Dicho programa permite encontrar las secuencias que tienen LCRs, aplicando una ventana deslizante a lo largo de la secuencia y utilizando medidas basadas en la entropía para determinar el grado de repetitividad de un segmento. El cálculo realizado se realiza mediante el algoritmo SEG, que a su vez se basa en la siguiente fórmula [141]:

$$K = \frac{1}{L} \log_N \left[\frac{L!}{\prod_{all\ i} n_i!} \right]$$

N = número de aminoácidos (= 20)

n_i = n número de aminoácidos (= 20), que

L = tamaro de aminoácido

0 <= K <= 1

En la fig. 3.2 podemos ver dos ejemplos de LCR donde se ve su naturaleza repetitiva. Como muestra dicho ejemplo, el tamaño y composición de las LCR puede variar substancialmente, así como los aminoácidos que las constituyen

Tandem-repeat insertion

PLABE 480 NPPETP gssvvtnpdnvascqnd EDCY 506
PLAFA 502 KPEENP epvekpnpeenpnpvekptpeenpnpvekptpeenpnpvekpepknpcinm EDCY 563

Homopolymer runs

PLABE 1379 CSEWGEWSACPQ lcdra ISKRERKRPyYTIKEEG 1413
PLAFA 1510 CSEWEEWSPCPE tcptinnnnnnssgnnssrtnirtyfpq LSKRERKGP YTLKGEE 1566

Fig. 3.2 Ejemplo de secuencia de dos LCR de distinto tamaño. En colores se señala la parte de la secuencia que corresponde a la LCR

Una vez obtenido el listado base de LCRs, lo restringimos a aquellos casos que tuviesen una estructura parcial o totalmente desordenada [133]. Siguiendo este procedimiento obtuvimos un total de 585 PDBs que almacenamos junto con el nombre del PDB, los lugares de inicio y final en la secuencia de la proteína, así como la secuencia de la propia LCR:

No	*PDB*	*InLcr*	*FinLcr*	*SecLcr*
0001	*1zwyA*	*6*	*17*	*IIKRRVMRKIII*
0002	*2qguA*	*16*	*35*	*VAAVAAVPAHAQEADAQATV*
0003	*1x1kF*	*1*	*30*	*PPGPPGPPGPPGPPGPPGPPGPPGPPGPPG*

…

PDB: Nombre de la proteína en formato PDB (4 primeras letras) y la cadena (última letra)

InLcr: Numero de inicio del residuo en secuencia de la proteína del LCR

FinLcr: Numero de fin del residuo en secuencia de la proteína del LCR

SecLcr: Secuencia de los residuos del LCR (formato 1 letra por residuo)

A partir de este listado aplicamos un segundo filtrado, más restrictivo, basándonos en los siguientes criterios:

- Se eliminan las LCRs con algún residuo no standard
- Se eliminan aquellas LCRs que están al principio o final de la cadena (el criterio es que la LCR tiene que empezar al menos en el tercer residuo de la cadena y acabar antes del penúltimo residuo)
- Se eliminan aquellas LCRs para las que el programa modeller no consiguió modelizar bien los gaps
- Se impone una proporción máxima entre 'longitud secuencia LCR/longitud secuencia Proteína' más adecuada (=10%)
- Eliminamos los no-monómero, i.e, aquellas para los que la estructura de la proteína completa formaba a su vez parte de una estructura mayor
- Se examinaron visualmente y se escogieron las mejores estructuralmente hablando. Los criterios visuales se basaron sobretodo en la posición de la LCR respecto al resto de la proteína. Si la LCR, o parte de ella, se encontra-

ba inmersa en el resto de la estructura de la proteína se descartó, ya que el programa tenía más dificultades para encontrar conformaciones diferentes. Ello es debido a la restricción de grados de libertad que presentaba la LCR. También se descartaron aquellos casos en los que había una gran distancia geométrica (en relación al tamaño de la proteína) entre los extremos de la LCR

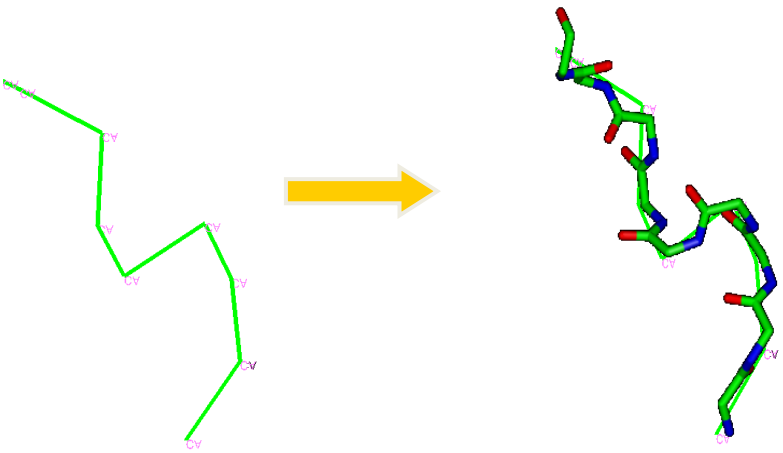
2.2 Modelizado y distribución de conformaciones

En un segundo paso, como las LCRs seleccionadas no tenían estructura y, para explorar su espacio conformacional, se necesitaba una estructura de partida, se procedió a asignarles una estructura inicial arbitraria. Para ello utilizamos el programa Modeller [134] que permite crear modelos a partir de la estructura original de rayos X y de la secuencia extraída de la información SEQRES del propio PDB (y, en casos posteriores, de la de aquellas variantes de secuencia de la LCR que vayamos a generar).

En el siguiente paso, procedemos a explorar el espacio conformacional de la LCR. Para ello obtenemos una colección de posibles conformaciones por cada modelo. Este proceso conlleva varias fases:

1.- Para cada LCR modelizada se construyen 1000 conformaciones de la cadena C_α con nuestro programa (ver capítulo anterior), variando únicamente la estructura correspondiente a la secuencia de la LCR. El resto de la proteína no se modifica.

2.- Para cada una de las conformaciones anteriores se construye la cadena principal con el programa BBQ [135].



3.- Se construyen las cadenas laterales mediante el programa SCWRL4 [136] (fig. 3.3)

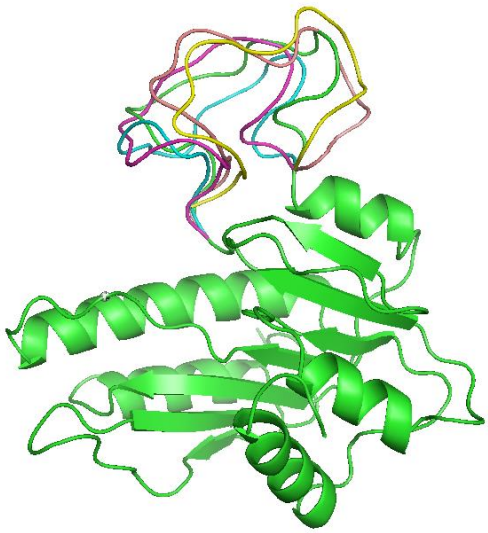


Fig. 3.3 Generación de 5 conformaciones de la LCR de la proteína 2plw. La secuencia de aminoácidos de la LCR es: KDNMNNIKNINYIDNMNNN.

Al final de este protocolo, para cada LCR hemos obtenido una colección de 1000 conformaciones posibles que proporcionan una muestra de su espacio conformacional

2.3 Generación de mutantes y distribución estructural

Con el objetivo de estudiar el efecto de los cambios de secuencia sobre el espacio conformacional de las LCRs, en el siguiente paso nos proponemos generar una serie de mutantes para cada una de las 5 LCRs escogidas. Dichos mutantes corresponden a tres categorías diferentes: deleción, inserción y sustitución. Una vez construida la secuencia correspondiente a cada mutante, se obtiene la muestra de su espacio conformacional siguiendo el método explicado en el apartado 2.1.

2.3.1 Modificación de secuencias de las LCR

Las secuencias de las LCRs se modifican para obtener 3 tipos de mutantes:

a.- Mutantes de sustitución

Se sustituye cada aminoácido de la LCR por otro en base a la siguiente matriz de sustitución

A	-> S	L	-> I
R	-> K	K	-> R
N	-> D	M	-> L
D	-> E	F	-> Y
C	-> A	P	-> D
Q	-> E	S	-> A
E	-> Q	T	-> S
G	-> A	W	-> Y
H	-> Y	Y	-> F
I	-> V	V	-> I

Esta matriz derivada a partir de la matriz BLOSUM62 [137], que comprende criterios de similitud funcional y de probabilidad de ocurrencia de las diferentes substituciones.

b.- Mutantes de inserción

Se introducen sistemáticamente Alaninas entre cada par de residuos de la LCR. Al final de este proceso habrá tantas LCR mutantes como posiciones tenga la LCR

c.- Mutantes de deleción

Se eliminan, sucesivamente, todos los amino ácidos generándose así tantas LCRs como posiciones tenga la LCR nativa.

Si dos o más mutantes dan lugar a la misma secuencia se deja solo uno (e.g. en TGGGA, la eliminación de cualquiera de las tres G dará lugar a tres mutantes con la misma secuencia: TGGA).

2.3.2. Generación de distribuciones estructurales

Una vez obtenidos los mutantes se procederá a crear una colección de estructuras para cada uno de ellos siguiendo el procedimiento del apartado 2.1.

Para cada una de las distribuciones estructurales resultantes, se calcula la energía de todas las conformaciones generadas mediante el programa CalRW basado en un potencial estadístico dependiente de la distancia [136]. Para facilitar la comparación de resultados, las energías calculadas se ponderan en función del valor medio obtenido. Igualmente se realizan cálculos de los crmsd (descrito en el apartado 2.2.2 del capítulo 2) entre las conformaciones de cada distribución y crmsd entre conformaciones mutante–nativa (fig. 3.4)

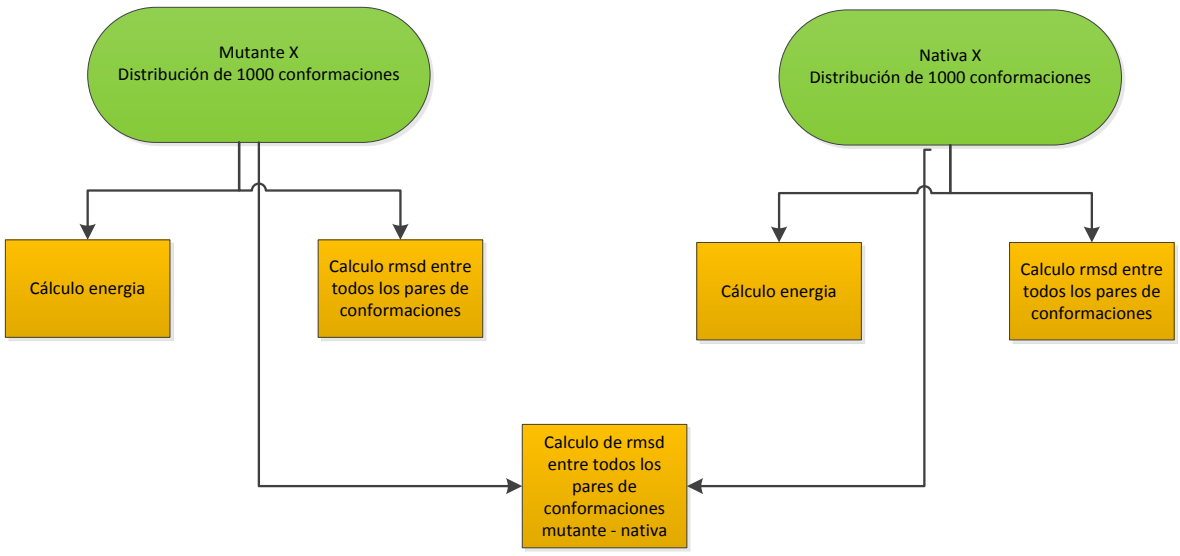


Fig. 3.4 Esquema ilustrativo del proceso seguido para los cálculos de las distribuciones de energía y crmsd de los mutantes y la LCR nativa de cada una de las proteínas estudiadas

Para el cálculo de los crmsd únicamente hemos utilizado la parte de estructura correspondiente a las LCR y no el PDB entero. Para ello hemos utilizado el programa PofitV3 [139].

Para dar más consistencia a nuestros resultados y reducir la posibilidad de que estos presenten un sesgo debido a la estructura de rayos X de las nativas, duplicamos el proceso de la fig.3.4 partiendo de unas estructuras nativas alteradas, con un crmsd un poco diferente del original. A partir de estas nuevas nativas se repite todo el proceso, volviéndose a generar la serie de mutantes, los cálculos de energía y los de crmsd (fig. 3.5).

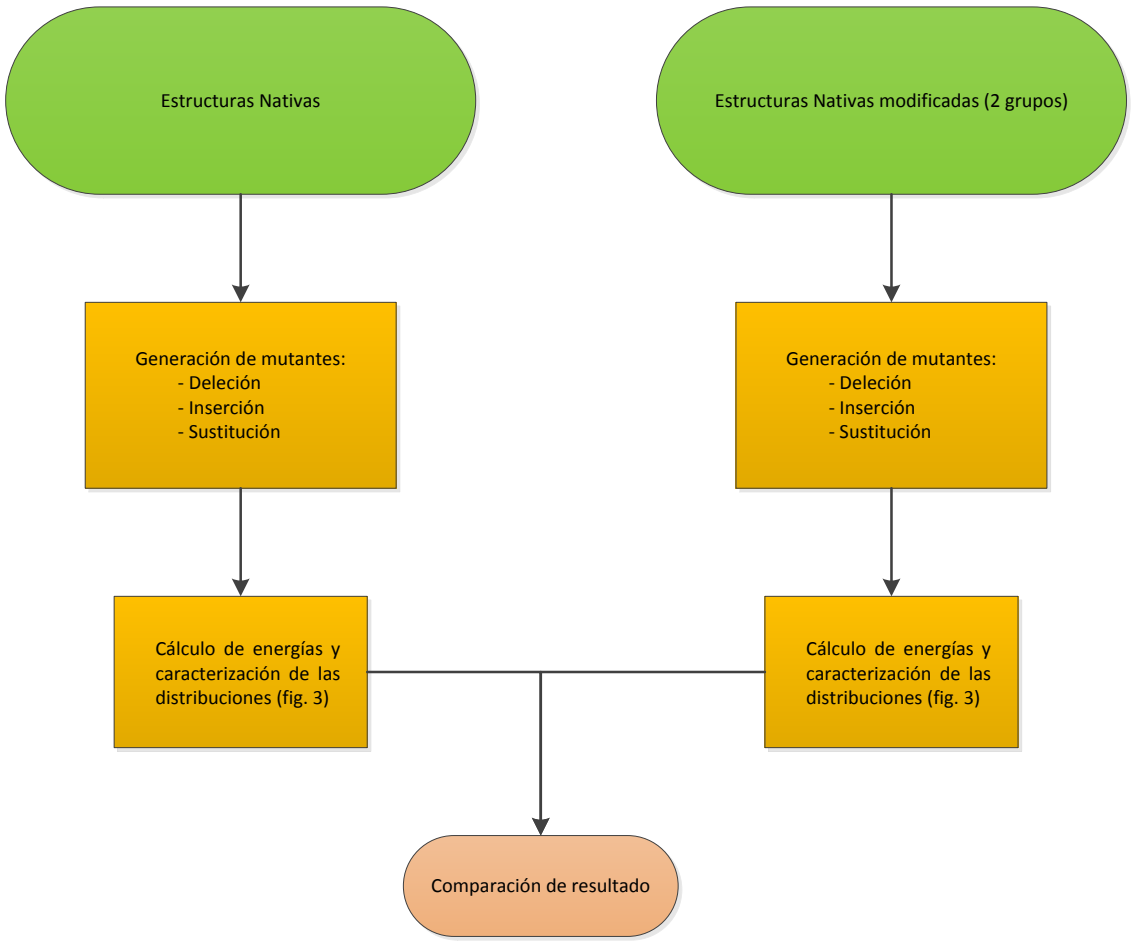


Fig. 3.5 Esquema del proceso de caracterización de distribuciones y energías para cada una de las 5 proteínas seleccionadas. Las estructuras nativas modificadas nos sirven para comprobar la robustez de nuestros resultados

3.- RESULTADOS

Siguiendo el proceso de selección explicado en el apartado 2.1 obtuvimos 5 proteínas con LCRs de interés con los que empezamos a trabajar (fig.3.6):

- Native Cytokinin Dehydrogenase (PDB: 1w1o)
- Deinococcus radiodurans maltooligosyltrehalose trehalohydrolase (PDB:2bhu)
- Flagellar motor switch protein FliM (PDB: 2hp7)
- Cholesterol Oxidase from Brevibacterium sterolicum - His121Ala Mutant (PDB: 2i0k)
- Ribosomal RNA methyltransferase, putative, from Plasmodium falciparum (PDB: 2plw)

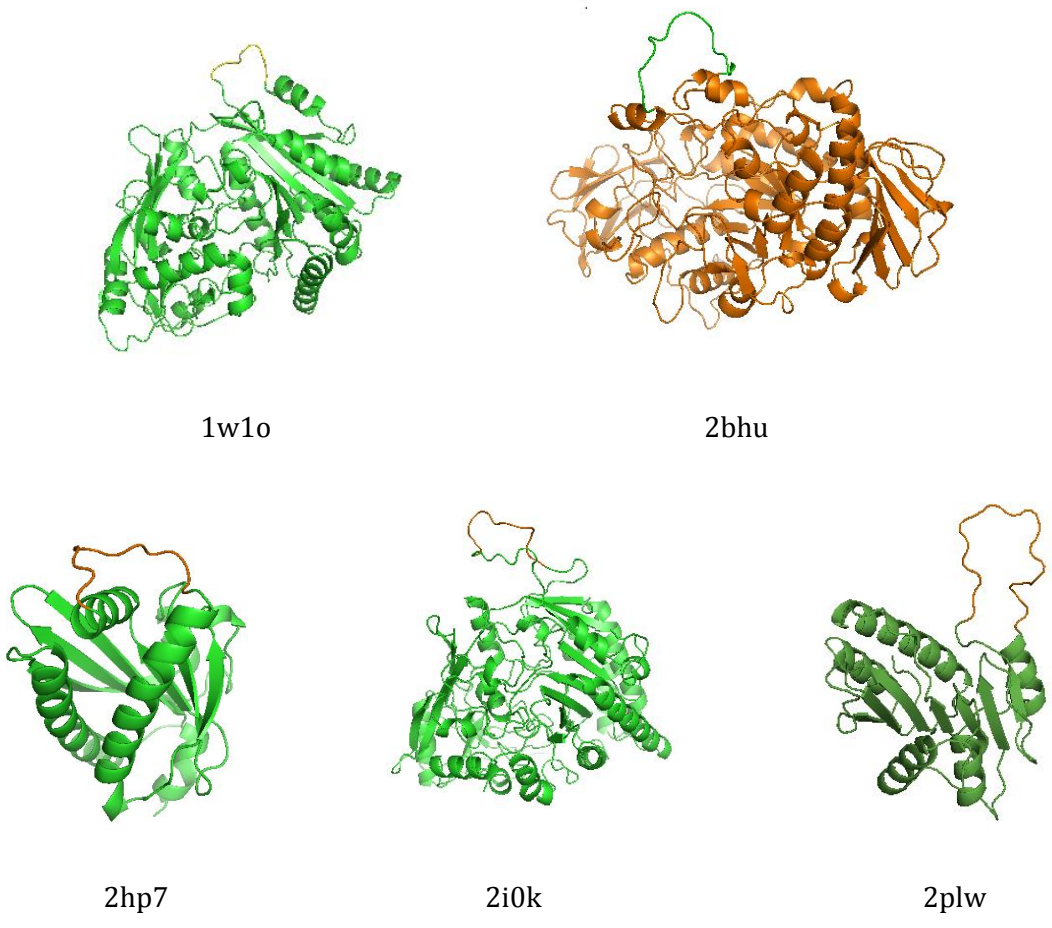


Fig. 3.6 Representación gráfica de las 5 proteínas escogidas para su estudio y el nombre PDB correspondiente. En cada caso, el LCR se identifica por su color diferente al resto de la estructura.

En la siguiente tabla podemos ver las secuencias de las LCR así como los mutantes construidos para la caracterización de las distribuciones (D: deleción; I: inserción; S: substitución).

PDB	LOCALIZACION LCR EN PDB	SECUENCIA LCR	MUTANTES CONSTRUIDOS D/I/S	Total
1w1o	339 - 345	ATAAAAA	3 / 2 / 6	11
2bhu	466 - 481	EGRKKEFGGFSGFSGE	14 / 15 / 16	45
2hp7	89 - 100	GGPGENPPNRPP	9 / 11 / 12	32
2i0k	287 - 298	VGSLGSAGSLVG	12 / 10 / 9	31
2plw	74 - 92	KDNMNNIKNINYIDNMNNN	16 / 18 / 19	53
			Total mutantes: 172	

En la fig. 3.7 podemos ver un ejemplo de dos conformaciones, una con la LCR con su secuencia nativa (en amarillo) y otra con la LCR con la secuencia mutada (en azul).

Fig. 3.7 2bhu con LCR nativa y una conformación mutante con inserción de Alanina. Las secuencias de aminoácidos son las siguientes: EGRKKEFGGFSGFSGE (Nativo, amarillo), EAGRKKEFGGFSGFSGE (Mutante, azul)

A continuación presentamos los resultados correspondientes a tres tipos de cálculo (Figuras 3.8 a 3.12): energías, crmsd auto-comparación y crmsd de la comparación con la nativa. Los primeros representan la energía de cada una de las 1000 conformaciones para una LCR, normalizada por la media del total. El crmsd de la auto-comparación proporciona una medida del espacio conformacional poblado por cada LCR (mutante o nativa); se obtiene, comparando entre si las 1000 conformaciones generadas para cada LCR y calculando el crmsd. El crmsd de la comparación con la nativa nos proporciona una medida complementaria del espacio conformacional poblado por las LCR mutantes (nos dice si es similar o diferente al de la estructura nativa); se obtiene, comparando las 1000 conformaciones generadas para cada LCR mutante con las 1000 conformaciones de la LCR nativa, y calculando el crmsd (se obtienen 1,000,000 de valores). En las figuras 3.8 a 3.12 mostramos un cuadro resumen de los resultados obtenidos. En el eje vertical representamos la frecuencia en millares y en las gráficas de energía se muestra en el eje horizontal la energía relativa a la energía máxima obtenida. En la leyenda de las gráficas se muestra el número de mutantes utilizados en los cálculos. Para una descripción más completa de los resultados, en el apéndice I se han añadido las gráficas más detalladas así como las tablas que permiten identificar a cada mutante (incluyendo el cambio a nivel de secuencia de aminoácidos), con las modificaciones respecto a la secuencia nativa señalado en color rojo. Por simplicidad, en la descripción posterior nos referiremos a las proteínas simuladas mediante su código PDB de cuatro letras.

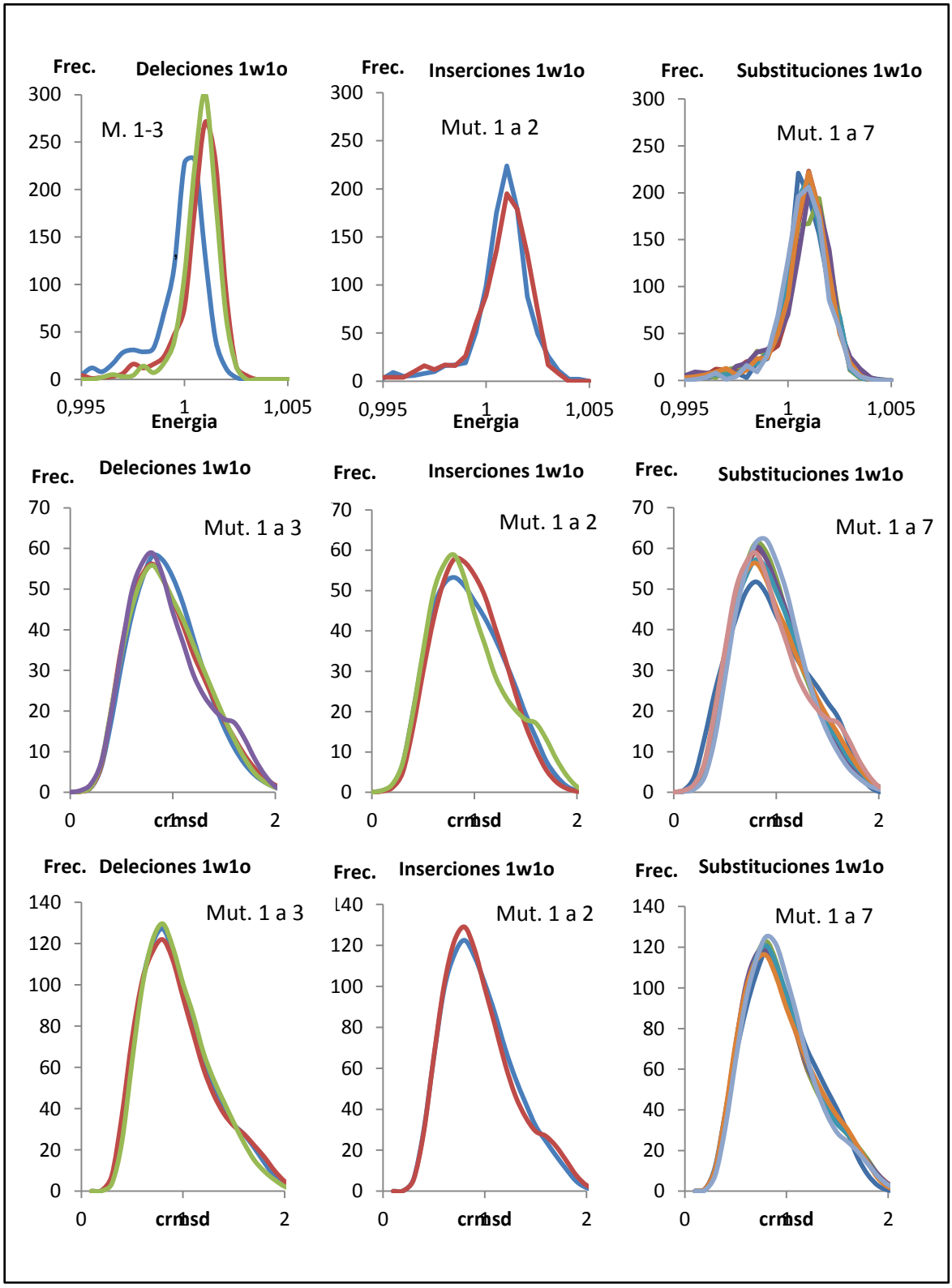


Fig. 3.8 Resultados 1w1o. Primera línea: Cálculo de Energías; segunda línea: Cálculo crmsd Mutante-Mutante; tercera línea: Cálculo crmsd Mutante-Nativa

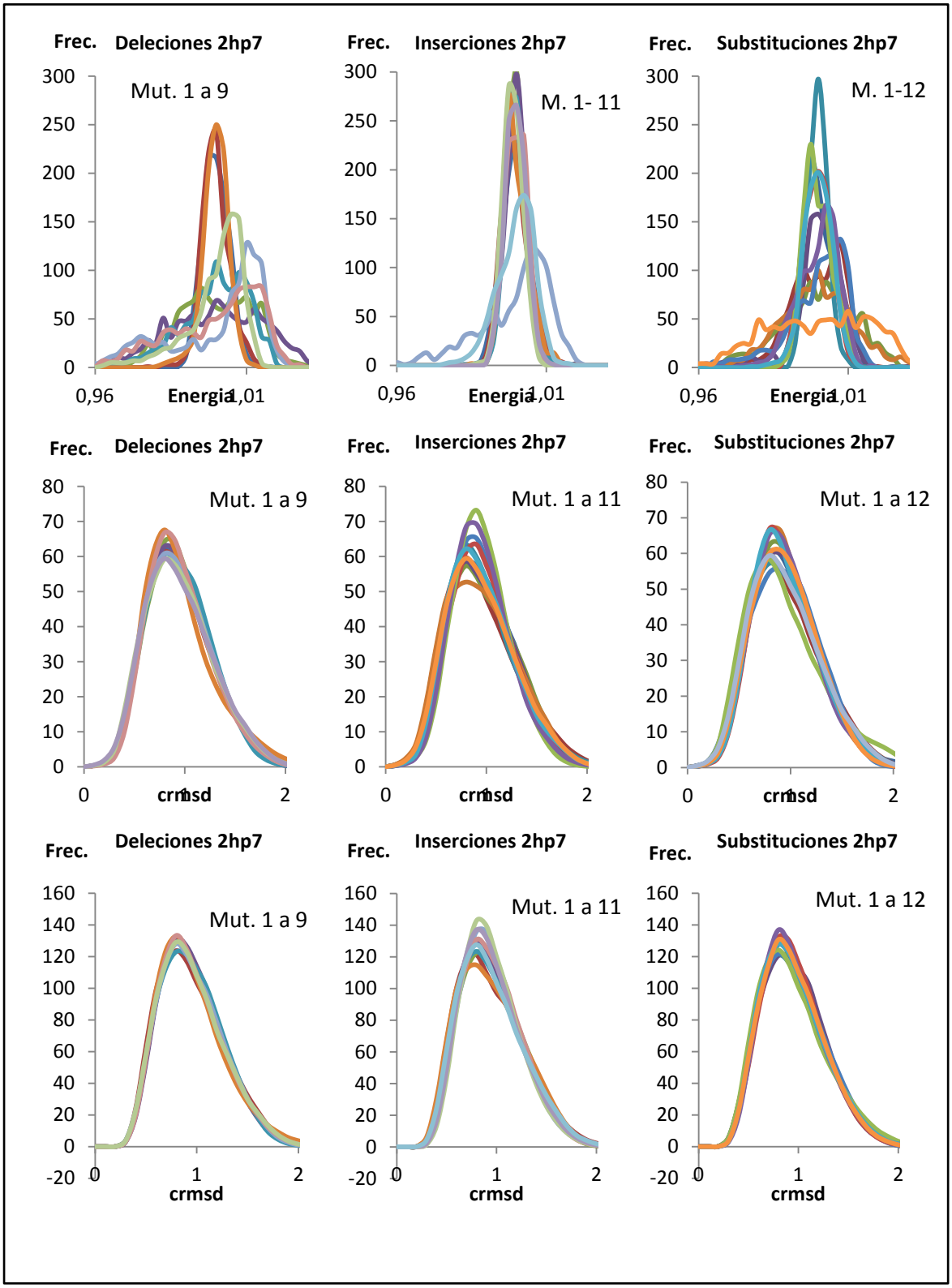


Fig. 3.9 Resultados 2hp7. Primera línea: Cálculo de Energías; segunda línea: Calculo crmsd Mutante-Mutante; tercera línea: Calculo crmsd Mutante-Nativa

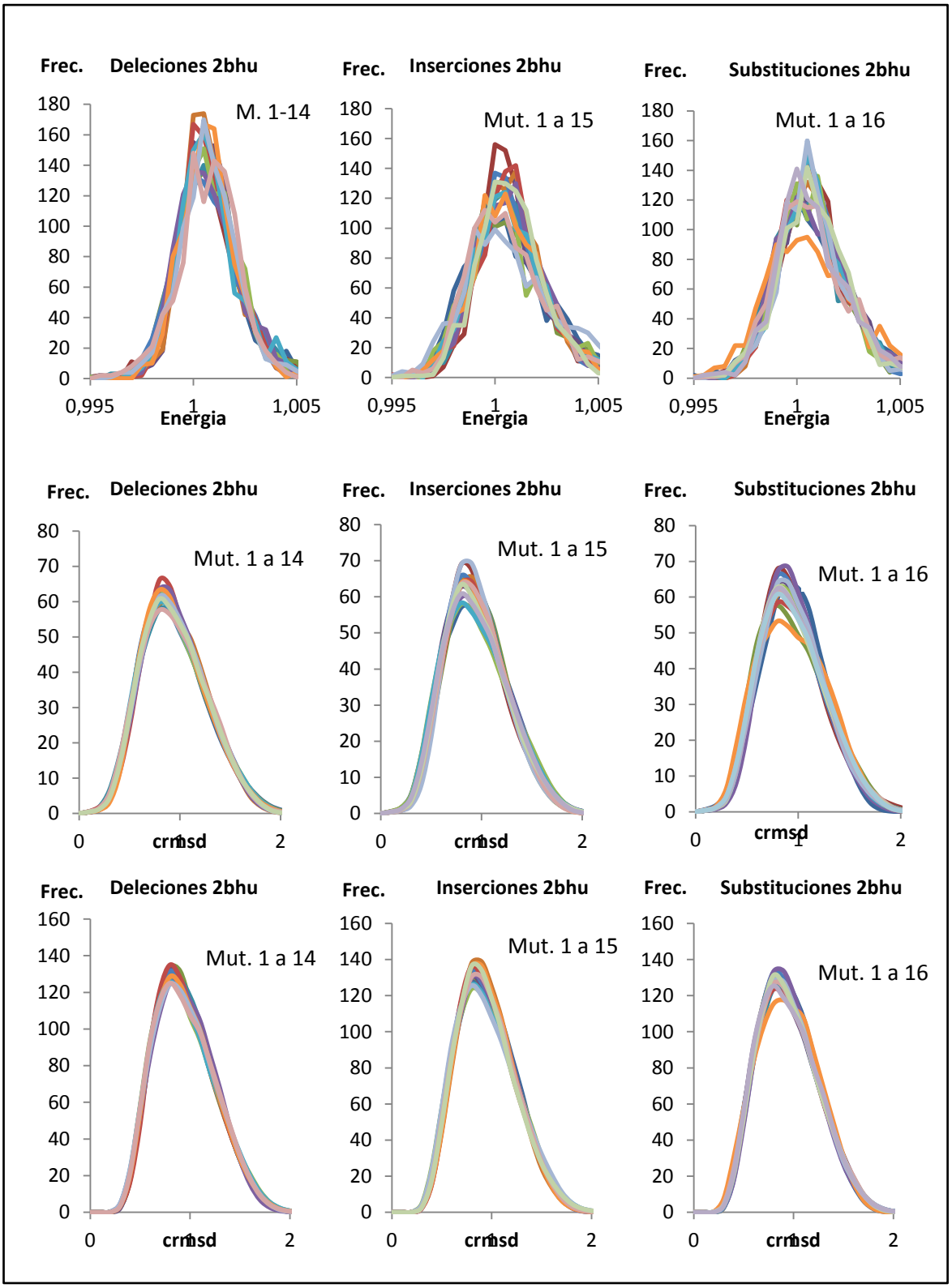


Fig. 3.10 Resultados 2bhu. Primera línea: Cálculo de Energías; segunda línea: Calculo crmsd Mutante-Mutante; tercera línea: Calculo crmsd Mutante-Nativa

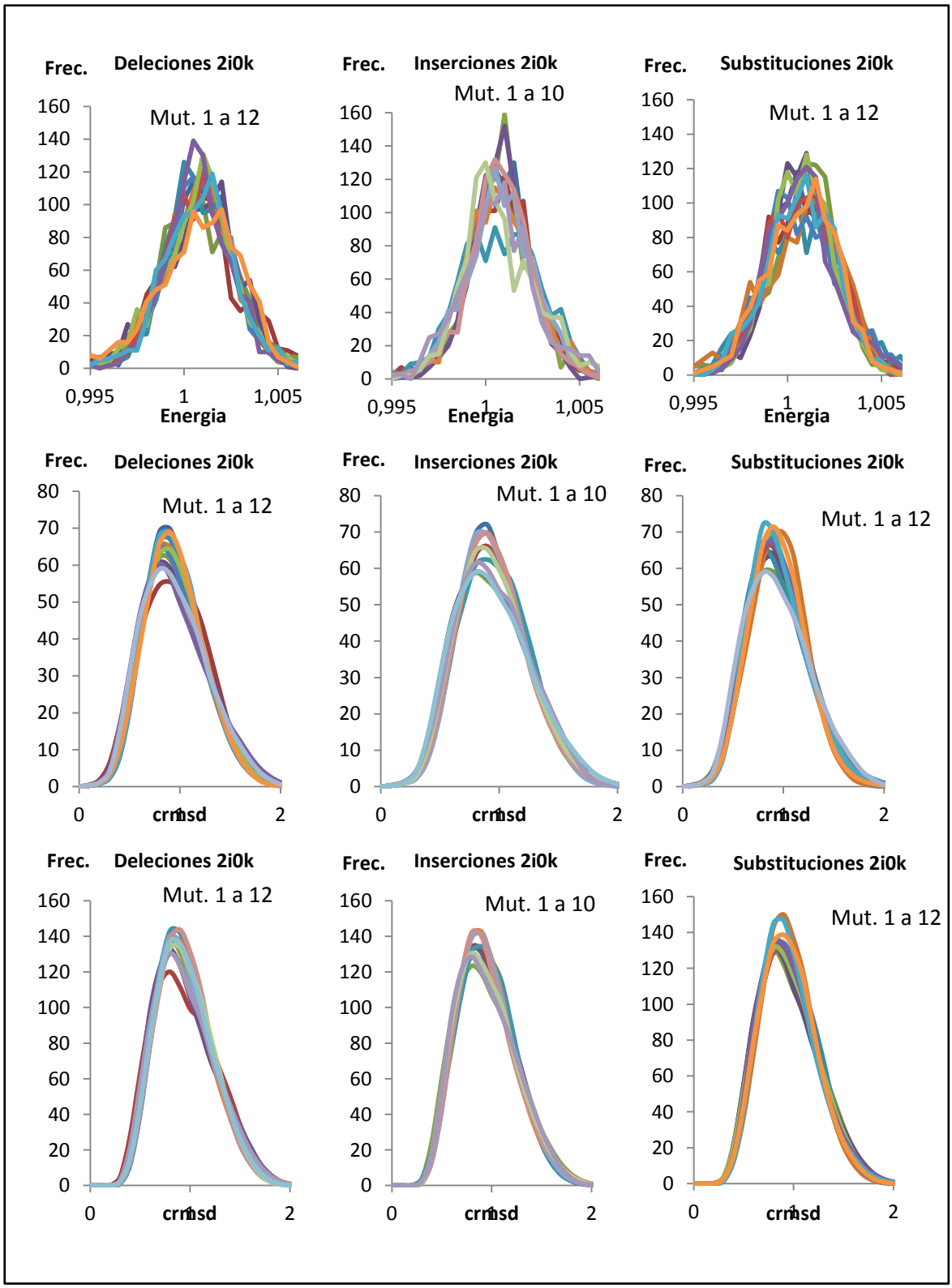


Fig. 3.11 Resultados 2i0k. Primera línea: Cálculo de Energías; segunda línea: Calculo crmsd Mutante-Mutante; tercera línea: Calculo crmsd Mutante-Nativa

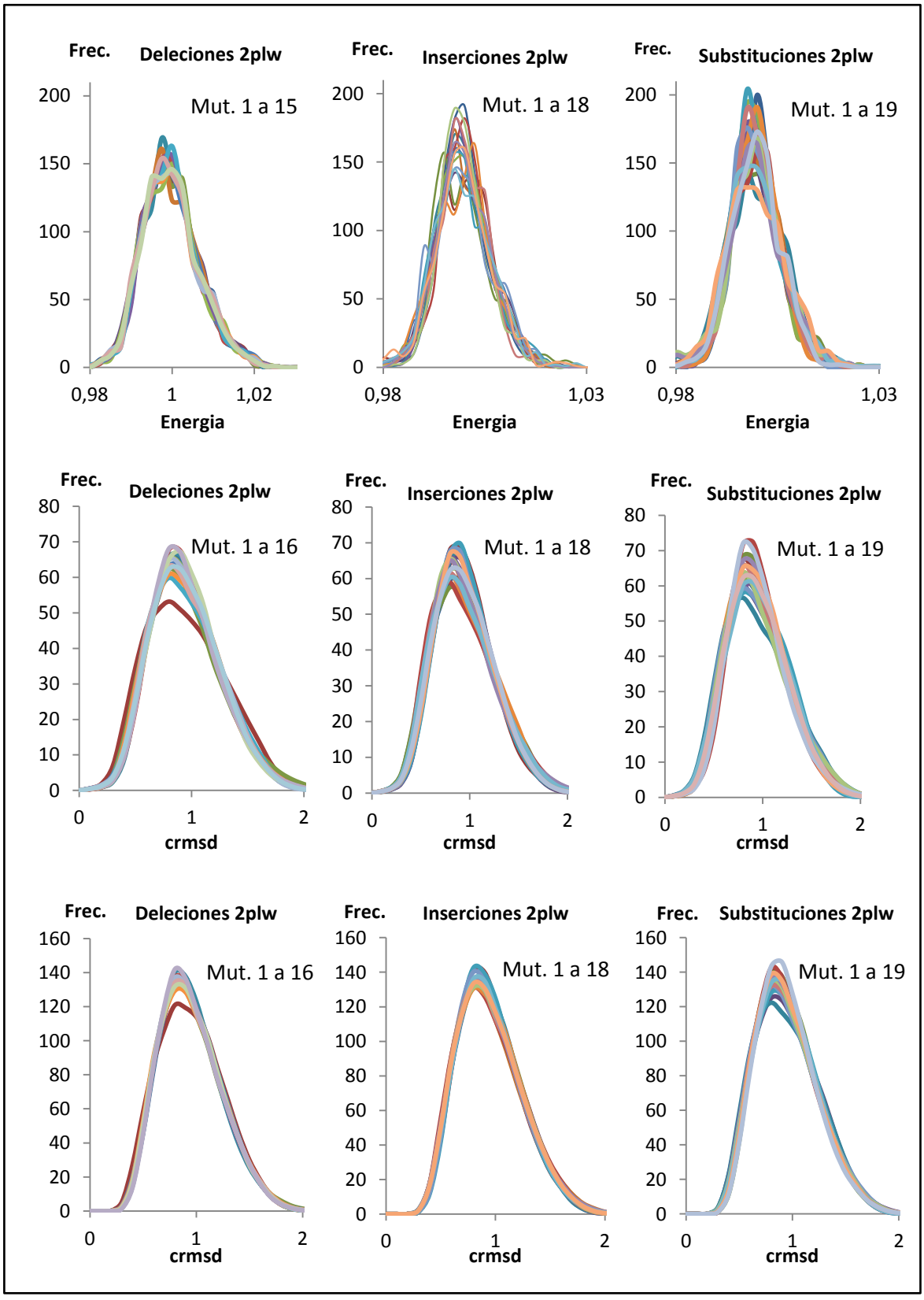


Fig. 3.12 Resultados 2plw. Primera línea: Cálculo de Energías; segunda línea: Calculo crmsd Mutante-Mutante; tercera línea: Calculo crmsd Mutante-Nativa

En general, para todas las LCRs vemos que las distribuciones de cada parámetro (energía, crmsd auto comparación y crmsd comparación con la nativa) globalmente tienen un aspecto similar, en forma de campana simétrica o asimétrica. Ello nos indica que en general, las LCR visitan con más frecuencia ciertos confórmeros que otros; en general, no observamos que este hecho cambie entre las LCRs nativas y las mutantes. Este es un carácter general, que observamos tanto en las curvas de energía como en los dos crmsd. Esta tendencia general debe, sin embargo modularse. En efecto, vemos que hay LCRs que presentan desviaciones apreciables para algunos de sus mutantes y para la función de energía. Por ejemplo, para la LCR de la proteína 2plw (Fig. 3.12), las curvas de energía presentan fluctuaciones claras en la altura de su máximo; un fenómeno que también se observa para 2bhu y para las otras, en mayor o menor medida. Este efecto contrasta con un comportamiento menos irregular para los crmsd, que tienen un aspecto no tan 'discontinuo'. Ello se explica por la naturaleza diferente de estos parámetros: mientras que la energía depende principalmente de las cadenas laterales, el crmsd se calcula para los átomos de la cadena principal. Así la energía recoge aspectos más sutiles de la estructura de las LCRs, mientras que los crmsd son más sensibles a los aspectos geométricos generales. De la comparación de estas gráficas, vemos que las LCRs mutantes no pueblan, generalmente, partes del espacio conformacional muy diferentes al de la nativa aunque, de forma local y alrededor de cada residuo, sí que puede haber diferencias destacables. Estas desviaciones reflejan la secuencia específica de cada LCR.

El objetivo principal de este estudio era ver si podíamos determinar la existencia de una relación entre cambios de secuencia y cambios estructurales significativos que nos pudieran dar una pista del porqué las mutaciones en las LCRs pueden dar lugar a efectos patogénicos [122]. Observamos que globalmente, el estado desordenado de las LCRs es muy robusto frente a los cambios de secuencia (Figs. 3.8 a 3.12). Este hecho sugiere que la variabilidad funcional introducida en las LCRs por las mutaciones podría ser debido a los cambios locales en las propiedades físico-químicas de la LCR, aspecto apoyado por las fluctuaciones mencionadas en las gráficas de energía, y por el hecho, obvio, de que las diferentes LCRs tienen diferente composición atómica.

4.- CONCLUSIONES

En primer lugar, hemos puesto a punto un protocolo para estudiar las propiedades estructurales generales de las LCRs desordenadas, y comprender el impacto que tienen sobre estas las variantes de secuencia más habituales (substituciones/inserciones/deleciones). En segundo lugar, hemos aplicado este protocolo a cinco LCRs, para las cuales hemos generado una elevada cantidad de variantes de secuencia. En general, el aspecto más destacado de estos resultados es que no hay diferencias substanciales en el comportamiento global de las LCRs, pero sí en aspectos locales que dependen de la secuencia específica de cada LCR.

CAPITULO 4: USO DE LA CORRELACIÓN EN EL ESTUDIO DE LAS TRAYECTORIAS DE DINÁMICA MOLECULAR

En este capítulo presentamos la segunda aplicación del programa presentado en esta tesis. Para esta aplicación escogimos un tema más fundamental que el de las LCR: establecer el efecto que tiene el volumen excluido (VE) de los átomos sobre las correlaciones que hay en sus movimientos. Como veremos en lo que sigue, este tema nos permite explotar la capacidad del programa para explorar amplios sectores del espacio conformacional de las proteínas de forma rápida, gracias a la representación simplificada que se utiliza.

1.- INTRODUCCIÓN

En el estudio de la dinámica de las biomoléculas a lo largo del tiempo se han utilizado las correlaciones interatómicas (ver sección 4.2 del capítulo de Introducción para la fórmula) como herramienta para identificar la existencia de movimientos colectivos de pares o de grupos de átomos [150]. Se ha visto que estas correlaciones, concretamente en el caso de las proteínas, son esenciales para determinar las características de su funcionalidad [151]. Ejemplos de ello son la transducción de señales alostérica [144-146] (mecanismo por el cual un estímulo que actúa en una zona de regulación de una proteína causa una respuesta conformacional en otra zona distante, provocando un movimiento coordinado entre partes de la proteína, fig. 4.1) o el transporte mecánico / termodinámico de la energía [149]. Adicionalmente, sabemos también que las correlaciones interatómicas dominan la parte entrópica de la función de energía de las proteínas [152]. En un plano más técnico, se ha establecido que la cuidadosa caracterización de los correlaciones de movimientos podría mejorar la interpretación de los experimentos de resolución de estructuras con resonancia magnética nuclear (NMR) y rayos X [142-143]. En resumen, en-

tender la naturaleza y los diferentes factores que pueden contribuir a las correlaciones entre átomos en la dinámica de las proteínas es un problema biológicamente y biofísicamente relevante.

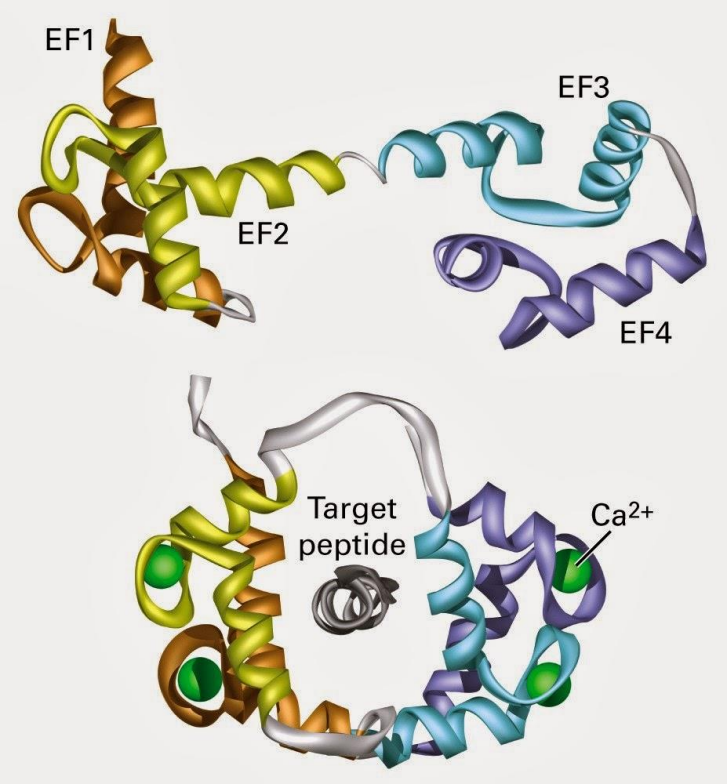


Fig. 4.1. Ejemplo de regulación alostérica: la calmodulina cambia su conformación estructural cuando se le unen cuatro Ca^{2+}. Una vez cambiada su conformación tiene la capacidad de unirse a otro péptido, generalmente CaM Kinasas, enzimas que necesitan unirse a la calmodulina para ser activas.

La aproximación habitualmente utilizada para el estudio de las correlaciones en proteínas es el uso de simulaciones de dinámica molecular (MD: Molecular Dynamics) [153-154]. Tal como hemos visto anteriormente, dichas simulaciones son técnicamente muy sofisticadas y computacionalmente costosas. Se basan en el uso de modelos atómicos que requieren, a nivel técnico, la discretización del movimiento en una serie de pasos sucesivos, cuya escala temporal es del orden de femtosegundos. Ello impide que las simulaciones exploren los eventos que tienen lugar en escala temporales que vayan más allá de los nanosegundos. Como consecuencia de esto, una simulación habitual de dinámica molecular no detectará aquellos mo-

vimientos y correlaciones de baja frecuencia que ocurran en una escala temporal mayor. Además, la correlación cruzada entre los desplazamientos de los diferentes átomos de una proteína dada no puede ser recogida de forma precisa [147]. En este contexto, pensamos que la aplicación de nuestro programa, que explora grandes volúmenes del espacio temporal de forma computacionalmente muy eficiente, puede aportar información adicional para la comprensión de la correlación entre residuos de las proteínas

La segunda aplicación del programa presentada en esta tesis se ha centrado en el estudio del impacto que tiene el volumen excluido [155] (fig. 4.3) en el valor de las correlaciones entre residuos. Como hemos apuntado, las correlaciones entre residuos se utilizan habitualmente en el estudio de las simulaciones de macromoléculas, para identificar efectos funcionales o estructurales relevantes [148]. Sin embargo, la complejidad de las estructuras moleculares, así como el elevado número de correlaciones que pueden definirse hacen que los valores de estas puedan verse afectados por efectos espurios, sin valor biológico real. En esta parte de la tesis, nos centramos en un factor cuyo efecto todavía no ha sido estudiado: el volumen excluido. Por volumen excluido nos referimos al hecho de que dos átomos no pueden ocupar simultáneamente la misma posición, o volumen, del espacio. Estrictamente hablando, este efecto introduce una correlación entre las posiciones de los átomos de la proteína, afectando directamente a la función de distribución de probabilidad de las distancias inter-residuales. Efectivamente, dicha distribución tiende a cero cuando la distancia entre átomos/residuos tiende a cero, en contra de lo que esperaríamos si no hubiese relación entre las posiciones de los átomos (se esperaría una distribución uniforme, con la misma probabilidad para cualquier distancia).

A nivel técnico, para cuantificar el impacto del volumen excluido sobre las correlaciones inter-residuo, hemos seguido una estrategia que se puede implementar fácilmente con nuestro programa:

1- hemos variado sistemáticamente el radio de los pseudo-átomos Ca, entre 0 y 3.8 Å
2- Para cada valor, hemos generado 1000 modelos;
3- Hemos calculado la correlación entre todos los pares de residuos posibles.

Este estudio se ha realizado para 200 estructuras de longitudes comprendidas entre los 100 y los 300 aminoácidos, a fin de tomar en consideración el efecto separación de secuencia entre residuos.

2.- MATERIALES Y MÉTODOS

2.1 Selección de estructuras para el estudio

Para construir el conjunto de proteínas empleadas en el estudio, escogimos al azar una serie de proteínas con estructura determinada experimentalmente, que cumpliesen con las siguientes características:

- Resolución (medida de calidad estructural estándar [156]): entre 0 y 3 angstroms
- R-value (medida de calidad estructural estándar [157]): 0.3
- Longitud cadena: entre 100 y 300
- Método experimental: difracción de rayos X

Del conjunto resultante de estructuras, seleccionamos 200 de ellas, distribuidas de la siguiente manera: 50 estructuras de longitud entre 101-150, 50 entre 151-200, 50 entre 201-250 y 50 entre 251-300. En la tabla 4.1 se proporciona la lista final de estructuras seleccionadas.

101-150 residuos	151-200 residuos	201-250 residuos	251-300 residuos
1AYOA	1AEPA	1EUVA	1EG4A
1B3TA	1DVOA	1JSSA	1FS2A
1GAKA	1FJRA	1K0MA	1FTRA
1GMUA	1GPRA	1K3YA	1G6HA
1GU2A	1JHSA	1OTKA	1HI9A
1HUFA	1LQVA	1RP3A	1K5NA
1IFRA	1O6DA	1SX5A	1KI0A
1IJYA	1QFTA	1XTTA	1NZJA
1J24A	1T4WA	2FB5A	1O9IA
1LKKA	1WDJA	2IIEA	1RWIA
1MG4A	2B99A	2IZWA	1UEKA
1MNMA	2FJRA	2PETA	1WLGA

1OO0A	2GDMA	2YPHA	1XQOA
1P57A	2O70A	3AOTA	1YT5A
1PSRA	2QDLA	3BGYA	2A14A
1S5UA	2W68A	3BPZA	2GB4A
1SLUA	2XI7A	3C5WA	2GTRA
1TU1A	2YOPA	3C6AA	2H26A
1VBVA	2Z0XA	3ETOA	2P26A
1XIWA	3DI2A	3SHGA	2RBKA
1XS0A	3FFVA	3UMHA	2X9KA
1YOZA	3FLDA	3W1DA	2XJ4A
2ASKA	3H6RA	3ZMLA	2Y3CA
2D59A	3HA4A	4BP0A	3DBYA
2E9XA	3OG6A	4BS6A	3DHAA
2F5GA	3QHPA	4BVXA	3GNEA
2G64A	3RGQA	4C47A	3R6UA
2GRCA	3RY4A	4C55A	3RKGA
2GU9A	3TOWA	4CT7A	3T9WA
2GUDA	3UB6A	4DOXA	3TCAA
2H1CA	3US6A	4F01A	3TDSA
2HY5A	4A7UA	4GJZA	3WK2A
2IGPA	4BTHA	4INKA	3ZNYA
2IP6A	4C0NA	4J3MA	4BKUA
2NPTA	4CVPA	4JGIA	4CS5A
2NSZA	4FDKA	4JXHA	4D1UA
2NUHA	4GFXA	4JZ5A	4E4WA
2OXOA	4GHTA	4KNKA	4HNOA
2PIEA	4H9WA	4LDZA	4HV3A
2PK8A	4IBQA	4LW5A	4ILLA
2QFEA	4IJTA	4M6GA	4JJPA
2VZCA	4JNBA	4M9CA	4JMWA
2WJ5A	4KDZA	4N67A	4JQXA
2XPPA	4MBUA	4N9JA	4K2AA
2YXYA	4MZIA	4NQTA	4KNBA
2ZHPA	4N1FA	4NUIA	4LAXA
3C5KA	4N1UA	4NUJA	4LO0A
3CTRA	4NAVA	4O6JA	4M76A
3D7AA	4O0NA	4OMVA	4MXDA
3DM3A	4O96A	4QMHA	4O46A

Tabla 4.1. Listado de las 200 estructuras utilizadas para el estudio, clasificadas según el número de residuos.

2.2 Protocolo de simulación

Para analizar el impacto del volumen excluido, para cada proteína de nuestro conjunto (Tabla 4.1) obtenemos una colección de simulaciones libres utilizando nuestro programa (Fig. 4.2). En estas simulaciones no se impone ninguna restricción geométrica, ni estructural, excepto la distancia de contacto mínima permitida (Fig. 4.3) que es la que determina el volumen excluido. Los valores de seleccionados para dicha distancia fueron: 0/1/2/3/3.2/3.4/3.6/3.8. El último valor corresponde a valor mínimo permitido entre C_α. Utilizando nuestro programa, hicimos una simulación (por proteína) de 1000 modelos estructurales para cada valor de la distancia de contacto.

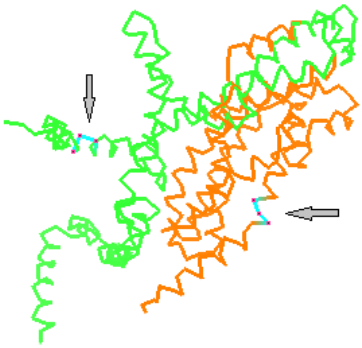


Fig. 4.2. Ejemplo de dos conformaciones de la cadena A de la proteína 1RP3 utilizada en el estudio. En color azul podemos ver una tripleta de átomos Cα para cada conformación correspondiente a la misma secuencia

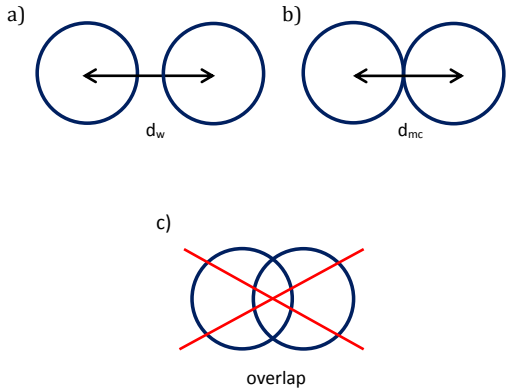


Fig. 4.3. Representación simplificada de situaciones de contacto entre átomos C_α. a) corresponde a la situación más normal donde los átomos están separados a una distancia d_w. B) Situación correspondiente a la distancia mínima de contacto. C) Situación prohibida físicamente en la que se superponen las esferas de contacto.

2.3 Cálculo de la correlación

Utilizamos los 1000 modelos de cada una de las simulaciones para calcular la correlación entre todos los pares de residuos a partir de la matriz de covarianza normalizada según se detalla a continuación [158]:

- Supongamos n conformaciones de una molécula M: M(1) .... M(n).

- Definimos el vector posición de los átomos i y j en la t-ésima conformación de la molécula M: $r_i(t)$, $r_j(t)$.

- El elemento de la matriz de covarianza c_{ij} se define como:

$$[\boldsymbol{A}] \qquad \boldsymbol{c_{ij}} = \langle (\boldsymbol{r_i} - \langle \boldsymbol{r_i} \rangle)(\boldsymbol{r_j} - \langle \boldsymbol{r_j} \rangle) \rangle = \langle \boldsymbol{r_i r_j} \rangle - \langle \boldsymbol{r_i} \rangle \langle \boldsymbol{r_j} \rangle$$

$$= \frac{\boldsymbol{1}}{\boldsymbol{n}} \left[\sum_{\boldsymbol{t=1}}^{\boldsymbol{n}} \boldsymbol{r_i(t) r_j(t)} - \frac{\boldsymbol{1}}{\boldsymbol{n}} \left(\sum_{\boldsymbol{t=1}}^{\boldsymbol{n}} \boldsymbol{r_i(t)} \right) \left(\sum_{\boldsymbol{t=1}}^{\boldsymbol{n}} \boldsymbol{r_j(t)} \right) \right]$$

El estudio se realiza utilizando matrices de covarianza normalizada. Cada elemento de la matriz C_{ij} se define como:

$$[\boldsymbol{B}] \qquad \boldsymbol{C_{ij}} = \frac{\boldsymbol{c_{ij}}}{\boldsymbol{c_{ii}^{1/2} c_{jj}^{1/2}}} = \frac{\langle \boldsymbol{r_i r_j} \rangle - \langle \boldsymbol{r_i} \rangle \langle \boldsymbol{r_j} \rangle}{\left[\left(\langle \boldsymbol{r_i^2} \rangle - \langle \boldsymbol{r_i} \rangle^2 \right) \left(\langle \boldsymbol{r_j^2} \rangle - \langle \boldsymbol{r_j} \rangle^2 \right) \right]}$$

2.4 Cálculo de la distribución de distancias interresiduo

Las distribuciones de distancias interresiduo corresponden a la distribución radial que se utiliza habitualmente en termodinámica estadística para estudiar las correlaciones entre las posiciones de diferentes partículas [159-161]. Para ello se calcularon las distancias entre pares de átomos Cα de cada proteína y se realizó en mismo cálculo para cada uno de los 1000 modelos de cada proteína. Posteriormente se realizó una representación gráfica de la frecuencia de las distancias calculadas para cada proteína y sus respectivos modelos. A fin de tener en cuenta las limi-

taciones impuestas por la estructura covalente de la proteína, tratamos por separado los pares de residuos situados a diferentes 'distancias' en la secuencia (Fig. 4.4). En este caso, la 'distancia' de secuencia entre dos residuos, corresponde al número de amino ácidos que hay entre ellos. Se introdujeron las siguientes categorías:

- Distancia entre secuencia <=3
- Distancia entre secuencia <=4
- Distancia entre secuencia <=5
- Distancia entre secuencia entre 6 y 10
- Distancia entre secuencia entre 11 y 20
- Distancia entre secuencia >= 21

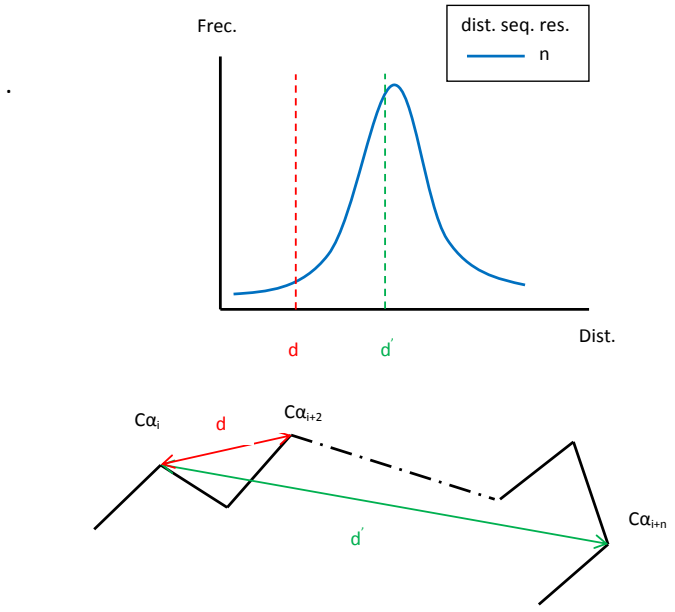


Fig. 4.4. Representación esquemática del cálculo de distancias entre Cα con una separación máxima en secuencia determinada

3.- RESULTADOS

Para este estudio calculamos dos descriptores que reflejan el efecto del vo

lumen excluido sobre la correlación entre las posiciones de los átomos: las correlaciones entre movimientos atómicos y la distribución de distancias entre residuos.

3.1 Correlaciones entre movimientos atómicos

Para cada simulación, se obtuvieron las matrices de correlación según la ecuación [B]. En la figura 4.5 representamos su distribución respecto a la distancia mínima de contacto. Recordemos que estas distancias cubren un rango de valores comprendido entre 0 y 3.8 angstroms, y que los valores mayores están asociados a un mayor efecto de volumen excluido.

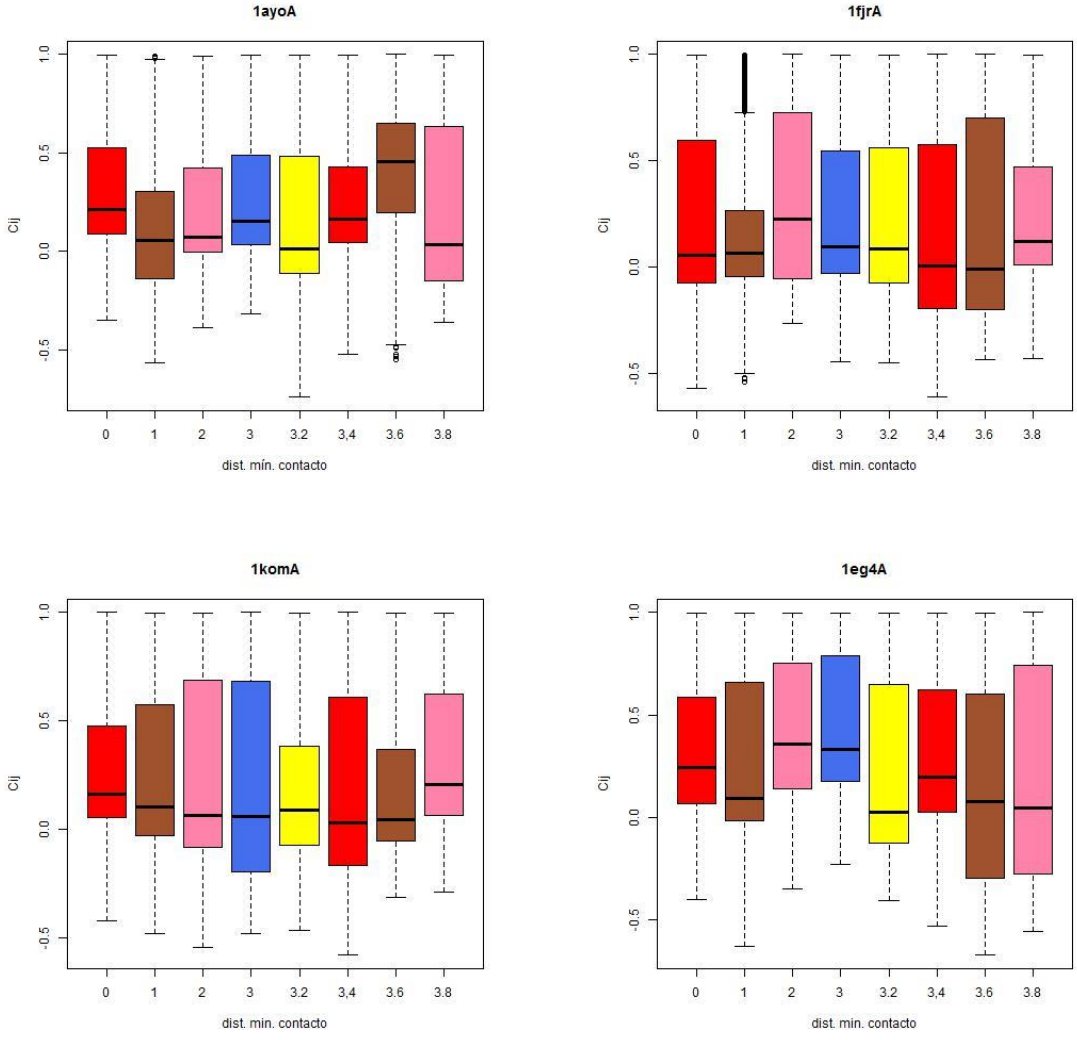


Fig. 4.5 Distribuciones de la correlación C_{ij} entre residuos, representadas en función de la distancia mínima de contacto. Las gráficas corresponden a cuatro proteínas representativas de cuatro rangos de tamaño diferentes: 1ayoA (grupo tamaño 101-150), 1fjrA (grupo tamaño 151-200), 1komA (grupo tamaño 201-250), 1eg4A (grupo tamaño 251-300).

En la figura 4.5 no se aprecia ninguna tendencia conspicua que nos permita establecer un vínculo entre volumen excluido y valores de la correlación; de hecho, para las cuatro proteínas vemos que la distribución de valores crece y decrece sin una relación aparente con la distancia mínima de contacto. Este resultado es parcialmente debido a que el rango de distancias interresiduo es muy amplio, e.g se dan distancias de 10, 20, 30, 40 o más angstroms y hay un gran número de conformaciones para las que se dan estas distancias, y en diferentes orientaciones relativas de los residuos. Frente a ello, el rango de distancias en el cual se manifiesta el volumen excluido es pequeño, con menos conformaciones posibles asociadas a él. Por lo tanto, este término tendrá un peso menor en el cálculo de las correlaciones, y su efecto queda diluido. Adicionalmente, pensamos que hay un efecto técnico debido al promediado implícito en el cálculo de la correlación (que se obtiene a partir de una nube heterogénea de estructuras, fig. 4.2), que puede distorsionar aspectos estructurales de la proteína, afectando parcialmente al cálculo de la correlación.

3.2 Distribución de distancias interresiduo

Para obtener un punto de vista diferente, libre de los problemas anteriores, decidimos trabajar con la distribución de distancias interresiduo. Para ver el efecto del volumen excluido (VE), utilizamos como referencia la curva obtenida para la distancia mínima de contacto igual a cero, ya que es el único caso en el que no se espera ver efecto VE. Para otras distancias mínimas de contacto esperamos ver un efecto VE creciente.

En la figura 4.6 mostramos los resultados de las distancias interresiduo sin normalizar, y en la figura 4.7 mostramos los resultados ya normalizados.

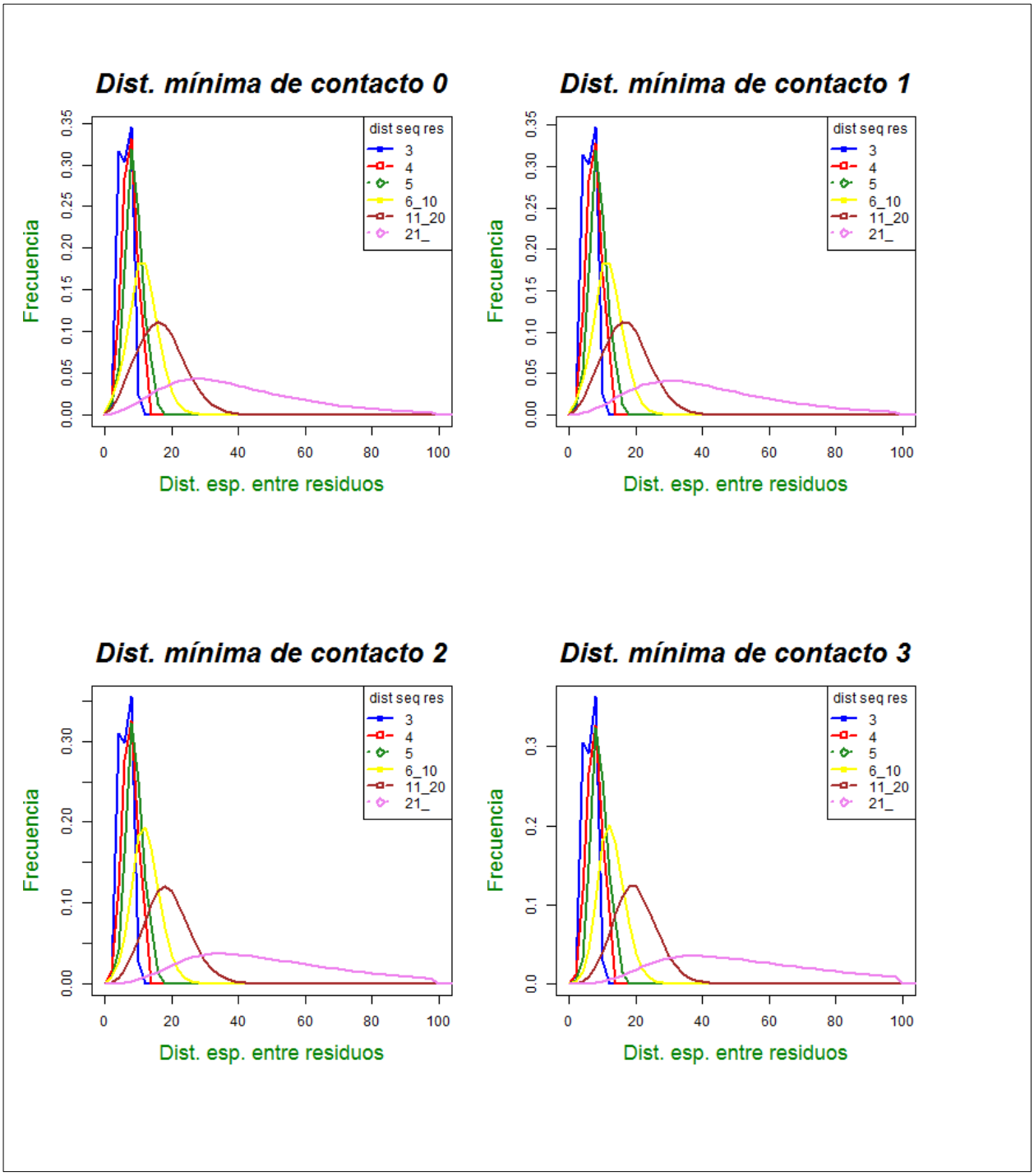
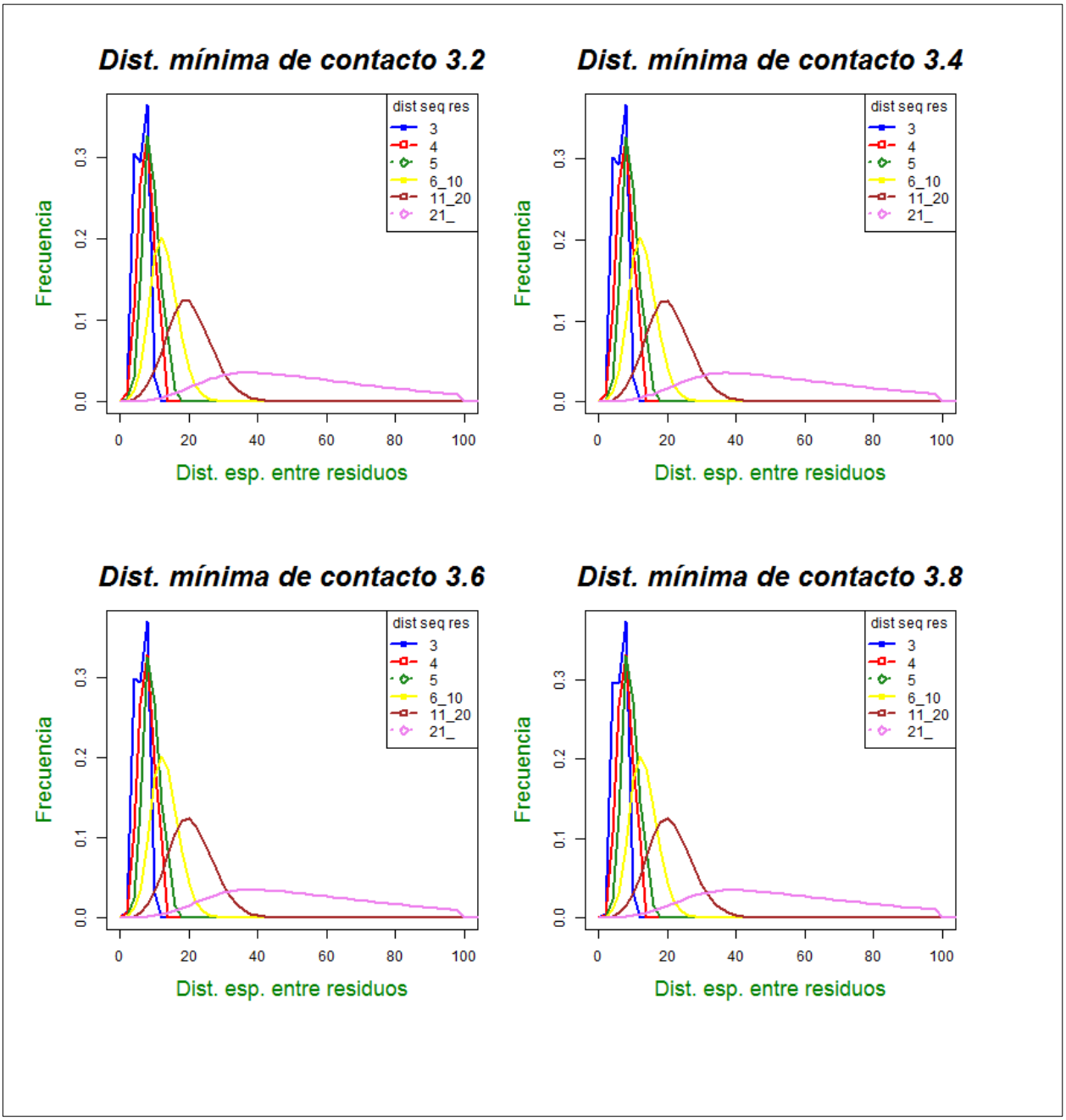


Fig. 4.6. Función de distribución de las distancias espaciales, organizada por separación entre residuos a lo largo de la cadena (dist.seq.res). Como vemos, presenta el comportamiento esperado, tendiendo a cero cuando la distancia tiende a cero, o bien cuando alcanza el máximo que permite la integridad de la estructura covalente. (Continua en página siguiente).



Cuando normalizamos las distribuciones respecto al caso en el que no hay efecto VE (Fig. 4.7), vemos que a medida que aumenta dicho efecto (para valores crecientes de distancias de contacto mínimas) se pueblan cada vez más las distancias grandes, y simultáneamente se deprime la población en las distancias cortas, más próximas a cero. Ello muestra la existencia de una desviación respecto al caso de no correlación (distancia mínima de contacto=0) que aumenta según crece la distancia de contacto. Este comportamiento indica que incluso en ausencia de in-

teracciones entre residuos de una misma cadena, existe una correlación que es únicamente debida al hecho de que dos, o más, átomos no pueden ocupar simultáneamente el mismo lugar. Esta correlación, es además mayor para aquellos pares de residuos que están a grandes distancias. Este aspecto es interesante, ya que normalmente en los estudios de correlación no se incluyen correcciones por efecto de distancia espacial.

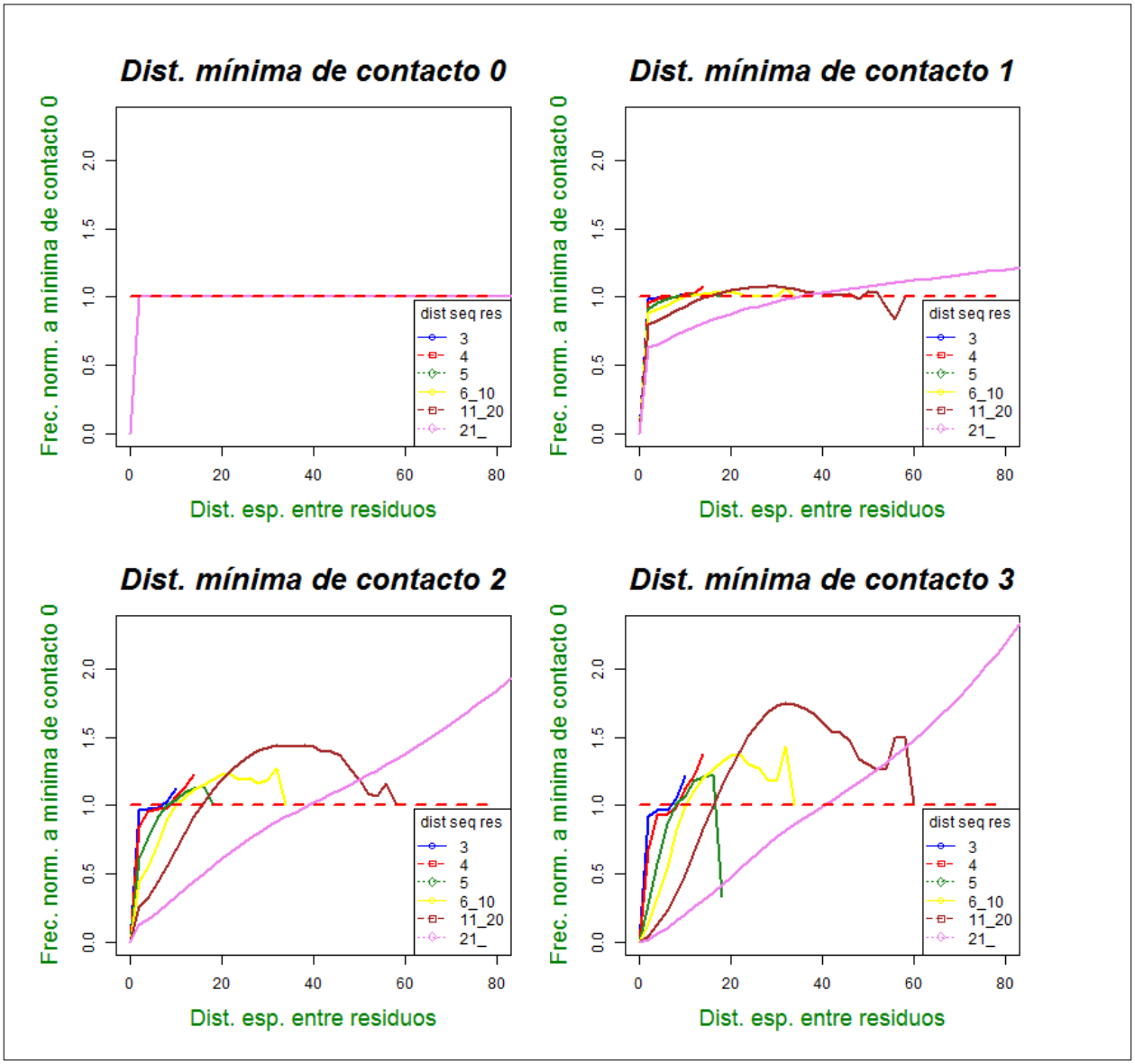
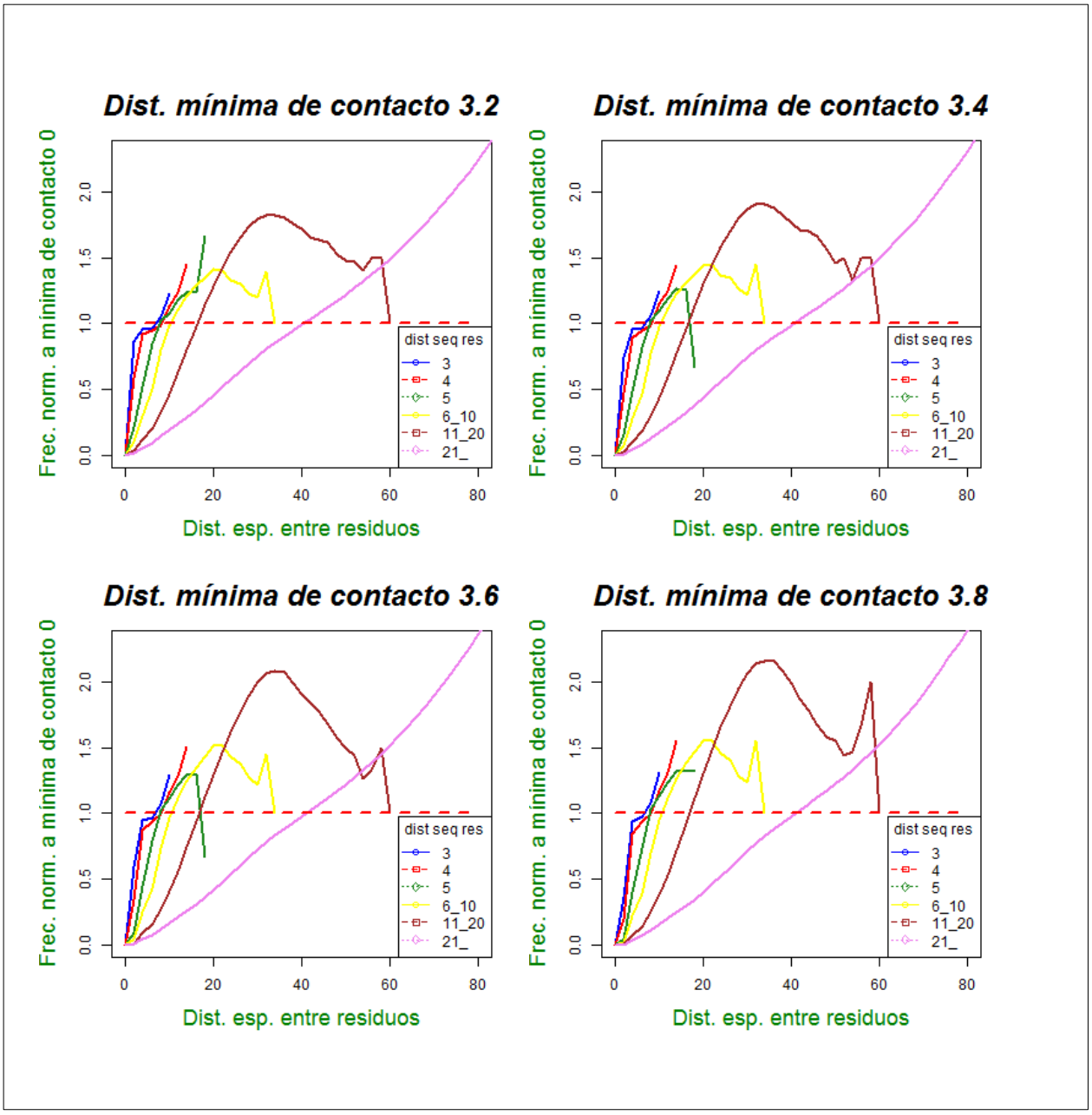


Fig. 4.7 Funciones de distribución de las distancias interresiduo, normalizadas utilizando la distribución correspondiente distancia mínima de contacto 0 y organizadas por separación a lo largo de la cadena (dist.seq.res). Vemos como las distancias geométricas elevadas entre residuos aumentan en frecuencia a medida que incrementamos la distancia mínima de contacto. (Continua en pg. Siguiente)


Dist. mínima de contacto 3.2
Dist. mínima de contacto 3.4
Dist. mínima de contacto 3.6
Dist. mínima de contacto 3.8
Frec. norm. a mínima de contacto 0
Dist. esp. entre residuos
dist seq res
3
4
5
6_10
11_20
21_
0.0
0.5
1.0
1.5
2.0
0
20
40
60
80

4.- CONCLUSIONES

El uso de nuestro programa, y concretamente la facilidad con la que se pueden realizar simulaciones con diferentes parámetros atómicos, nos ha permitido identificar y cuantificar la existencia de un término en las correlaciones interresiduo que se debe al efecto del volumen excluido. Este efecto es de origen meramente estructural, y no tiene ningún valor funcional, por lo tanto debe de tenerse en cuenta en los análisis de simulaciones moleculares en los que se utiliza sistemáticamente la correlación interresiduo para describir efectos funcionales.

CAPITULO 5: DISCUSION Y CONCLUSIONES

1.- DISCUSIÓN

En esta tesis se describe el desarrollo de un programa para la simulación de la estructura de las proteínas y su aplicación a dos problemas biológicos de interés. El programa está basado en el uso de una representación simplificada de las proteínas desarrollada en el grupo del Dr. Xavier de la Cruz [101]. Científicamente, este trabajo se sitúa en el campo del análisis estructural orientado a responder preguntas biológicamente relevantes [172]. Este campo es realmente complejo y las aproximaciones disponibles, tanto experimentales [94,95,57] como computacionales [62,65] tienen todavía limitaciones claras. A nivel experimental, el estudio estructural se ve dificultado por diferentes problemas técnicos; por ejemplo, la obtención de cristales adecuados para la difracción estructural [59], el límite en el tamaño de la proteína o complejos estudiados [170], la dificultad de analizar los aspectos dinámicos de la estructura [171], etc. Ello ha llevado a la creación de modelos simplificados de baja resolución o 'coarse-grained' [5,80] que nos permiten obtener simulaciones de comportamiento y estructura a escalas de tiempo cada vez mayores, a un coste computacional asumible, aunque sea con un menor nivel de detalle. En este sentido, es importante tener en cuenta que es precisamente el balance entre detalle y objetivo científico el que debe determinar el uso de este tipo de modelo.

Como hemos visto en la Introducción, existen varias formas de simplificar la estructura de las proteínas [10,25-30,70,96]. La aproximación presentada en esta tesis se basa en utilizar un único átomo por residuo, el carbono Cα, definiendo ángulos pseudo-torsionales, que confieren al modelo de la proteína una flexibilidad estructural análoga a la de una proteína real, y que de forma implícita recogen parte de las propiedades estéricas asociadas a cada amino ácido. A continuación se discute más en detalle la aportación concreta de este trabajo, siguiendo el orden de presentación de los capítulos.

1. El programa

Como se ha visto anteriormente, el primer paso de esta tesis fue el desarrollo de un programa que implementase nuestro modelo de la estructura de las proteínas y permitiese explorar, de forma computacionalmente eficiente, el espacio conformacional de estas.

Como se ha descrito en el capítulo 2, el programa utiliza el algoritmo de Metropolis para ir encontrando nuevas conformaciones, partiendo de una proteína de referencia. Nos decantamos por el algoritmo de Metropolis porque ha mostrado su valor en programas de simulación similares. Un aspecto interesante de nuestro trabajo, es que nos permite definir varias funciones de energía (ver capítulo 2), que permiten analizar diferentes aspectos estructurales de los problemas considerados por el usuario. Por ejemplo, podemos representar el espacio conformacional de los estados desnaturalizados combinando un término de quiralidad (chirmsd) y otro del radio de giro (rdgrmsd); o bien, permitir fluctuaciones conformacionales respecto a una estructura promedio restringiendo las distancias de contacto entre residuos; etc. Una ventaja del programa de cara al usuario, es que las diferentes funciones de energía se pueden utilizar de forma individual o en combinación, dándoles un peso específico a cada una de ellas.

De la misma manera que existe una cierta flexibilidad al escoger la función de energía, también se han definido tipos de movimientos consistentes con la estructura local de las proteínas, que permiten explorar grandes volúmenes de espacio conformacional natural, no artificial como en el caso de otros modelos. Los tipos de movimientos implementados nos permiten modificar las conformaciones tanto de forma global, como de forma local, lo que facilita el estudio de diferentes problemas. Por ejemplo, hemos analizado los movimientos globales para comprender el rol del volumen excluido en el cálculo de las correlaciones, y los movimientos locales para estudiar el espacio conformacional de las LCRs.

Dentro de la clasificación de los modelos simplificados expuesta en el capítulo 2, podemos considerar que el modelo que hemos utilizado pertenece a la categoría de modelos basados en la estructura. Este tipo de modelos han sido útiles para investigar escenarios de plegamiento y movimientos funcionales de proteínas

[19-21] aunque en nuestro caso lo hemos utilizados como una herramienta para explorar el espacio conformacional de la estructura de la proteína. Una ventaja respecto a los modelos basados en mecánica molecular [69.70] es su rapidez de ejecución y la capacidad para explorar una gran cantidad de conformaciones en tiempos comparativamente menores. Ello queda demostrado por la enorme cantidad de simulaciones realizadas en el capítulo 3. Por el contrario, la falta de descripción de las interacciones fisicoquímicas limita el tipo de problema estudiado a aquellos casos en los que no se requiere una ponderación energética de las estructuras generadas.

Un aspecto positivo de nuestro modelo, si lo comparamos con otros modelos basados en estructura como las redes cúbicas [108-110], su capacidad para generar estructuras secundarias plausibles, lo que hace que sea más eficaz para las conformaciones generadas.

Por todo ello, consideramos la validez de nuestro modelo para aplicaciones en las que se requiera explorar el espacio conformacional de la proteína con un gran número de conformaciones a generar, y poder deducir propiedades estadísticas de las conformaciones. De entre las aplicaciones posibles, escogimos dos que cumplen estas condiciones: el estudio del espacio conformacional de las LCR y el análisis de las correlaciones

2. Las LCR y sus variantes

Nuestra primera aplicación se encaminó a esclarecer un problema biológico de interés: el impacto de las mutaciones sobre el espacio conformacional de las LCRs. Como se explica en el capítulo 3, las regiones de baja complejidad o LCRs son zonas de la secuencia de la proteína especialmente ricas en uno o unos pocos aminoácidos. Estas regiones se caracterizan en muchos casos por ser altamente inestables, o estructuralmente desordenadas, habiéndose asociado su expansión (generalmente cuando corresponden a motivos pequeño de secuencia) a diferentes tipos de enfermedades humanas [122]. Por ejemplo, un grupo de factores de transcripción que contienen motivos en secuencia ricos en Glutamina (GLN) han sido implicados en enfermedades de poliglutamina [162]. También se ha visto que las proteínas humanas tienen numerosos tramos de regiones de muy baja complejidad

(muchos de ellos enriquecidos en GLN, SER o residuos ácidos) que están relacionadas con enfermedades neurodegenerativas y cáncer [163]. En general, se han descrito varias propiedades de las regiones desordenadas que contribuirían a la aparición de enfermedades: 1- su habilidad para plegarse al entrar en contacto con otra proteína o ácido nucleico [165], 2- su regulación funcional vía splicing alternativo o vía modificaciones postraduccionales [166,167] y 3- su plasticidad y promiscuidad a la hora de formar interacciones[164] (una región desestructurada puede unirse a diferentes moléculas).

Por todo lo anterior, resulta hasta cierto punto paradójica la abundancia de las LCRs, dado el potencial daño que pueden causar. Para explicar esta situación, se ha postulado que las LCRs tendrían un valor adaptativo, ya que mutaciones en su secuencia podrían generar fácilmente nuevas estructuras/funcionalidades de las proteínas, con valor adaptativo [119]. En este contexto, decidimos aplicar nuestro programa para estudiar, de forma sistemática, el impacto de las mutaciones en el espacio conformacional de las LCRs. Los resultados de este estudio nos muestran que los mutantes de las LCRs, aunque tienen un comportamiento similar al de las LCR nativas, pueden efectivamente explorar nuevas regiones del espacio conformacional. Ello apoyaría la idea de que las variantes de secuencia de las LCRs tienen el potencial de crear novedades estructurales, de acuerdo con la hipótesis de (Green et al. [119]). También detectamos algunas diferencias en los términos de energías, atribuibles a la dependencia de estos de la composición de la cadena lateral. Este último aspecto es relevante desde el punto de vista funcional, ya que la energética de estas regiones puede afectar a su capacidad de unión con otras proteínas.

Nuestro estudio sugiere que las mutaciones en las LCRs podrían inducir cambios estructurales locales más que generales, siendo este tipo de cambios los que afectarían a su funcionalidad. Adicionalmente, este trabajo confirma la utilidad de nuestro programa en el estudio de problemas estructurales de interés biológico, sobre todo de aquellos en los que prima la exploración global del espacio conformacional sobre el estudio detallado de un pequeño número de estructuras.

3. El efecto del volumen excluido sobre las correlaciones interresiduo

Finalmente, hemos querido mostrar que nuestro programa también permite abordar problemas técnicos del campo de la bioinformática estructural. Para ello, lo aplicamos al estudio del efecto del volumen excluido sobre las correlaciones entre residuos. Este es un problema interesante, porque las correlaciones son una herramienta que se utiliza habitualmente en el análisis de las simulaciones de macromoléculas [168,169], con la finalidad de identificar movimientos colectivos asociados a la función de las proteínas. De hecho, la correlación de movimientos entre diferentes partes de la proteína puede mediar procesos bioquímicos fundamentales como la transducción de señales y el alosterismo [144-146]. Estos procesos son estudiados frecuentemente mediante el uso de simulaciones de dinámica molecular, combinadas con el análisis de correlaciones. Para este último, se calcula la matriz de covarianzas o más recientemente la correlación generalizada [169]. La correcta utilización de estas herramientas es fundamental para la interpretación biológica de los resultados computacionales, siendo particularmente importante la identificación de los posibles efectos de origen técnico, ajenos a la función biológica que queremos estudiar. En este contexto, la identificación de un posible artefacto en el cálculo de estas correlaciones es relevante, ya que puede ayudar a interpretar más correctamente la información que contienen.

En el estudio, aprovechamos la capacidad de nuestro programa para generar rápidamente grandes cantidades de conformaciones, gracias a su representación simplificada de la proteína. De hecho, generamos 1000 modelos para una muestra de 200 proteínas del PDB. Sobre estos modelos, realizamos dos tipos de cálculos de correlación: utilizando las coordenadas cartesianas de los átomos y utilizando la distribución de distancias interresiduo. En el primer caso, no apreciamos ninguna relación entre los valores de la correlación y el volumen excluido. Sin embargo, sí que detectamos un efecto del volumen excluido al estudiar la distribución de distancias interresiduo. Efectivamente, para estas distribuciones vemos como, a medida que aumentamos el efecto de volumen excluido, las distancias interresiduo grandes tienden a poblarse más que las pequeñas, mostrando así un efecto de correlación. En base a este resultado, pensamos que sería interesante utilizar alguna medida complementaría a la correlación convencional, como el uso de funciones de distribución radial, en el análisis de problemas biológicos, para evitar la inclusión de efectos exclusivamente de origen técnico.

2.- CONCLUSIONES

En la presente tesis presentamos un programa para la simulación rápida de estructuras de proteínas, y su aplicación a dos problemas de interés biológico y técnico en bioinformática estructural. Las principales conclusiones de este trabajo son:

1.- El programa desarrollado presenta una implementación efectiva de la representación de las proteínas basada en los Cα.

2.- Este programa nos permite generar una gran cantidad de conformaciones en un tiempo reducido.

3.- La rapidez y flexibilidad en su parametrización permiten su aplicación a una gran variedad de problemas en los que se requiere explorar el espacio conformacional de las estructuras de proteínas.

4.- En el estudio de las LCRs hemos desarrollado un protocolo específico para el análisis de sus mutaciones con nuestro modelo

5.- El espacio conformacional que exploran los mutantes de las LCRs no es esencialmente muy diferente del que exploran las LCRs nativas

6.- Las funciones biológicas que se observan en estos mutantes no serían debidas a cambios globales en su estructura general sino a cambios locales en esta.

7.- La forma convencional de analizar las correlaciones entre residuos no permite detectar el efecto del volumen excluido sobre estas.

8- El uso de la función de distribución radial permite detectar la existencia de un efecto debido al volumen excluido.

APENDICE I – GRAFICAS EN DETALLE DE RESULTADOS DEL CAP. 3

En este apéndice se proporciona una versión más detallada de los resultados correspondientes a las figuras 3.8 a 3.12 del capítulo 3, así como unas tablas explicativa de las leyendas. En estas tablas se pueden identificar los cambios de secuencia asociados a cada mutante.

Resultados 1W1O

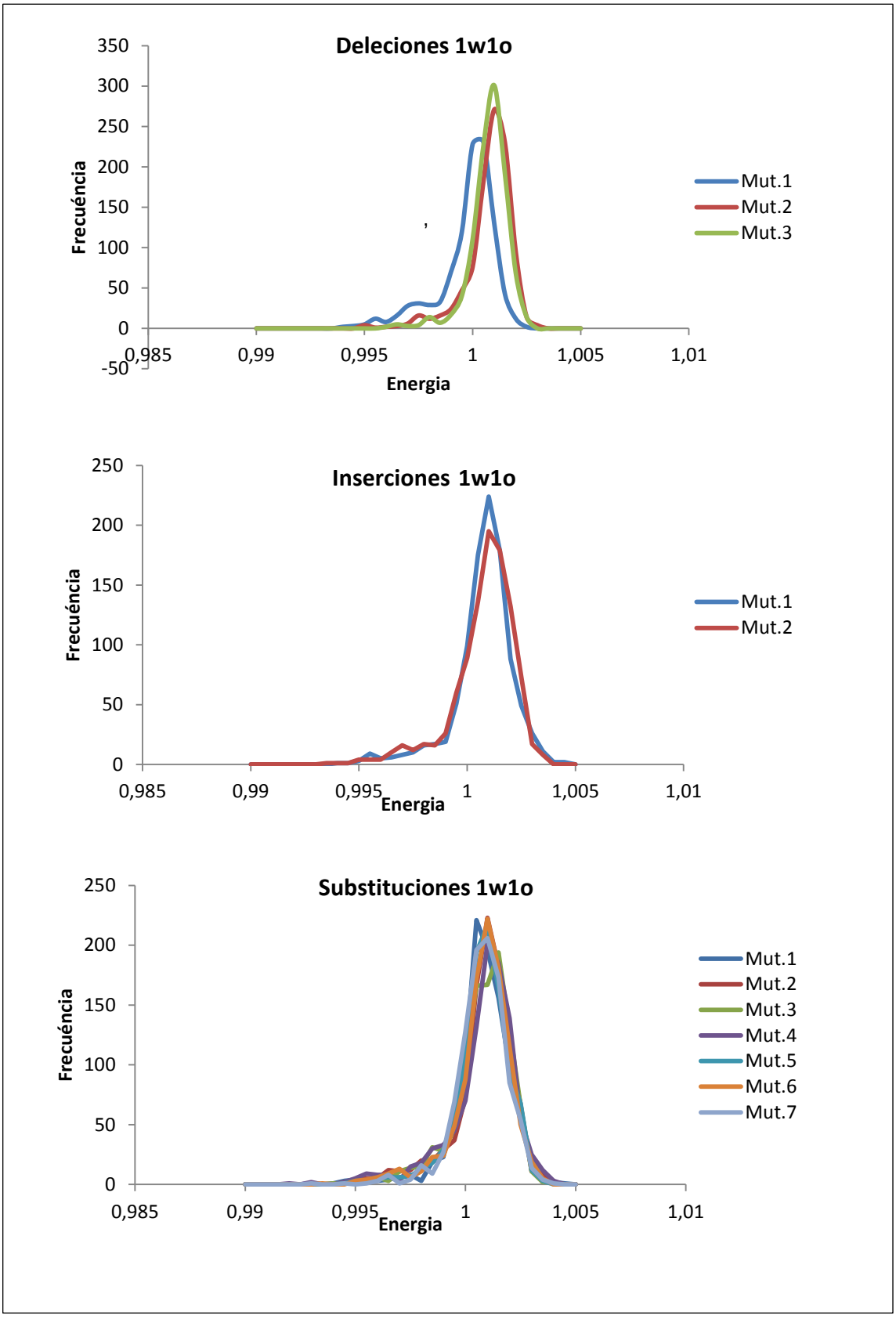


Fig. I.1 Energías 1w1o

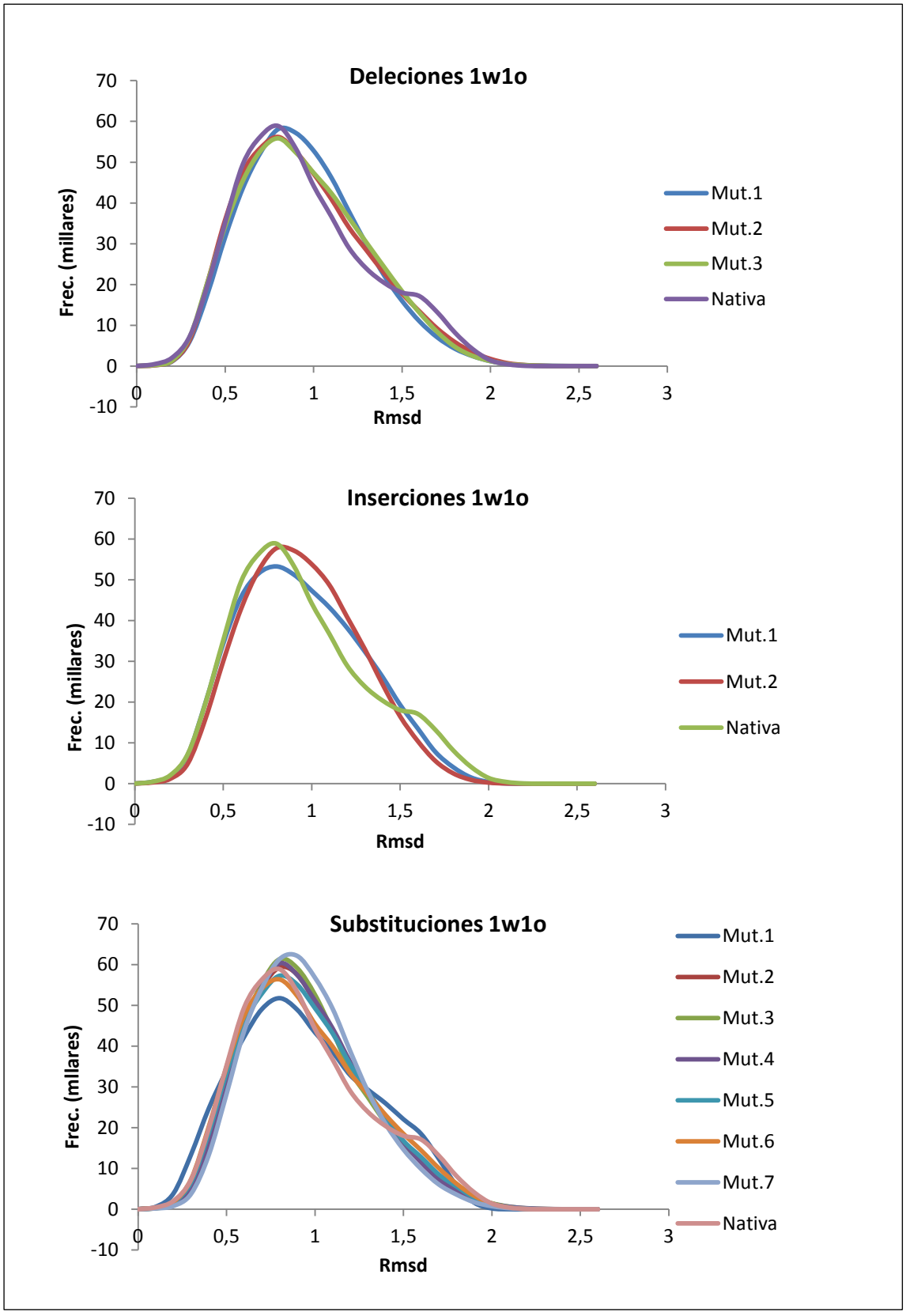


Fig. I.2 crmsd Mutante – Mutante 1w1o

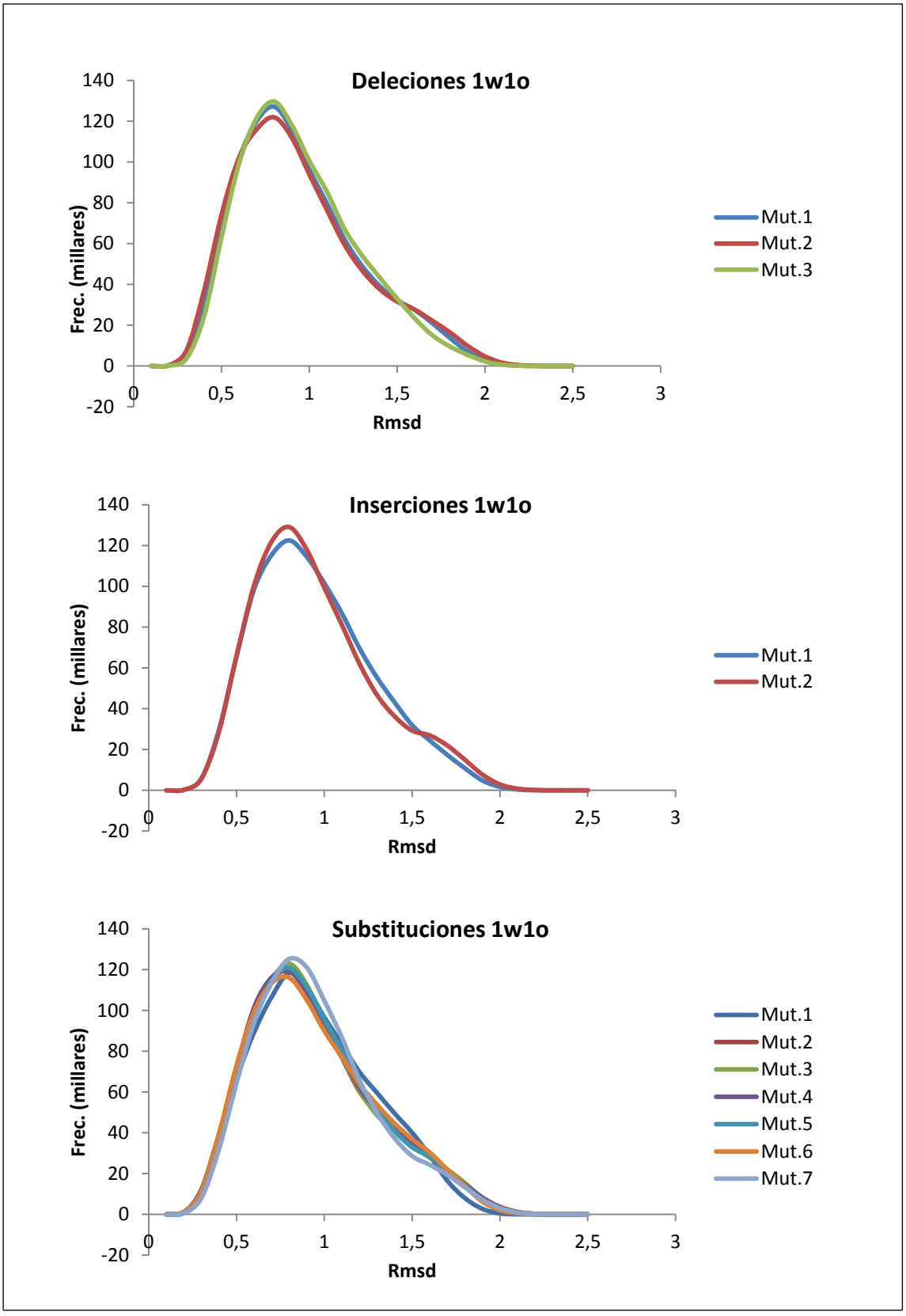


Fig. I.3 crmsd Mutante – Nativa 1w1o

Resultados 2HP7

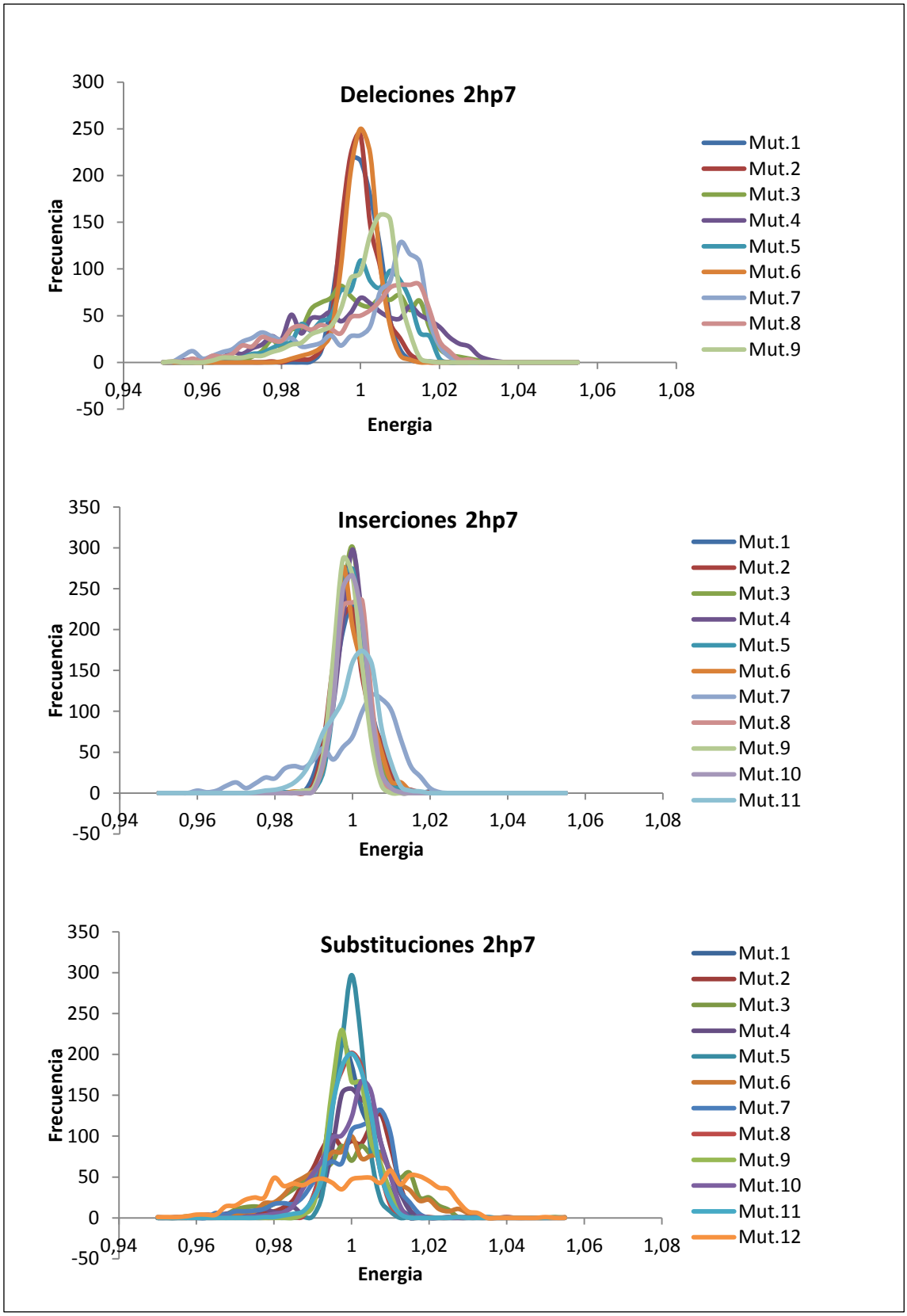


Fig. I.4 Energías 2hp7

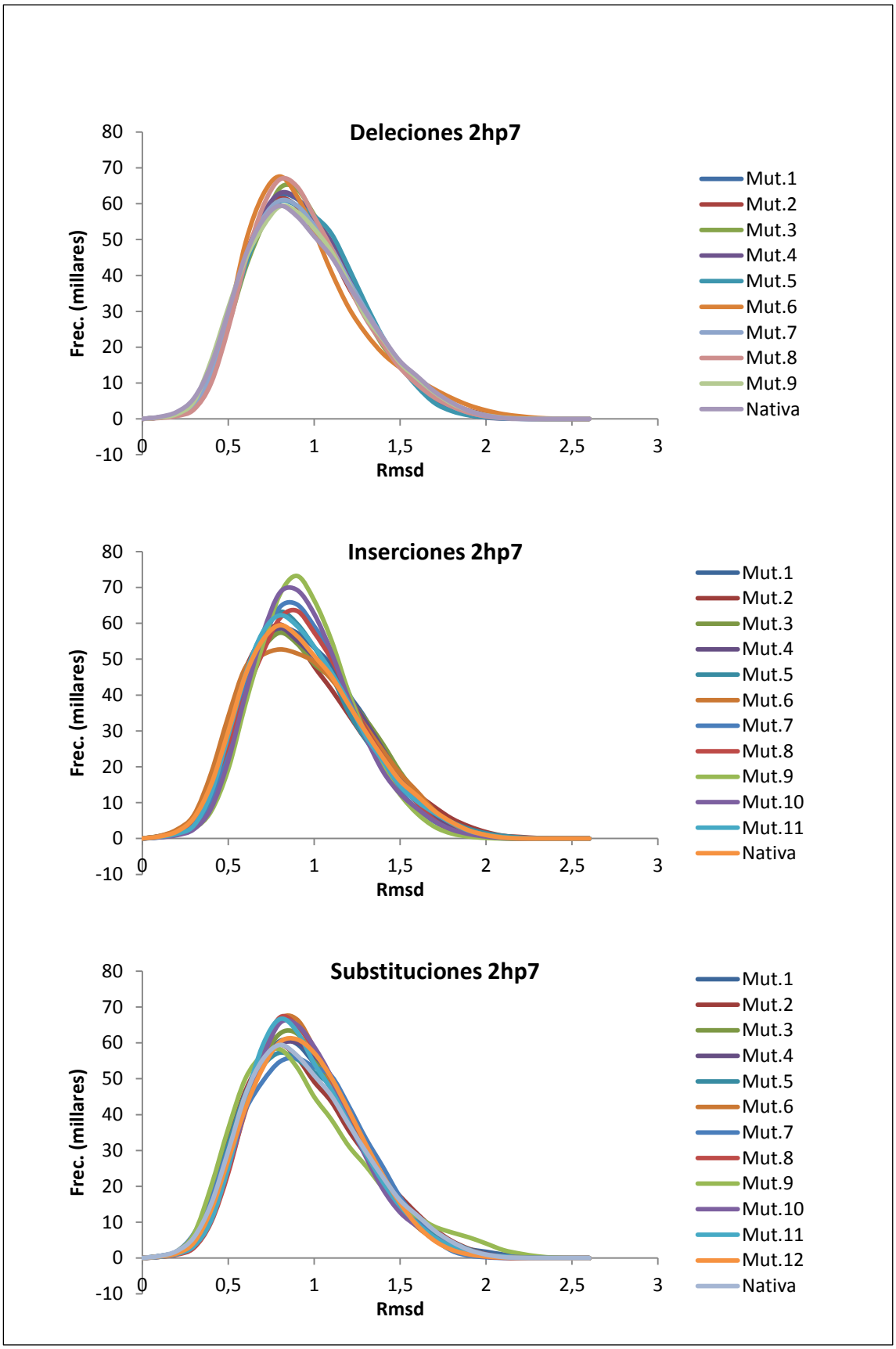


Fig. I.5 crmsd Mutante – Mutante 2hp7

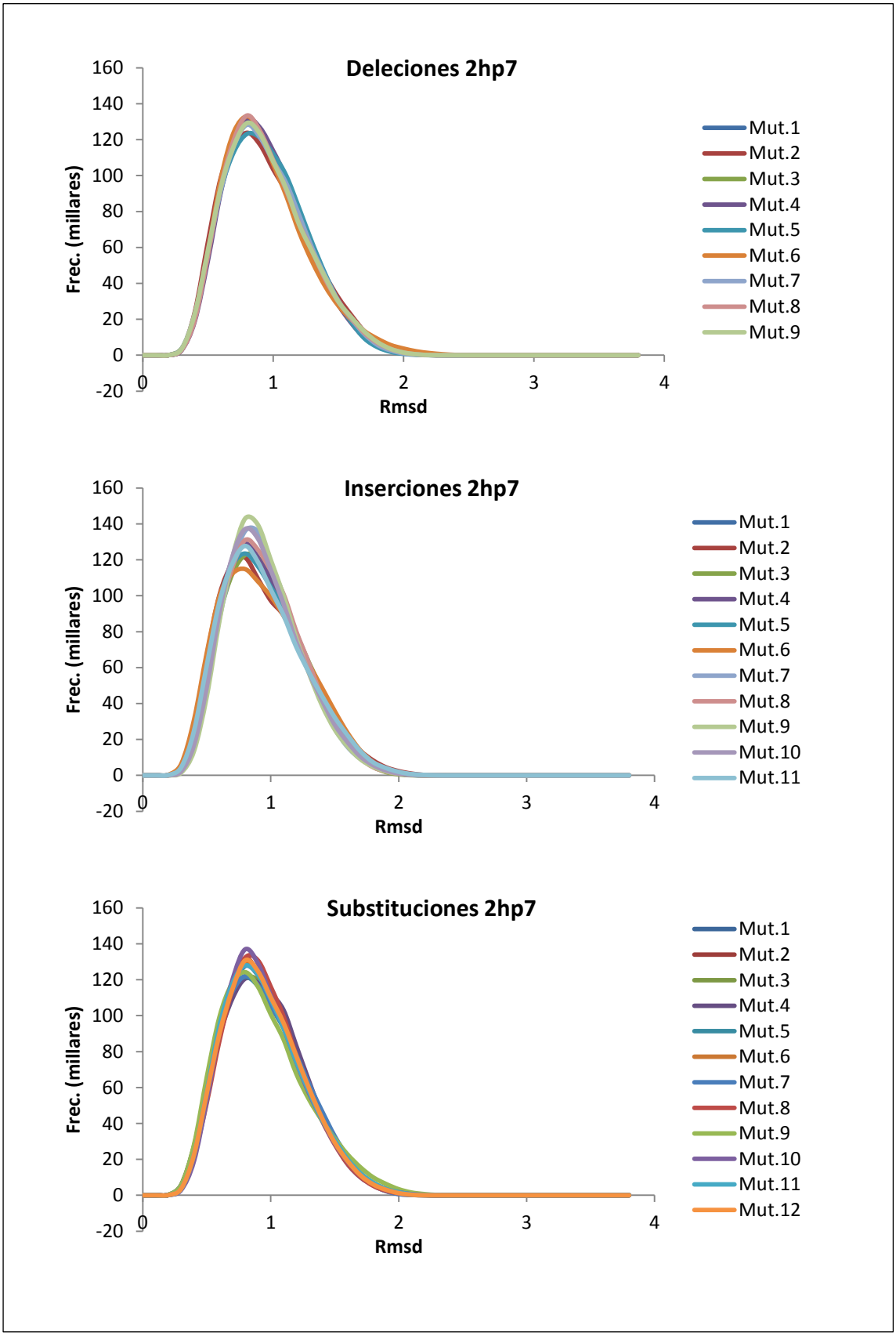


Fig. I.6 crmsd Mutante – Nativa 2hp7

Resultados 2HBU

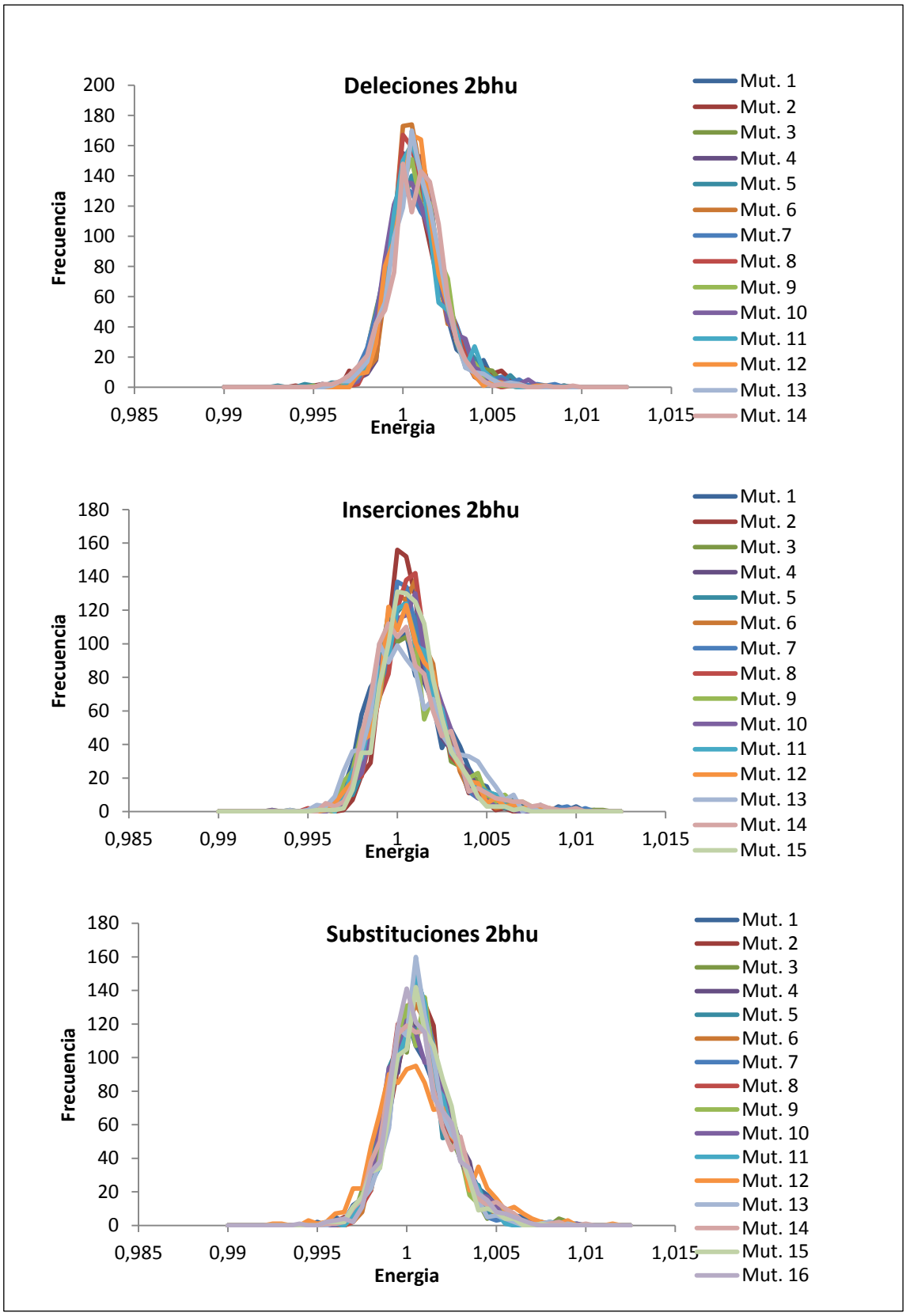


Fig. I.7 Energías 2bhu

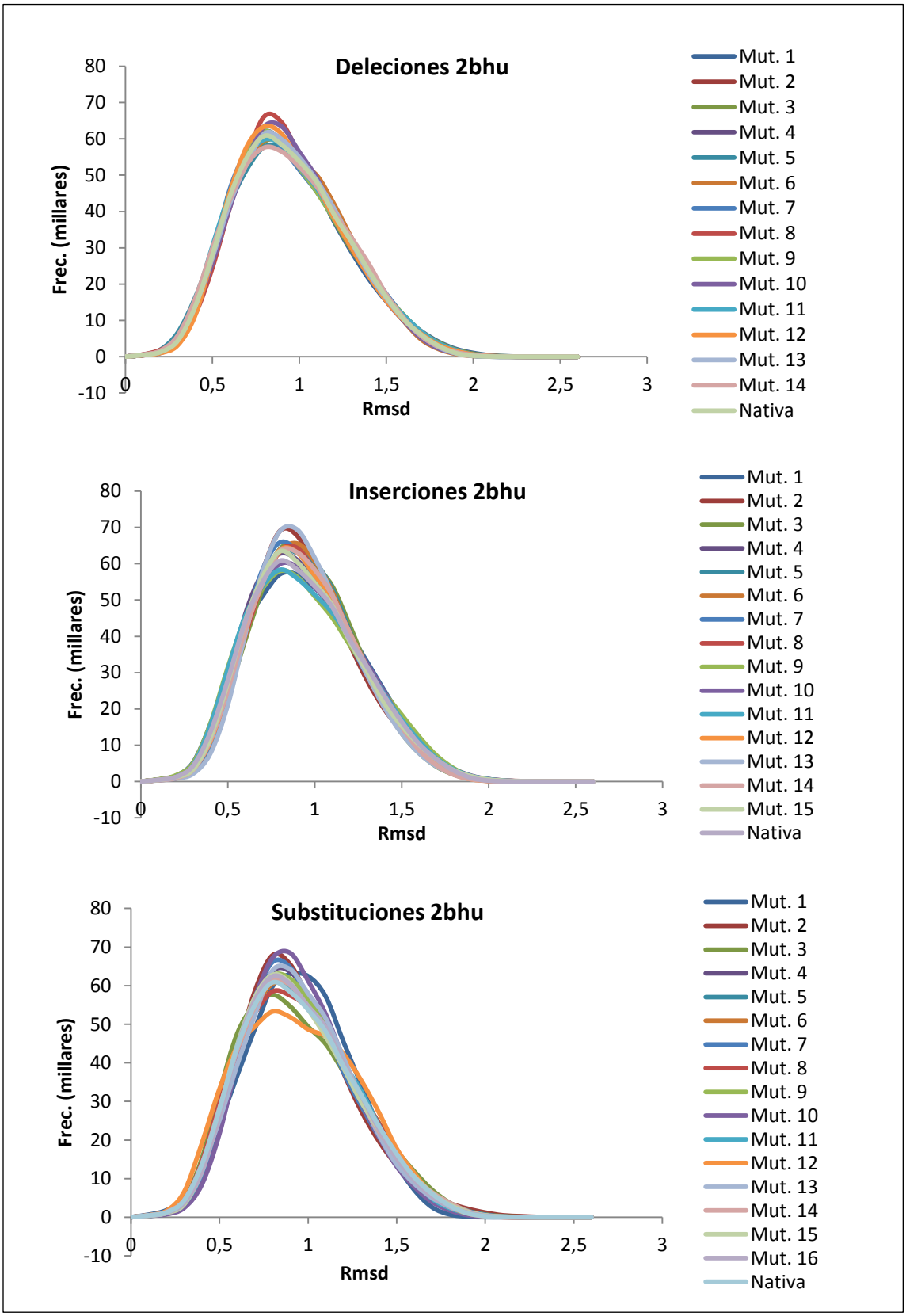


Fig. I.8 crmsd Mutante – Mutante 2bhu

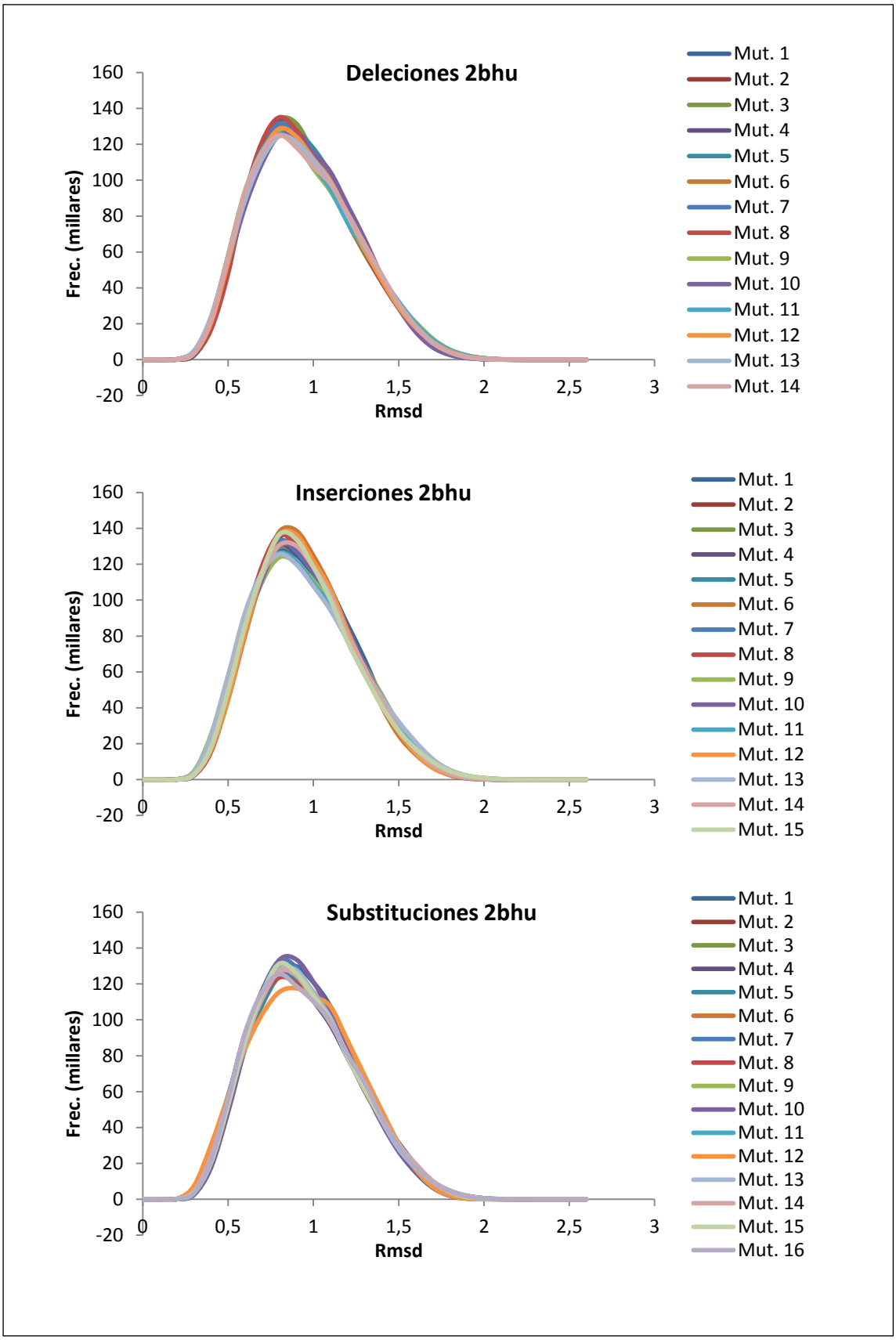


Fig. I.9 crmsd Mutante – Nativa 2bhu

Resultados 2I0K

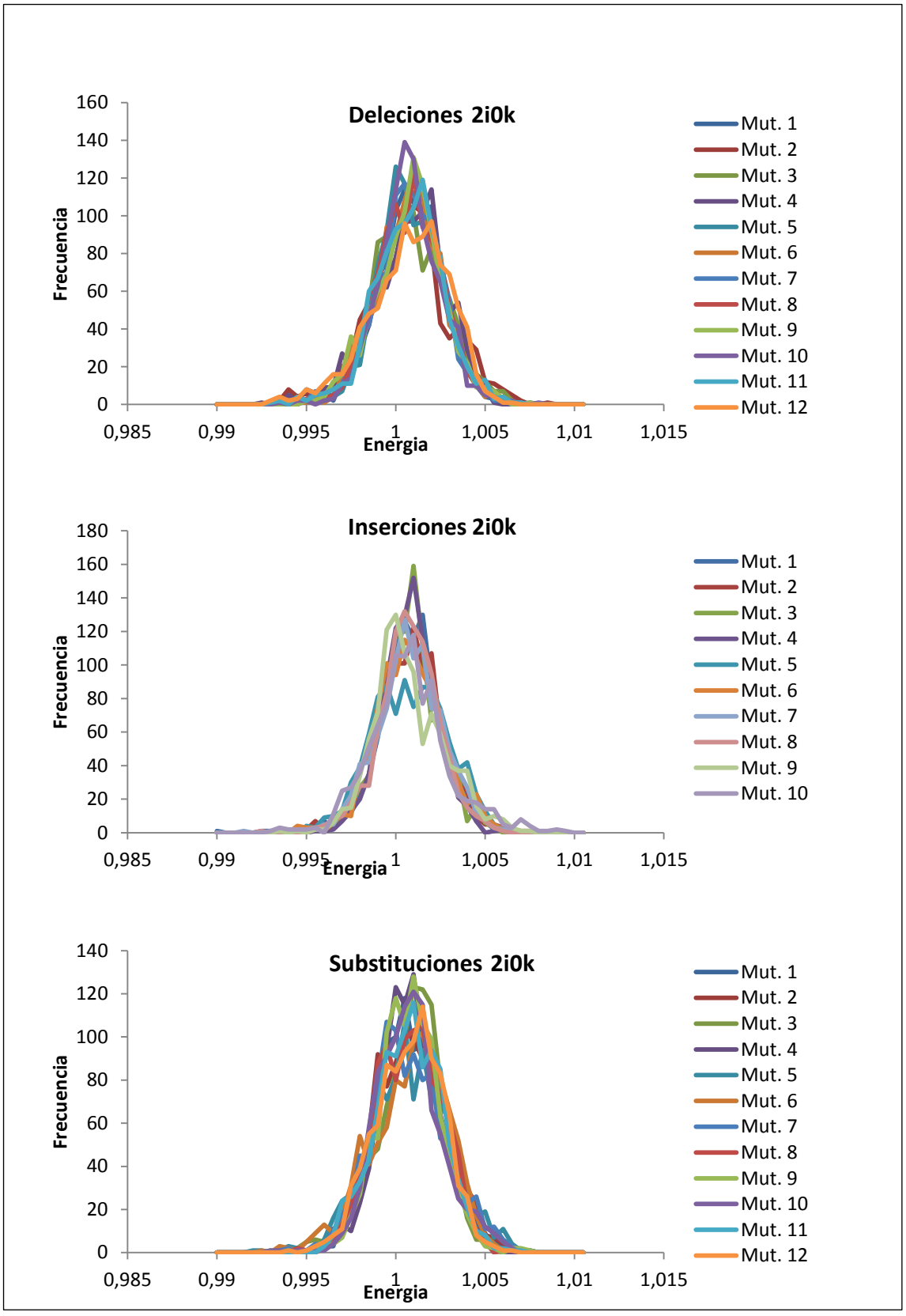


Fig. I.10 Energías 2i0k

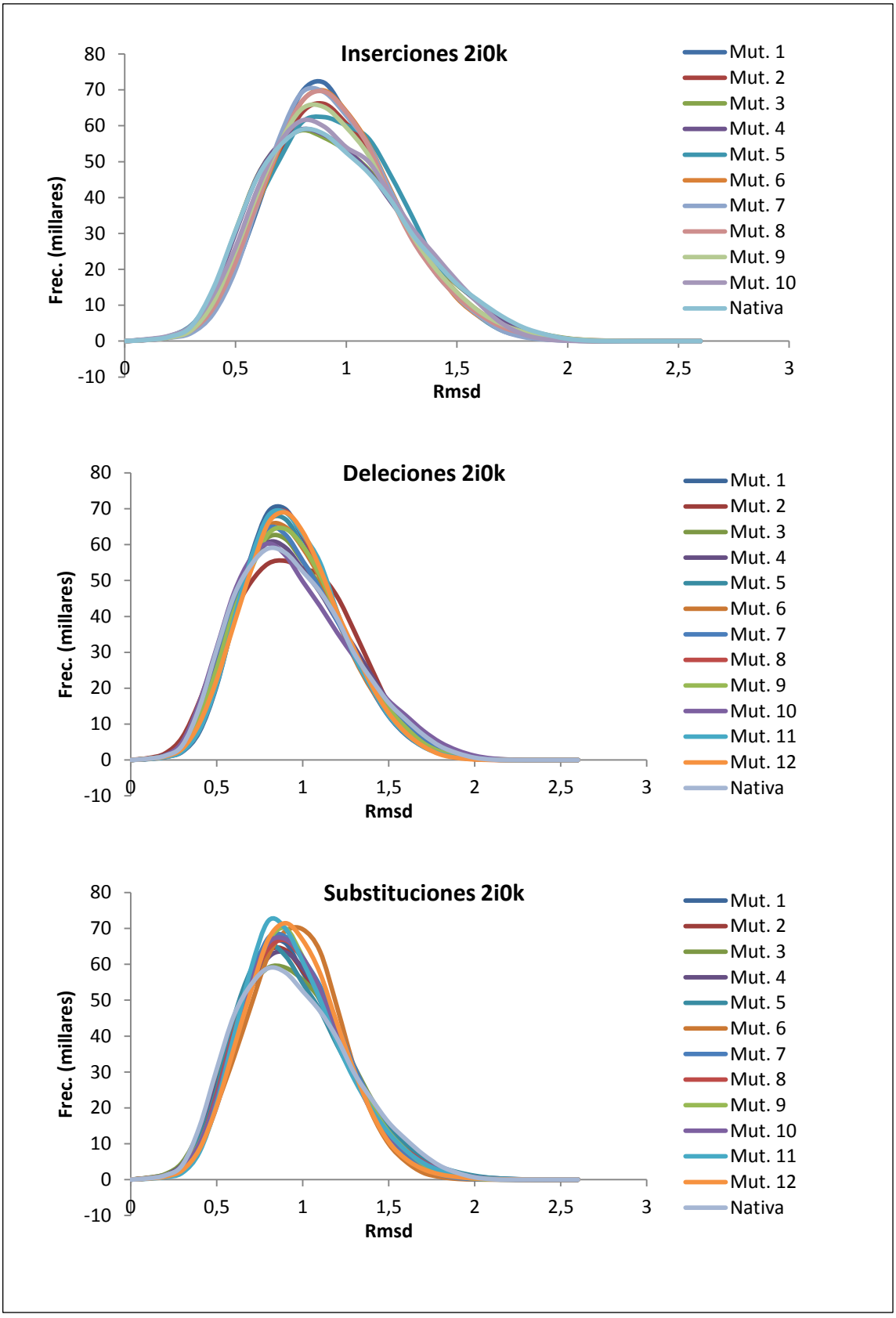


Fig. I.11 crmsd Mutante – Mutante 2i0k

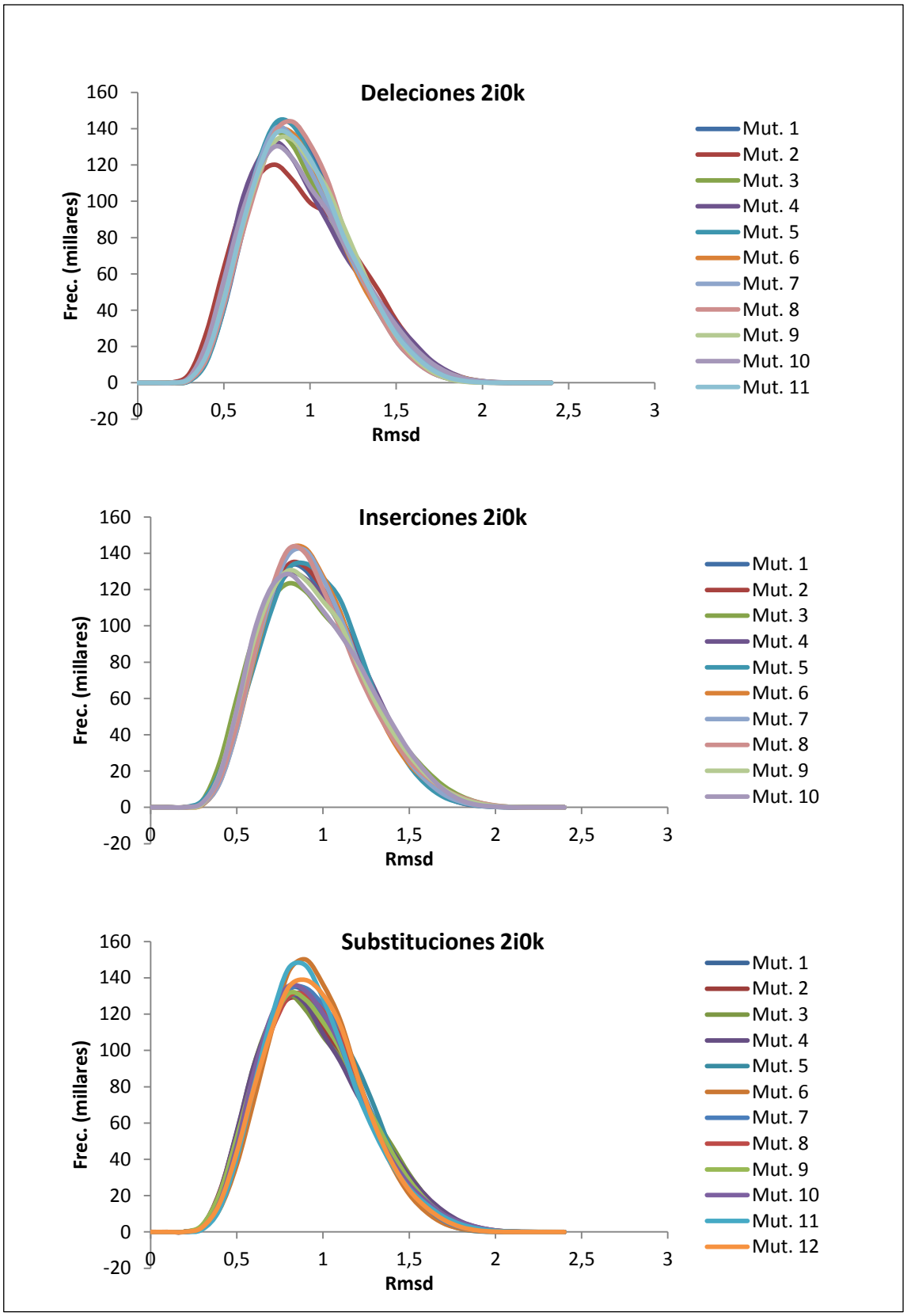


Fig. I.12 crmsd Mutante – Nativa 2i0k

Resultados 2PLW

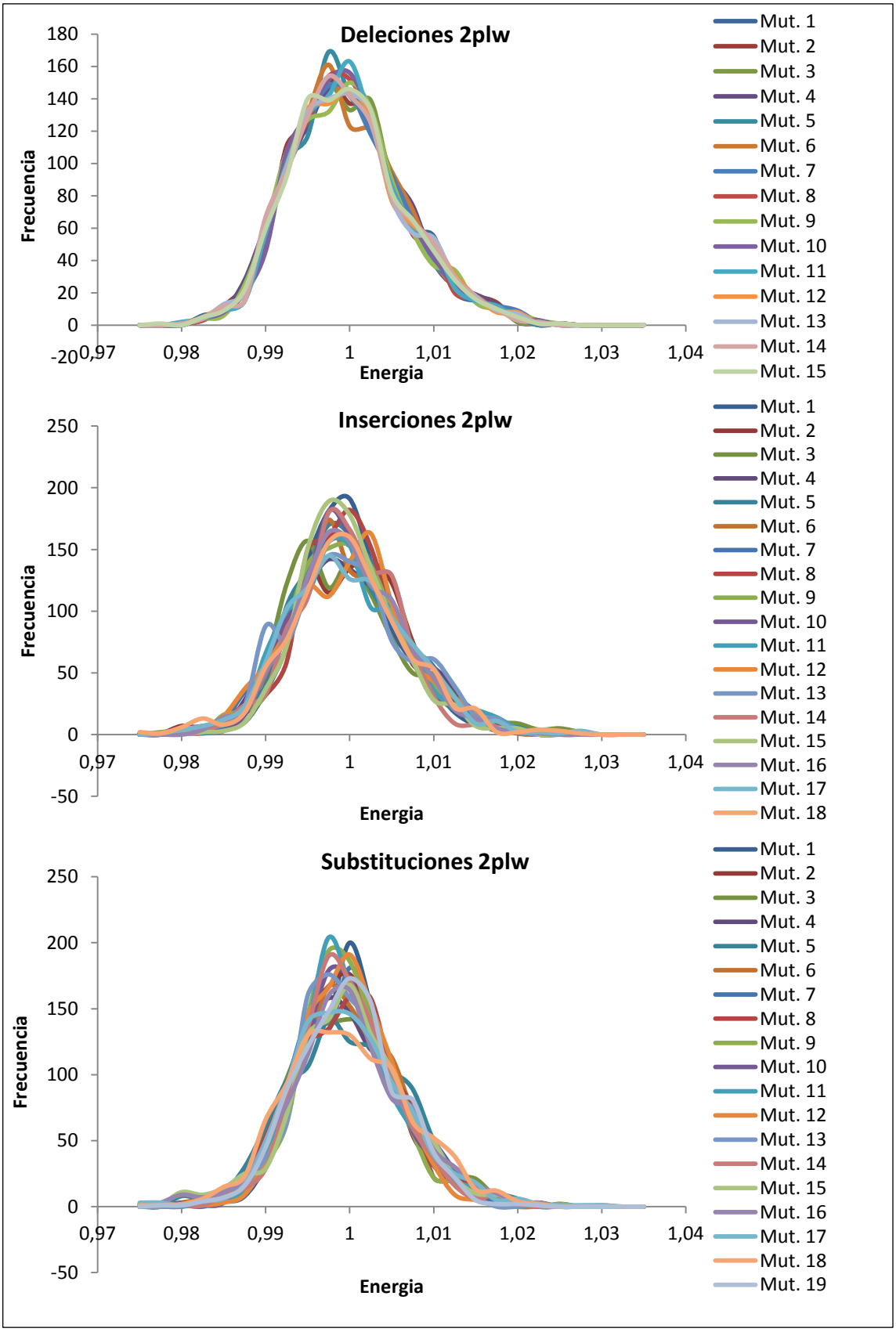


Fig. I.13 Energías 2plw

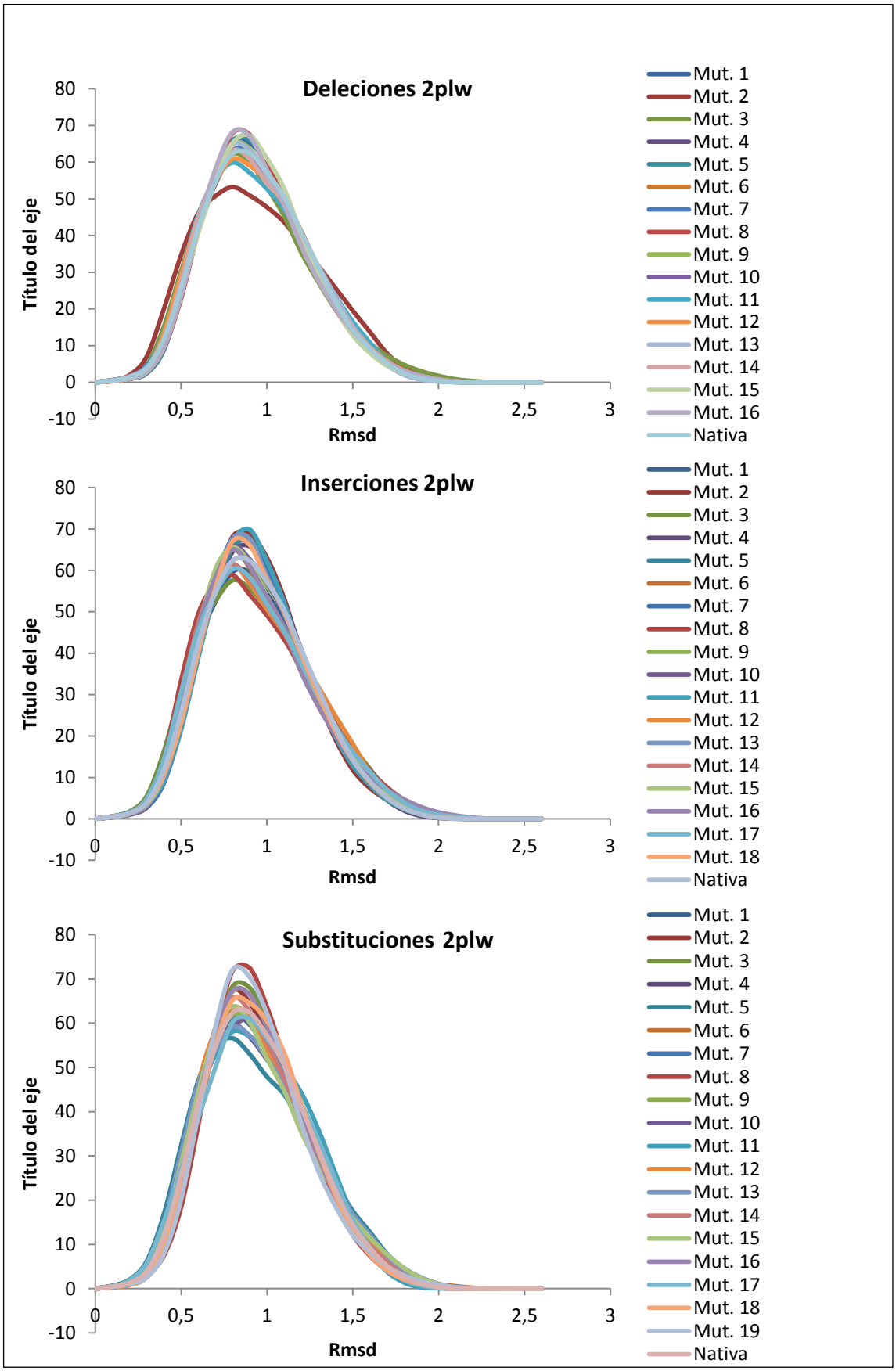


Fig. I.14 crmsd Mutante – Mutante 2plw

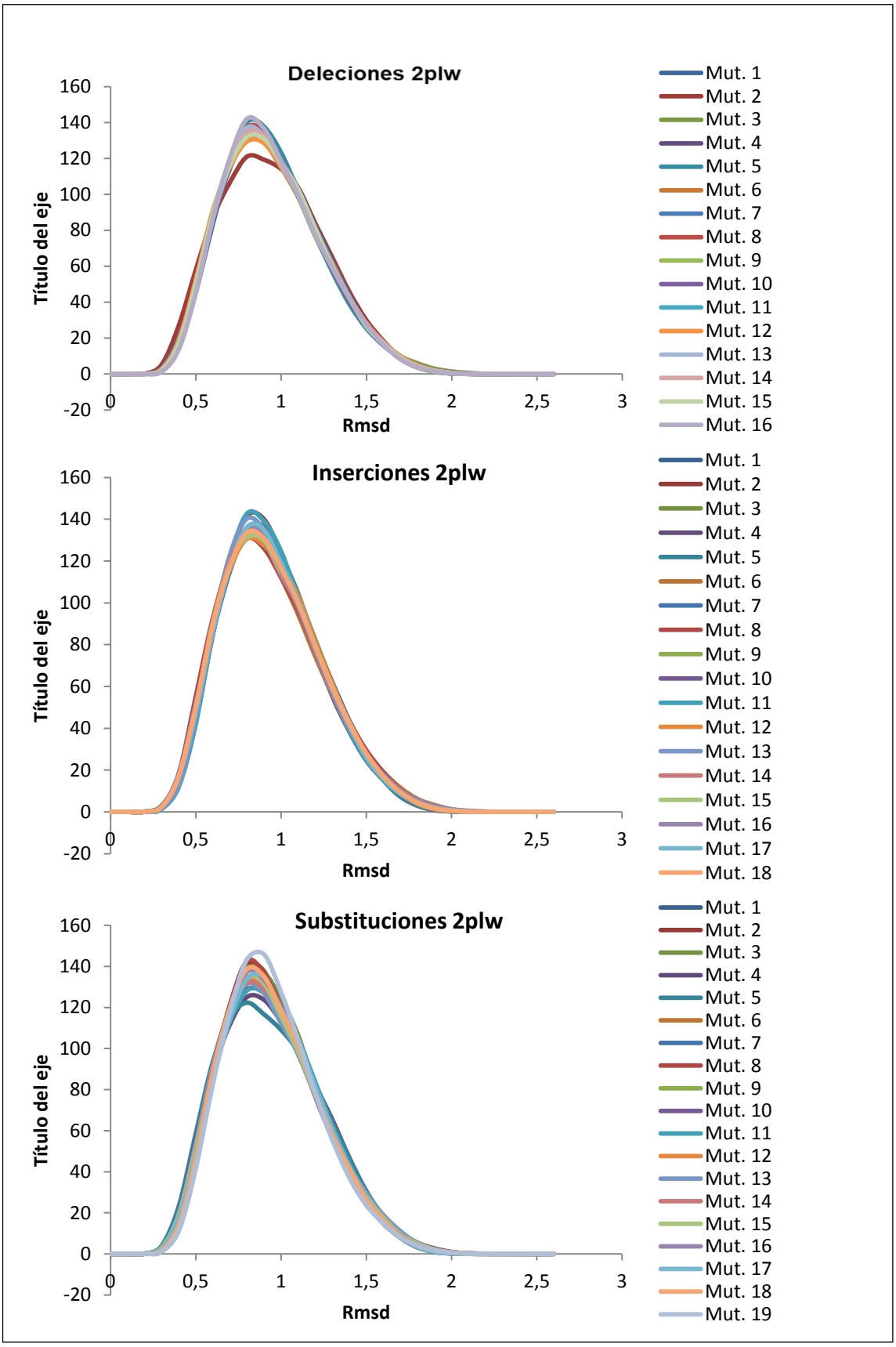


Fig. I.15 Crmsd Mutante – Nativa 2plw

Tablas de equivalencia mutante-secuencia

1w1o			
Nativa	ATAAAAA		
Mutantes	**Deleciones**	**Inserciones**	**Sustituciones**
Mut: 1	_TAAAAA	AATAAAAA	STAAAAA
Mut: 2	A_AAAAA	ATAAAAAA	ASAAAAA
Mut: 3	AT_AAAA		ATSAAAA
Mut: 4			ATASAAA
Mut: 5			ATAASAA
Mut: 6			ATAAASA
Mut: 7			ATAAAAS

Tabla I.1 Secuencia mutantes proteína 1w1o

2hp7			
Nativa	GGPGENPPNRPP		
Mutantes	**Deleciones**	**Inserciones**	**Sustituciones**
Mut: 1	G_PGENPPNRPP	GAGPGENPPNRPP	AGPGENPPNRPP
Mut: 2	GG_GENPPNRPP	GGAPGENPPNRPP	GAPGENPPNRPP
Mut: 3	GGP_ENPPNRPP	GGPAGENPPNRPP	GGDGENPPNRPP
Mut: 4	GGPG_NPPNRPP	GGPGAENPPNRPP	GGPAENPPNRPP
Mut: 5	GGPGE_PPNRPP	GGPGEANPPNRPP	GGPGQNPPNRPP
Mut: 6	GGPGEN_PNRPP	GGPGENAPPNRPP	GGPGEDPPNRPP
Mut: 7	GGPGENPP_RPP	GGPGENPAPNRPP	GGPGENDPNRPP
Mut: 8	GGPGENPPN_PP	GGPGENPPANRPP	GGPGENPDNRPP
Mut: 9	GGPGENPPNR_P	GGPGENPPNARPP	GGPGENPPDRPP
Mut: 10		GGPGENPPNRAPP	GGPGENPPNKPP
Mut: 11		GGPGENPPNRPAP	GGPGENPPNRDP
Mut: 12			GGPGENPPNRPD

Tabla I.2 Secuencia mutantes proteína 2hp7

2bhu			
Nativa	EGRKKEFGGFSGFSGE		
Mutantes	**Deleciones**	**Inserciones**	**Sustituciones**
Mut: 1	_GRKKEFGGFSGFSGE	EAGRKKEFGGFSGFSGE	QGRKKEFGGFSGFSGE
Mut: 2	E_RKKEFGGFSGFSGE	EGARKKEFGGFSGFSGE	EARKKEFGGFSGFSGE
Mut: 3	EG_KKEFGGFSGFSGE	EGRAKKEFGGFSGFSGE	EGKKKEFGGFSGFSGE
Mut: 4	EGRK_EFGGFSGFSGE	EGRKAKEFGGFSGFSGE	EGRRKEFGGFSGFSGE
Mut: 5	EGRKK_FGGFSGFSGE	EGRKKAEFGGFSGFSGE	EGRKREFGGFSGFSGE
Mut: 6	EGRKKE_GGFSGFSGE	EGRKKEAFGGFSGFSGE	EGRKKQFGGFSGFSGE
Mut: 7	EGRKKEF_GFSGFSGE	EGRKKEFAGGFSGFSGE	EGRKKEYGGFSGFSGE
Mut: 8	EGRKKEFGG_SGFSGE	EGRKKEFGAGFSGFSGE	EGRKKEFAGFSGFSGE
Mut: 9	EGRKKEFGGF_GFSGE	EGRKKEFGGAFSGFSGE	EGRKKEFGAFSGFSGE
Mut: 10	EGRKKEFGGFS_FSGE	EGRKKEFGGFASGFSGE	EGRKKEFGGYSGFSGE
Mut: 11	EGRKKEFGGFSG_SGE	EGRKKEFGGFSAGFSGE	EGRKKEFGGFAGFSGE
Mut: 12	EGRKKEFGGFSGF_GE	EGRKKEFGGFSGAFSGE	EGRKKEFGGFSAFSGE
Mut: 13	EGRKKEFGGFSGFS_E	EGRKKEFGGFSGFASGE	EGRKKEFGGFSGYSGE
Mut: 14	EGRKKEFGGFSGFSG_	EGRKKEFGGFSGFSAGE	EGRKKEFGGFSGFAGE
Mut. 15		EGRKKEFGGFSGFSGAE	EGRKKEFGGFSGFSAE
Mut. 16			EGRKKEFGGFSGFSGQ

Tabla I.3 Secuencia mutantes proteína 2bhu

2i0k			
Nativa	VGSLGSAGSLVG		
Mutantes	**Deleciones**	**Inserciones**	**Sustituciones**
Mut: 1	_GSLGSAGSLVG	VAGSLGSAGSLVG	IGSLGSAGSLVG
Mut: 2	V_SLGSAGSLVG	VGASLGSAGSLVG	VASLGSAGSLVG
Mut: 3	VG_LGSAGSLVG	VGSALGSAGSLVG	VGALGSAGSLVG
Mut: 4	VGS_GSAGSLVG	VGSLAGSAGSLVG	VGSIGSAGSLVG
Mut: 5	VGSL_SAGSLVG	VGSLGASAGSLVG	VGSLASAGSLVG
Mut: 6	VGSLG_AGSLVG	VGSLGSAAGSLVG	VGSLGAAGSLVG
Mut: 7	VGSLGS_GSLVG	VGSLGSAGASLVG	VGSLGSSGSLVG
Mut: 8	VGSLGSA_SLVG	VGSLGSAGSALVG	VGSLGSAASLVG
Mut: 9	VGSLGSAG_LVG	VGSLGSAGSLAVG	VGSLGSAGALVG
Mut: 10	VGSLGSAGS_VG	VGSLGSAGSLVAG	VGSLGSAGSIVG
Mut: 11	VGSLGSAGSL_G		VGSLGSAGSLIG
Mut: 12	VGSLGSAGSLV_		VGSLGSAGSLVA

Tabla I.4 Secuencia mutantes proteína 2i0k

2plw			
Nativa	KDNMNNIKNINYIDNMNNN		
Mutantes	**Deleciones**	**Inserciones**	**Sustituciones**
Mut: 1	_DNMNNIKNINYIDNMNNN	KADNMNNIKNINYIDNMNNN	RDNMNNIKNINYIDNMNNN
Mut: 2	K_NMNNIKNINYIDNMNNN	KDANMNNIKNINYIDNMNNN	KENMNNIKNINYIDNMNNN
Mut: 3	KD_MNNIKNINYIDNMNNN	KDNAMNNIKNINYIDNMNNN	KDDMNNIKNINYIDNMNNN
Mut: 4	KDN_NNIKNINYIDNMNNN	KDNMANNIKNINYIDNMNNN	KDNLNNIKNINYIDNMNNN
Mut: 5	KDNM_NIKNINYIDNMNNN	KDNMNANIKNINYIDNMNNN	KDNMDNIKNINYIDNMNNN
Mut: 6	KDNMNN_KNINYIDNMNNN	KDNMNNAIKNINYIDNMNNN	KDNMNDIKNINYIDNMNNN
Mut: 7	KDNMNNI_NINYIDNMNNN	KDNMNNIAKNINYIDNMNNN	KDNMNNVKNINYIDNMNNN
Mut: 8	KDNMNNIK_INYIDNMNNN	KDNMNNIKANINYIDNMNNN	KDNMNNIRNINYIDNMNNN
Mut: 9	KDNMNNIKN_NYIDNMNNN	KDNMNNIKNAINYIDNMNNN	KDNMNNIKDINYIDNMNNN
Mut: 10	KDNMNNIKNI_YIDNMNNN	KDNMNNIKNIANYIDNMNNN	KDNMNNIKNVNYIDNMNNN
Mut: 11	KDNMNNIKNIN_IDNMNNN	KDNMNNIKNINAYIDNMNNN	KDNMNNIKNIDYIDNMNNN
Mut: 12	KDNMNNIKNINY_DNMNNN	KDNMNNIKNINYAIDNMNNN	KDNMNNIKNINFIDNMNNN
Mut: 13	KDNMNNIKNINYI_NMNNN	KDNMNNIKNINYIADNMNNN	KDNMNNIKNINYVDNMNNN
Mut: 14	KDNMNNIKNINYID_MNNN	KDNMNNIKNINYIDANMNNN	KDNMNNIKNINYIENMNNN
Mut. 15	KDNMNNIKNINYIDN_NNN	KDNMNNIKNINYIDNAMNNN	KDNMNNIKNINYIDDMNNN
Mut. 16	KDNMNNIKNINYIDNM_NN	KDNMNNIKNINYIDNMANNN	KDNMNNIKNINYIDNLNNN
Mut. 17		KDNMNNIKNINYIDNMNANN	KDNMNNIKNINYIDNMDNN
Mut. 18		KDNMNNIKNINYIDNMNNAN	KDNMNNIKNINYIDNMNDN
Mut. 19			KDNMNNIKNINYIDNMNND

Tabla I.5 Secuencia mutantes proteína 2plw

APENDICE II – MANUAL DE FUNCIONAMIENTO DEL PROGRAMA DE SIMULACIÓN

Menú principal	
go	Empieza una ejecución dinámica
revu	Revisa los parámetros dinámicos establecidos
rorg	Cambia el residuo origen (origen del sistema de coordenadas) Una vez establecido, esta posición no cambia durante la ejecución dinámica
sprob	Establece probabilidades para los estados discretos de cada residuo Solo tiene una opción disponible: establece las probabilidades igual a los pesos relativos dados con cada estado ds (puede verse con el comando-ds- dentro del menu –enq-) Se ha de ejecutar antes de la ejecución dinámica
sstr	(igual que en el menú -io-)
Menú –io-	**(invoca al menú de Entrada / Salida)**
rref / rr	Lee una estructura de referencia de un fichero El fichero tiene que estar en formato PDB Se lee después de 'rs'. Si la secuencia es diferente de la introducida en 'rs' da error Solo se consideraran los átomos Ca Parámetros: -b: Se construye una estructura de referencia usando los estados discretos con el método descrito en : Park & Levitt, JMB, 249, 493-507 -c: Se usará el 'COIL state' en la simulación -bt: Se asume que la estructura ha sido construida a partir de un set de puntos de referencia ds
rseq / rs	Lee una secuencia de aminoácidos de un fichero

	Las líneas encabezadas por el símbolo # se consideran comentarios y se ignoran Se ignoran los espacios, tabulaciones y saltos de línea La secuencia de aminoácidos es una cadena de 1-letra en mayúsculas por cada aminoácido Ejemplo: #2CRO 1 65 MQTLSERLKKRRIALKMTQTELATKAGVKQQSIQLIEAGVTKRPRFLFEIA-MALNCDPVWLQYGT
sstr / ss	Establece la estructura inicial. Puede ser la dada inicialmente o se puede usar para establecer la de una estructurar arbitrariamente escogida de una ejecución anterior Parámetros: -d (default): Establece la estructura por defecto (conformación beta) Este parámetro se asume por defecto si no se le da ninguno -s: Establece cada residuo de la estructura a un estado especificado por un fichero El fichero tiene el mismo formato que el que saca el comando 'wds' -r: Igual que la opción -s. Sin embargo se selecciona los IC como los puntos representativos de cada estado en lugar de escoger un ángulo aleatorio de la zona continua que dicho estado discreto pertenece -a: Establece las coordenadas internas de la estructura a los valores especificados por un fichero. El formato del fichero es el mismo que el que saca el comando 'wic' -b: Igual que -d pero se asume que la estructura ha sido construida a partir de un set de puntos de referencia 'ds'
wcc	Escribe las coordenadas cartesianas en formato PDB a un fichero Parámetros: -r: Escribe la estructura de referencia
wds	Escribe la estructura en símbolos de estado discreto a un fichero Parámetros: -r: Escribe la estructura de referencia
wic	Escribe en coordenadas internas (ángulos y distancias) a un fichero Parámetros: -r: Escribe la estructura de referencia

Menú –dpar- (se invocan los parámetros dinámicos)	
bump	Establece la distancia de mínima de contacto entre dos Ca La distancia de mínima de contacto es la distancia más cercana permitida para un par de átomos de Ca Los Ca tiene que tener una distancia en secuencia >=3 Valor por defecto: 3.8
dbmax	Establece el máximo permito cambio en magnitud de los cambios de ángulo 'bond' Parámetros: angle: valor del ángulo en grados (180 por defecto)
dbor	Establece el parámetro dbor, que define el espacio de vecinos. El espacio de vecinos de un residuo -i- son todos los residuos -j- que están dentro de la esfera de radio -dbor- centrado en i. Este parámetro se utiliza para calcular -ldrms- Parámetros: d: valor de dbor (8.0 por defecto)
dtmax	Establece el máximo permito cambio en magnitud de los cambios de ángulo torsional Parámetros: angle: valor del ángulo en grados (180 por defecto)
enmps	Establece el número efectivo de movimientos de Monte Carlo que define un paso de Montecarlo Un número efectivo de movimientos de Monte Carlo es el número de residuos que es movido en un movimiento de Monte Carlo Un paso de MC se devine como la sucesión de movimientos de MC (enm) tales que la suma de todos estos enm's para todos los movimientos esta justo más allá del enmps Ejemplos de movimientos de MC (enm) enm = 1 : movimiento que rota el residuo -i- sobre el eje que lo une con el i-1 y i+1 enm = 2: movimiento que rota residuos sobre el eje que une los residuos i e i+3 enm = i o nres -i -1 (el más pequeño): movimiento que cambia el ángulo tau para el residuo i Notas: El chequeo de la distancia mínima de contacto se realiza después de cada movimiento (enm) El energy check se realiza solo después de cada paso de MC
eout	Establece los parámetros de escritura de la energía Parámetros:

	- : no escribe nada -s: escribe para cada paso -B: escribe para cada batch -b: escribe para cada batch pero sin estadísticas de movimientos fname: nombre del fichero de salida (si no se pone se coge el del parámetro 'seed') filename.eo
ereout	Establece los parámetros de escritura de la energía por residuo Parámetros: - : no escribe nada -r : escribe la energía promedio de cada residuo por cada batch fname: nombre del fichero de salida (si no se pone se coge el del parámetro 'seed') filename.ero
loop	Establece los Parámetros de modelado de loops Solo se puede modelar hasta 5 loops al mismo tiempo Parámetros: (sin Parámetros): Muestra el número de loops a modelar, el primer loop y últimos residuos y el máximo número de variaciones de ángulos diédricos - : Elimina todos los loops almacenados n1: Primer residuo del loop. No ha de ser mayor que 3 n2: Último residuo del loop: Tiene que ser menor que 'nres -3' thr: umbral de variación del ángulo diédricos m: Indica que el loop será modelado de forma aleatoria "with a clousure condition"
mout	Establece Parámetros para la salida de la información referentes a los movimientos. Parámetros: (sin argumentos): Muestra los Parámetros actuales - : no muestra la salida filename: fichero de salida (*..mo). * (nombre por defecto (ver 'seed' comando para nombre por defecto) El fichero de salida está en formato GPLOT
mtf	Establece la frecuencia relativa de cada tipo de movimiento Mc Se establece por orden: primera frec. --> movimiento tipo 0 segunda frec --> movimiento tipo 1 # tipos de movimientos:

	# 0. NBT /* cambio de los ángulos de enlace y torsión en el 'N-terminus' */ # 1. TBC /* cambio de los ángulos de enlace y torsión en el 'C-terminus' */ # 2. BTB /* cambio 'ba-ta-ba' dentro del mismo ds */ # 3. DSJ /* salto entre estados discretos */ # 4. LOOP /* fixed ends loop move */ # 5. LOC /* nov. aleatorio local combinado con la condición de 'loop closure' */
mxenr	Establece un umbral para un término de energía durante el la ejecución MC Si la energía es mayor que el umbral la ejecución se detendrá Parámetros: mxe: término de energía mxthr: umbral de energía
nbor	Establece el parámetro -nbor- que define los vecinos de secuencia Los vecinos de sec. de un resido -i- son: (i - nobor) a (i + nbor). Se utiliza para calcular -sdrms- Valor por defecto: 8
nspb	Agrupa los pasos de Montecarlo de forma que en cada agrupamiento la temperatura se mantiene constante y cambia al siguiente agrupamiento
nstep	Establece el número total de pasos de Monte Carlo El valor por defecto es 5000
R	Establece la constante de gas R. Se usa en el chequeo de Metrópolis: (new E - old E)/ R*T De esta forma R actúa solo como factor de escala de energía. Las unidades de R son las mismas que las usadas para el cálculo de energías
revu	Revisa los Parámetros dinámicos establecidos
seed	Semilla de generador de números aleatorios (cadena de 6 caracteres -> número 48 bits) Si no se ponen todos los caracteres se rellena con ceros por la derecha El nombre usado por la semilla se utiliza para la salida de fichero .trj
Tempe-ratura / T	Establece el principio y final de la temperatura. Parámetros: T1: temperatura inicial T2: temperatura final -l: La temperatura varia de forma logarítmica (por defecto es lineal)

	Cuando la temperatura varia linealmente, esta cambia cada batch dT= (T2-T1/num. batches Cuando varía logarítmicamente la temperatura cambia por un factor -f- cada batch: f=(T2-T1)^(1/num. batches)
thenr	Establece un umbral para cada término de energía. Si el término de energía es más bajo que el umbral, la ejecución se parará
trj	Establece los Parámetros de salida de la trayectoria Parámetros: - : no escribe el output de la trayectoria -B: escribe antes del chequeo de la distancia mínima de contacto -b: escribe después del chequeo de la distancia mínima de contacto -M: escribe antes del Metrópolis check -m: escribe después del Metrópolis check (i.e. después de cada paso MC) -n: escribe de cada n-ésimo paso MC nombre de fichero de salida .trj (si no se especifica se pone el de la semilla)
wet	Pesos de cada término de la energía. Los números especifican los pesos relativos La energía se da como suma de los términos siguientes: drmsd: drmsd desde la estructura de ref. crmsd: rmsd después de superposición utilizando el procedimiento Kabsch w1drmsd: drmsd ponderado por 1/(distancia en la estructura de ref.). w2drmsd: drmsd ponderado por 1/(distancia en la estruc. de ref..)^2 sdrmsd: drmsd entre vecinos en secuencia lndrmsd: drmsd entre vecinos espaciales en la nueva estructura lrdrmsd: drmsd entre vecinos espaciales en la estructura de referencia dihrmsd: rmsd de ángulos diédricos (entre la nueva estruc. y la de ref.) chirmsd: rmsd que refleja la desviación de la quiralidad en cada residuo rdgrmsd: rmsd del radio de giro (entre la nueva estruc. y la de ref.)
Menú enq (comandos de informaciones varias)	
aofn	Muestra el nombre del fichero actual de salida acumulado
dij	Calcula y muestra las distancias entre Ca (excepto las i -> i+1) que son más pequeñas que un umbral dado Parámetros: -r: Se calcula para la estructura de referencia dmax: umbral de distancias Ca

ds	Muestra la información de los estados discretos usados (fichero -ds.table-)
fname	Muestra el actual nombre de fichero de secuencia
mt	Lista los tipos de movimientos disponibles
nres	Muestra el número de residuos
prob	Muestra las probabilidades de un estado discreto para un residuo dado Parámetros: r: número del residuo
rorg	(igual que en el menú principal)
seq	Muestra la secuencia (1 letra por aminoácido) de la estructura actual
Menú man (comandos de manipulación manual)	
1ba	Cambia un ángulo de 'bond' de un residuo Parámetros: r: número del residuo ba: ángulo a cambiar (número entero)
1bt	Cambia el ángulo de 'bond' y el torsional (r -> r+1) de un residuo Parámetros: r: número del residuo ba: ángulo de 'bond' a cambiar (número entero) ta: ángulo de torsión a cambiar (número entero)
1ta	Cambia el ángulo torsional (r -> r+1) de un residuo Parámetros: r: número del residuo ta: ángulo de torsión a cambiar (número entero
1tb	Cambia el ángulo de 'bond' y el torsional (r -1 -> r) de un residuo Parámetros: r: número del residuo ba: ángulo de 'bond' a cambiar (número entero) ta: ángulo de torsión a cambiar (número entero)

btb	Cambia los ángulos 'bond' de r y r+1 y el torsional (r -> r+1) Parámetros: r: número del residuo ba1: ángulo de 'bond' a cambiar de r (número entero) ta: ángulo de torsión a cambiar (número entero) ba2: ángulo de 'bond' a cambiar de r+1 (número entero)
dsj	Cambio el estado discreto de un residuo. Si el residuo es 0, el último o penúltimo el comando aborta El comando cambia los ángulos 'bond' de r y r+1 y el ángulo torsional (r -> r+1) a los valores por defecto de la tabla -ds- escogida Parámetros: r: número del residuo dscode: nombre (1 letra) del estado discreto
rorg	(igual que en el menú principal)
sstr	(igual que en el menú -io-)
w2ic	Escribe dos coordenadas internas para cada residuo: una es la almacenada y la otra la calculada a partir de las coord. cartesianas Se realiza un test para comprobar la consistencia entre las dos coord. Parámetros: fname: fichero de salida -r: Se utiliza la estructura de referencia

GLOSARIO

- **Campo de fuerzas** (modelos moleculares): Se refiere a la función de forma y los parámetros que se establecen para calcular el potencial de energía de un sistema de átomos o de partículas 'de grano grueso'. Los campos de fuerzas se utilizan para simulaciones de mecánica molecular y dinámica molecular. Los parámetros pueden ser derivados de trabajos experimentales o establecidos de forma teórica.

- **Superficie de Energía potencial (PES)**: Describe la energía de un sistema, i.e. colección de átomos, en términos de ciertos parámetros, normalmente la posición de dichos átomos. La superficie puede definir la energía como función de una o más coordenadas. Si solo hay una coordenada, la superficie se llama curva de energía potencial. La PES se utiliza como herramienta para el análisis de la geometría molecular y dinámica de reacciones químicas.

- **Potencial efectivo**: Se refiere a una expresión matemática que combina múltiples efectos un potencial simple. Se usa de forma común para el cálculo de órbitas de planetas y para cálculos semi-clásicos de sistemas compuestos por átomos. Permite muchas veces reducir el número de grados de libertad de dichos sistemas.

- **Potencial estadístico**: También llamado 'knowledge-based potential', es una función de energía basada en el análisis de estructuras de proteínas conocidas. Existen muchos métodos para obtener dichos potenciales como la aproximación 'quasi-química' o el potencial de fuerzas medias. Los potenciales estadísticos se utilizan como funciones de energía en la evaluación de un conjunto de modelos estructurales producidos por modelaje por homología, así como en modelos de caminos de plegamiento de proteínas.

- **Proteínas homólogas**: Son aquellas cuyas secuencias de aminoácidos son similares entre si debido a que presentan un mismo origen evolutivo. En bioinformática se utilizan las proteínas homólogas para determinar qué partes de una proteína son importantes en la formación de la estructura y en la interacción

con otras proteínas. Esta información también se utiliza en modelaje por homología para predecir la estructura de una proteína una vez conocida la estructura de la proteína homóloga

- **Frustración de la energía**: La frustración en física estadística se trata de la incapacidad de un sistema de hacer mínimos todos los términos de la energía de simultáneamente. En el plegamiento de proteínas se introduce el principio de 'mínima frustración'. Este principio dice que la naturaleza ha escogido aquellas secuencias de aminoácidos de forma que el estado plegado de la proteína es muy estable. Igualmente, las interacciones entre aminoácidos a lo largo del proceso de plegamiento son reducidos de forma que la llegada al estado plegado se convierte en un proceso muy rápido.

- **Volumen excluido**: Se refiere a la idea que una parte de la cadena molecular no puede ocupar espacio que ya está ocupado por otra parte de la misma molécula. Esto hace, por ejemplo, que los extremos de una cadena en solución están en promedio más separados de lo que estarían si no existiera el fenómeno del volumen excluido.

- **Efecto estérico**: En química orgánica, es un impedimento causado por la influencia de un grupo funcional de una molécula en el curso de una reacción química. Existen diferentes clases: 1. Impedimento estérico: porque el volumen ocupado por una parte de la molécula impide que la otra parte reaccione; 2. Repulsión estérica: porque un grupo de una molécula es aparentemente debilitado o protegido por grupos funcionales menos cargados o con carga eléctrica opuesta; 3. Atracción estérica: cuando las moléculas tienes formas o geometrías optimizadas para sus interacciones

- **Replication slippage**: Es un mecanismo por el que se produce una expansión o contracción de los trinucleótidos o dinucleótidos durante la replicación de ADN. Esto normalmente ocurre cuando se encuentran secuencias repetitivas de nucleótidos en el lugar de la replicación.

REFERENCIAS

1. T.E. Creighton (1992). “Proteins: Structures and Molecular Properties”. Freeman, New York.
2. L. Mirny, E. Shakhnovich (2001). “Protein folding theory: from lattice to all-atom models”. Annual Review of Biophysics and Biomolecular Structure, vol. 30, no. 1, pp. 361–396.
3. Zhang Y., Skolnick J. (2005). “The protein structure prediction problem could be solved using the current PDB library”. Proc Natl Acad Sci USA 102 (4): 1029–1034.
4. Kolinski A. (2004). “Protein modeling and structure prediction with a reduced representation”. Acta Biochim. Pol. 51:349–371.
5. Tozzini V. (2005). “ Coarse-grained models for proteins”. Curr Opinion Struct. Biol 15:144–150.
6. Colombo G., Micheletti C. (2006). “Protein folding simulations: combining coarse-grained models and all-atom molecular dynamics”. Theor Chem Acc 116:75–86.
7. Liwo A., Khalili M., Scheraga HA. (2005). “Ab initio simulations of protein-folding pathways by molecular dynamics with the united-residue model of polypeptide chains”. Proc Natl Acad Sci USA 102:2362–2367.
8. Bhella D., Ralph A., Yeo RP (2004). “Conformational flexibility in recombinant measles virus nucleocapsids visualised by cryo-negative stain electron microscopy and real-space helical reconstruction”. J Mol Biol 340:319-331.
9. White SH., Von Heijne G (2004). ”The machinery of membrane protein assembly”. Curr Opin Struct Biol 14:397-404.
10. Atilgan AR., Durell SR., Jernigan RL., Demirel MC., Keskin O., Bahar I. (2001). “Anisotropy offluctuations dynamics of proteins with an elastic network model”. Biophys J 80:505-515.
11. Bahar I., Atilgan AR., Erman B. (1997). “Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential”. Fold Des 2:173–181.

12. Hinsen K. (1998). "Analysis of domain motions by approximate normal mode calculations". Proteins: Struct Func Genet 33:417–429.

13. Chu JW., Voth GA. (2007). "Coarse-grained free energy functions for studying protein conformational changes: a double-well network model". Biophys J 11:3860–3871.

14. Moritsugu K., Smith JC. (2008). "REACH coarse-grained biomolecular simulation: transferability between different protein structural classes". Biophys J 95:1639–1648.

15. Tobi D., Bahar I. (2005). "Structural changes involved in protein binding correlate with intrinsic motions of proteins in the unbound state". Proc Natl Acad Sci USA 102:18908–18913.

16. Baker D. (2000). "A surprising simplicity to protein folding". Nature 405:39-42.

17. Cheung MS., Garcia AE., Onuchic J. (2000). "Protein folding mediated by solvation: water expulsion and formation of the hydrophobic core occur after structural collapse". Proc Natl Acad Sci USA 99:685-690.

18. Kaya H., Chan HS. (2003). "Solvation effects and driving forces for protein thermodynamics and kinetic cooperativity: how adequate is native-centric topological modeling?". J Mol Biol 326:911-931.

19. Das P., Wilson CJ., Fossati G., Wittung-Stafshede P., Matthews KS., Clementi C. (2005). "Characterization of the folding landscape of monomeric lactose repressor: quantitative comparison of theory and experiment". Proc Natl Acad Sci USA 102:14569-14574.

20. Levy Y., Wolynes PG., Onuchic JN. (2004). "Protein topology determines binding mechanism". Proc Natl Acad Sci USA 101:511-516.

21. Lu Q., Lu PH., Wang J. (2007). "Exploring the mechanism of flexible biomolecular recognition with single molecule dynamics". Phys Rev Lett 98:128105.

22. Hyeon C., Onuchic JN. (2007). "Internal strain regulates the nucleotide binding site of the kinesin leading head". Proc Natl Acad Sci USA 104:2175-2180.

23. Koga N., Takada S. (2006). "Folding-based molecular simulations reveal mechanisms of the rotary motor F1-ATPase". Proc Natl Acad Sci USA 103:5367-5537.

24. Levitt M. (1976). "A simplified representation of protein conformations for rapid simulation of protein folding". J Mol Biol 104(1):59-107.

25. Derreumaux P. (1999). "From polypeptide sequences to structures using Monte Carlo simulations and an optimized potential". J Chem Phys 111:2301.

26. DeWitte R., Shakhnovich E. (1994). "Pseudodihedrals: simplified protein backbone representation with knowledge-based energy". Protein Sci 3:1570

27. Haliloglu T., Bahar I. (1998). "Coarse-grained simulations of conformational dynamics of proteins: application to apomyoglobin". Proteins 31:271

28. Srinivasan R., Rose GD. (1999). " A physical basis for protein secondary structure". Proc Natl Acad Sci USA 96:14258.

29. Van Giessen A., Straub J. (2006). "Coarse-grained model of coil-to-helix kinetics demonstrates the importance of multiple nucleation sites in helix folding". J Chem Theory Comput 2:674.

30. Voegler Smith A., Hall C. (2001). "Alpha-helix formation: discontinuous molecular dynamics on an intermediate-resolution protein model". Proteins 44:344.

31. Hoang TX., Seno F., Banavar JR., Cieplak M., Maritan A. (2003). "Assembly of protein tertiary structures from secondary structures using optimized potentials". Proteins 52:155–165.

32. DeWitte R., Shakhnovich E. (1994). "Pseudodihedrals: simplified protein backbone representation with knowledge-based energy". Protein Sci 3:1570.

33. Liwo A., Pincus M., Wawak R., Rackovsky S., Scheraga H. (1993). "Prediction of protein conformation on the basis of a search for compact structures: test on avian pancreatic polypeptide". Protein Sci 2:1715.

34. Ha-Duong T. (2010). "Protein backbone dynamics simulations using coarse-grained bonded potentials and simplified hydrogen bonds". J Chem Theory Comput 6:76.

35. Tozzini V., Trylska J., Chang C-e., McCammon JA. (2007). " Flap opening dynamics in HIV-1 protease explored with a coarse-grained model". J Struct Biol 157:606.

36. Anfinsen C. (1973). "Principles that govern the folding of protein chain". Science 181:223–230.

37. Scheraga HA., Liwo A., Ołdziej S., Czaplewski C., Pillardy J., Ripoll DR., Vila JA., Kazmierkiewicz R., Saunders JA., Arnautova YA., Jagielska A., Chinchio M., Nanias M. (2004). "The protein folding problem: global optimization of force fields". Front Biosci 9:3296–3323.

38. Liwo A., Oldziej S., Pincus MR., Wawak RJ., Rackovsky S., Scheraga HA. (1997). "A united-residue force field for off-lattice protein-structure simulations. I. Functional forms and parameters of long-range side-chain nteraction potentials from protein crystal data". J Comput Chem 18:849–873.

39. Maisuradze GG., Senet P., Czaplewski C., Liwo A., Scheraga HA. (2010). "Investigation of protein folding by coarse-grained molecular dynamics with the UNRES force field". J Phys Chem A 114: 4471–4485.

40. Oldziej S., Czaplewski C., Liwo A., Chinchio M., Nanias M., Vila JA., Khalili M., Arnautova YA., Jagielska A., Makowski M. et al (2005). "Physics-based proteinstructure prediction using a hierarchical protocol based on the UNRES force field: assessment in two blind tests". Proc Natl Acad Sci USA 102:7547–7552.

41. He Y., Liwo A., Weinstein H., Scheraga HA. (2011). "PDZ binding to the BAR domain of PICK1 is elucidated by coarse-grained molecular dynamics". J Mol Biol 405:298–314.

42. Rojas A., Liwo A., Browne D., Scheraga HA. (2010). "Mechanism of fiber assembly: treatment of a beta peptide aggregation with a coarse-grained united-residue force field". J Mol Biol 404:537–552.

43. Oldziej S., Czaplewski C., Liwo A., Chinchio M., Nanias M., Vila JA., Khalili M., Arnautova YA., Jagielska A., Makowski M. et al (2005). "Physics-based proteinstructure prediction using a hierarchical protocol based on the UNRES force field: assessment in two blind tests". Proc Natl Acad Sci USA 102:7547–7552.

44. Liwo A., Khalili M., Scheraga HA. (2005). "Ab initio simulations of protein-folding pathways by molecular dynamics with the united-residue model of polypeptide chains". Proc Natl Acad Sci USA 102:2362–2367.

45. Rojas AV., Liwo A., Scheraga HA. (2007). "Molecular dynamics with the united-residue force field: ab initio folding simulations of multichain proteins". J Phys Chem B 111:293–309

46. Sukharev SI., Blount P., Martinac B., Blattner FR., Kung C. (1994). "A large-conductance mechanosensitive channel in E.coliencoded by mscL alone". Nature 368:265–268.

47. Yefimov S., van der Giessen E., Onck PR., Marrink SJ. (2008). "Mechanosensitive membrane channels in action". Biophys J 94:2994–3002.

48. Ollila OHS., Louhivuori M., Marrink SJ., Vattulainen I. (2011). "Protein shape change has a major effect on the gating energy of a mechanosensitive channel". Biophys J 100:1651–1659.

49. Lennard-Jones, J. E. (1931). "Cohesion". Proceedings of the Physical Society 43, 461-482.

50. Verstrepen K., Jansen A., Lewitter F., Fink G. (2005). "Intragenic tandem repeats generate functional variability". Nat Genet 37(9):986-90.

51. Phatnani H., Greenleaf A. (2006). "Phosphorylation and functions of the RNA polymerase II CTD". Genes Dev 20:2922-2936.

52. S. Istrail., F. Lam (2009). "Combinatorial algorithms for protein folding in lattice models: a survey of mathematical results". Communications in Information and Systems, vol. 9, no. 4, article 303.

53. Minton AP. (1992). "Confinement as a determinant of macromolecular structure and reactivity". Biophysical Journal 63 (4): 1090–100.

54. Orsi N. (2004). "The antimicrobial activity of lactoferrin: current status and perspectives. Biometals" 17: 189-96.

55. Drenth J. (1994). "Principles of protein X-ray crystallography". Springer-Verlag, Nueva York.

56. Evans, J. N. S. (1995). "Biomolecular NMR spectroscopy:a multifaceted approach to macromolecular structure". Quarter Rev Biophys 33: 29-65.

57. Van Heel, M. et al (2000). "Single-particle electron cryo-microscopy: towards atomic resolution". Q. Rev Biophys 33: 307-36

58. Glusker J.P. (1994). "X-ray crystallography of proteins Methods Biochem". Anal. 37: 1-72.

59. Acharya KR., Lloyd MD. (2005). "The advantages and limitations of protein crystal structures". Trends Pharmacol Sci 26(1):10-4.

60. Tucker T., Marra M., Friedman JM. (2009). "Massively Parallel Sequencing: The Next Big Thing in Genetic Medicine". The American Journal of Human Genetics 85 (2): 142–154.

61. Martí-Renom MA., Stuart AC., Fiser A., Sánchez R., Melo F., Sali A. (2000). **"**Comparative Protein Structure Modeling of Genes and Genomes". Annual Review of Biophysics and Biomolecular Structure 29: 291-325.

62. Chothia C., Lesk A.M. (1986). “The use of sequence homologies to predict protein structures”. Computer graphics and molecular modeling. Cold Spring, Harbor Laboratory, New York, pp 33-37.

63. Samudrala R., Xia Y., Huang ES., Levitt M. (1999). “Ab initio prediction of protein structure using a combined hierarchical approach”. Proteins: Structure, Function, and Genetics S3: 194-198.

64. Zhang Y., Skolnick J. (2005). "The protein structure prediction problem could be solved using the current PDB library". Proc Natl Acad Sci USA 102 (4): 1029–34.

65. Hardin C1., Pogorelov TV., Luthey-Schulten Z. (2002). “Ab initio protein structure prediction”. Curr Opin Struct Biol 12(2):176-81.

66. Karplus M., McCammon JA. (2002). “Molecular dynamics simulations of biomolecules”. Nature Structural Biology 9:788.

67. Tozzini V. (2005). “Coarse-grained models for proteins”. Current Opinion in Structural Biology 15:144–150.

68. Taketomi H., Ueda Y. & Gō N. (1975). “Studies on protein folding, unfolding and fluctuations by computer simulation. I. The effect of specific amino acid sequence represented by specific inter-unit interactions”. International journal of peptide and protein research 7(6):445–59.

69. Alder BJ., Wainwright TE. (1959). "Studies in Molecular Dynamics. I. General Method". J. Chem. Phys. 31 (2): 459.

70. Levitt M. , Warshel A. (1975). "Computer Simulations of Protein Folding". Nature 253 (5494): 694–8.

71. Berman, H. M. (2008). "The Protein Data Bank: a historical perspective". Acta Crystallographica Section A: Foundations of Crystallography A64 (1): 88–95.

72. Siewert J. Marrink (2007). “The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations”. J. Phys. Chem. B 111, 7812-7824.

73. Singer SJ., Nicolson GL. (1972). “The Fluid Mosaic Model of the Structure of Cell Membranes”. Science 175(4023) :720-731.

74. Wootton JC., (1994). “Non-globular domains in proteína sequences:automated segmentations using complexity mesures”. Comput. Chem. 18:269-285.

75. Mark A. DePristo, Martine M., Daniel L. (2006). “On the abundance, amino acid composition, and evolutionary dynamics of low-complexity regions in proteins”. Gene 378: 19–30

76. Hynes RO., (2002). "Integrins: Bidirectional, Allosteric Signaling Machines". Cell, 110(6):673–687.

77. Leitner DM. (2008). "Energy Flow in Proteins". Annual Review of Physical Chemistry 59: 233-259.

78. Ha-Duong T. (2014). "Coarse-grained models of the proteins backbone conformational dynamics". Adv Exp Med Biol 805:157-69.

79. Liwo A. (2011). "Coarse-grained force field: general folding theory". Phys Chem Chem Phys 13:16890–16901.

80. Ingólfsson Helgi I., Lopez Cesar A., Uusitalo Jaakko J., Jong Djurre H. de, Gopal Srinivasa M., Periole Xavier, Marrink Siewert J. (2014). "The power of coarse graining in biomolecular simulations". WIREs Comput Mol Sci 4: 225-248.

81. Klingelhoefer JW., Carpenter T., Sansom MSP. (2009). "Peptide nanopores and lipid bilayers: interactions by coarse-grained molecular-dynamics simulations". Biophys J 96:3519–3528.

82. Balali-Mood K., Bond PJ., Sansom MSP. (2009). "Interaction of Monotopic Membrane Enzymes with a Lipid Bilayer: A Coarse-Grained MD Simulation Study". Biochemistry 48:2135-2145.

83. Lumb CN., Sansom MSP. (2013). "Defining the Membrane-Associated State of the PTEN Tumor Suppressor Protein". Biophys J 104(3):613-621.

84. Qin SS., Yu YX., Li QK., Yu ZW. (2013). "Interaction of Human Synovial Phospholipase A2 with Mixed Lipid Bilayers: A Coarse-Grain and All-Atom Molecular Dynamics Simulation Study". Biochemistry 52(8): 1477-1489.

85. Tjörnhammar R., Edholm O. (2013). "The shape and free energy of a lipid bilayer surrounding a membrane inclusion". Chem. Phys. Lipids 169:2-8.

86. Michelitsch MD., Weissman JS. (2000). "A census of glutamine/asparagine-rich regions: implications for their conserved function and the prediction of novel prions". Proc Natl Acad Sci USA 97(22):11910–11915.

87. Dunker A., Lawson J., Brown C., Williams R., Romero P., Oh J., Oldfield C., Campen A., Ratliff C., Hipps K., Ausio J., Nissen M., Reeves R., Kang C., Kissinger C., Bailey R., Griswold M., Chiu W., Garner E., Obradovic Z. (2001). "Intrinsically disordered protein". J. Mol Graph Model 19(1):26–59.

88. Kemp DJ., Coppel RL., Anders RF. (1987). "Repetitive proteins and genes of malaria". Annu Rev Microbiol 41:181–208.

89. Bocola M., Schwaneberg U, Jaeger KE., Krauss U. (2015). "Light-induced structural changes in a short light, oxygen, voltage (LOV) protein revealed by molecular dynamics simulations—implications for the understanding of LOV photoactivation". Front Mol Biosci 2:55.

90. Periole X., Huber T., Marrink SJ., Sakmar TP. (2007). "G protein-coupled receptors self-assemble in dynamics simulations of model bilayers". J Am Chem Soc 129:10126–10132.

91. Arnarez C., Marrink SJ., Periole X. (2013). "Identification of cardiolipin binding sites on cytochrome c oxidase at the entrance of proton channels". Sci Rep 3:1263.

92. Periole X., Knepp AM., Sakmar TP., Marrink SJ., Huber T. (2012). "Structural determinants of the supra-molecular organization of G protein-coupled receptors in bilayers". J Am Chem Soc 134:10959–10965.

93. Domanski J., Marrink SJ., Schafer LV. (2012). "Transmembrane helices can induce domain formation in crowded model biomembranes". Biochim Biophys Acta—Biomembr 1818:984–994.

94. Blundell TL., Johnson LN. (1976). "Protein crystallography". London: Academic Press.

95. Wüthrich K. (2001). "The way to NMR structures of proteins". Nature Structural & Molecular Biology 8 (11): 923–5.

96. Taketomi H., Ueda Y., Gō, N. (1975). "Studies on protein folding, unfolding and fluctuations by computer simulation. I. The effect of specific amino acid sequence represented by specific interunit interactions". International journal of peptide and protein research 7(6):445–59.

97. Barnoud J., Monticelli L. (2015). "Coarse-grained force fields for molecular simulations". Methods Mol Biol. 1215:125-49.

98. Ramachandran G.N., Ramakrishnan C., Sasisekharan V. (1963). "Stereochemistry of polypeptide chain configurations". Journal of Molecular Biology **7**: 95–9.

99. Levitt M. (1976). "A simplified representation of protein conformations for rapid simulation of protein folding". J Mol Biol 104(1):59-107.

100. Park B., Levitt M. (1995). "The Complexity and Accuracy of Discrete State Models of Protein Structure". J Mol Biol 249:493-507.

101. Xavier F de la Cruz, Michael W Mahoney and Byungkook Lee (1997). "Discrete representations of the protein Cα chain". Folding & Design 2:223–234.

102. Metropolis N., Rosenbluth AW. Rosenbluth MN., Teller AH., Teller E. (1953). "Equations of State Calculations by Fast Computing Machines". Journal of Chemical Physics 21(6):1087–1092.

103. Minary P., Levitt M (2010). "Conformational Optimization with Natural Degrees of Freedom: A Novel Stochastic Chain Closure Algorithm". Journal of Computational Biology 17(8):993–1010.

104. Baker D (2000). "A surprising simplicity to protein folding". Nature 405:39-42.

105. DeWitte RS., Shakhnovich EI. (1994). "Pseudodihedrals: Simplified protein backbone representation with knowledge-based energy". Protein Science 3:1570-1581.

106. Rooman MJ., Kocher JA., Wodak SJ. (1991). "Prediction of protein backbone conformation based on seven structure assignments: Influence of local interactions". J Mol Biol 221:961-979.

107. Ramachandran G.N., Ramakrishnan C., Sasisekharan V. (1963). "Stereochemistry of polypeptide chain configurations". Journal of Molecular Biology 7: 95–9.

108. Bromberg S., Dill KA. (1994). "Side chain entropy and packing in proteins". Prot Sci 3(7):997–1009.

109. Hart WE., Istrail S. (1997). "Lattice and off-lattice side chain models of protein folding: Linear time structure prediction better than 86% of optimal". Journal of Computational Biology 4(3):241–259.

110. Volker Heun (2003). "Approximate protein folding in the HP side chain model on extended cubic lattices". Discrete Applied Mathematics 127(1):163–177.

111. Unger R., Harel D., Wherland,S., Sussman JL. (1989). "A 3D building blocks approach to analyzing and predicting structure of proteins". Proteins 5(4):355-73.

112. Gordon H., Somorjai R. (1992). "Fuzzy cluster analysis of molecular dynamics trajectories". Proteins 14(2), 249–6.

113. Wallin S., Farwer J., Bastolla U. (2003). "Testing similarity measures with continuous and discrete protein models". Proteins: Structure, Function, and Bioinformatics 50 (1):144-157.

114. Holm L., Sander C. (1993). "Protein structure comparison by alignment of distance matrices". J Mol Biol 233(1):123-38.

115. Vendruscolo M., Subramanian B., Kanter I., Domany E., Lebowitz J. (1999). "Statistical properties of contact maps". Phys. Rev. E 59:977.

116. Koehl P. (2001). "Protein structure similarities". Curr Opin Struct Biol 11 (3):348-53.

117. Kabsch W. (1976) "A solution for the best rotation to relate two sets of vectors". Acta Crystallographica **32**:922.

118. Wootton JC. (1994). "Non-globular domains in proteína sequences:automated segmentations using complexity mesures". Comput. Chem 18:269-285.

119. Green H., Wang N. (1994). "Codon reiteration and the evolution of proteins". Proc Natl Acad Sci USA 91:4298–4302.

120. Alba MM., Tompa P., Veitia RA. (2007). "Amino acid repeats and the structure and evolution of proteins". Genome Dyn. 3:119–130.

121. Ellegren H. (2004). "Microsatellites: simple sequences with complex evolution". Nat Rev Genet. 5:435–445.

122. Brown LY., Brown SA. (2004). "Alanine tracts: the expanding story of human illness and trinucleotide repeats". Trends Genet 20:51–58.

123. Kashi Y., King DG. (2006). "Simple sequence repeats as advantageous mutators in evolution". Trends Genet 22:253–259.

124. Xiao H., Jeang KT. (1998). "Glutamine-rich domains activate transcription in yeast Saccharomyces cerevisiae". J Biol Chem 273:22873–22876.

125. Salichs E., Ledda A., Mularoni L., Alba MM., de la Luna S. (2009). "Genome-wide analysis of histidine repeats reveals their role in the localization of human proteins to the nuclear speckles compartment". PLoS Genet 5(3):e1000397.

126. Michelitsch MD., Weissman JS. (2000). "A census of glutamine/asparagine-rich regions: implications for their conserved function and the prediction of novel prions". Proc Natl Acad Sci USA 97(22):11910–11915.

127. Dunker A., Lawson J., Brown C., Williams R., Romero P., Oh J., Oldfield C., Campen A., Ratliff C., Hipps K., Ausio J., Nissen M., Reeves R., Kang C., Kissinger C., Bailey R., Griswold M., Chiu W., Garner E., Obradovic Z. (2001). "Intrinsically disordered protein". J. Mol Graph Model 19(1):26–59.

128. Pizzi E., Frontali C. (2001). "Low-complexity regions in Plasmodium falciparum proteins". Genome Res. 11:218–229.

129. Freitas-Junior LH., Bottius E., Pirrit LA., Deitsch KW., Scheidig C., Guinet F., Nehrbass U., Wellems TE., Scherf A.. (2000). "Frequent ectopic recombination of

virulence factor genes in telomeric chromosome clusters of P. falciparum". Nature 407: 1018–1022.

130. Kemp DJ., Coppel RL., Anders RF. (1987). "Repetitive proteins and genes of malaria". Annu Rev Microbiol 41:181–208.

131. Hughes AL. (2004). "The evolution of amino acid repeat arrays in Plasmodium and other organisms". J Mol Evol 59:528–535.

132. Wooton C., Federhen S. (1993). "Statistics of local complexity in aminoacid sequences and sequence database". Computers and Chemistry **17**(2):149-163.

133. Dyson HJ., Wright PE. (2005). "Intrinsically unstructured proteins and their function". Nature Reviews Molecular Cell Biology 6:197-208.

134. Fiser A., Do RK., Sali A. (2000). "Modeling of loops in protein structures". Protein Science 9: 1753-1773.

135. Dominik G., Sebastian K., Andrzej K. (2007). "Backbone building from quadrilaterals". Journal of Computational Chemistry 28:1593–1597.

136. Georgii G., Maxim V., Roland L. (2009). "Improved prediction of protein side-chain conformations with SCWRL4". Proteins 77(4): 778–795.

137. Sean R. (2004). "Where did the BLOSUM62 alignment score matrix come from?". Nature Biotechnology 22,:1035-1036

138. Zhang J., Zhang Y. (2010). "A Novel Side-Chain Orientation Dependent Potential Derived from Random-Walk Reference State for Protein Fold Selection and Structure Prediction". PLoS ONE 5(10): e15386.

139. McLachlan A.D. (1982). "Rapid Comparison of Protein Structres''. Acta Cryst A38: 871-873.

140. Alonso PL. et al (2004). "Efficacy of the RTS, S/AS02A vaccine against Plasmodium falciparum infection and disease in young African children: randomised controlled trial". The Lancet 364(9443):1411-20.

141. Wootton JC., Federhen S. (1993). "Statistics of Local Complexity in Amino Acid Sequences and Sequence Databases". Comput Chem 17:149-163.

142. Case DA. (2002). "Molecular dynamics and NMR spin relaxation in proteins". Acc Chem Res 35:325–331.

143. Meinhold L., Smith JC. (2005). "Fluctuations and correlations in crystalline protein dynamics: a simulation analysis of staphylococcal nuclease". Biophys J 88:2554 –2563.

144. Berg J., Tymoczko J., Stryer L. (2006). "Biochemistry". W. H. Freeman, New York, 6th Ed.

145. Gunasekaran B., Ma B., Nussinov R. (2004). "Is allostery an intrinsic property of all dynamic proteins?". Proteins Struct Funct Genet 57:443–453.

146. Krauss G. (2003). "Biochemistry of Signal Transduction and Regulation". Wiley, Weinheim, Germany, 3rd Ed.

147. Clarage JB., Romo, T., Andrews BK., Pettitt BM., Phillips JGN. (1995). "A sampling problem in molecular dynamics simulations of macromolecules". Proc. Natl. Acad. Sci USA 92: 3288-3292.

148. Barbany M. Meyer T. Hospital A. Faustino I. D'Abramo M. Morata J. et al. (2015). "Molecular Dynamics Study of Naturally Existing Cavity Couplings in Proteins". PLoS ONE 10(3): e0119978.

149. Leitner DM. (2008). "Energy Flow in Proteins". Annual Review of Physical Chemistry 59:233-259.

150. Lange OF., Grubmüller H. (2006). "Generalized correlation for biomolecular dynamics". Proteins. 62(41):1053–10.

151. Lange OF., Grub H. (2008). "Full correlation analysis of conformational protein dynamics". Proteins 70(4):1294–1312.

152. Kassem S. (2015). "Entropy in bimolecular simulations: A comprehensive review of atomic fluctuations-based methods". Journal of Molecular Graphics and Modelling 62:105-117.

153. Hünenberger PH., Mark AE. (1995). "Fluctuation and Cross-correlation Analysis of Protein Motions Observed in Nanosecond Molecular Dynamics Simulations". Journal of Molecular Biology 252(4):492-503.

154. Karplus M., McCammon JA. (200). "Molecular dynamics simulations of biomolecules". Nature Structural Biology 9:646–652.

155. Minton AP. (1997). "Influence of excluded volume upon macromolecular structure and associations in 'crowded' media". Current Opinion in Biotechnology 8:65-69.

156. Massa W. (2004). "Crystal Structure Determination". Berlin: Springer. ISBN 3-540-20644-2.

157. Brunger AT. (1992). "Free R value: a novel statistical quantity for assessing the accuracy of crystal structures". Nature. 355 (6359):472–475.

158. Bahar I., Erman B. (1997). "Efficient Characterization of Collective Motions and Interresidue Correlations in Proteins by Low-Resolution Simulations". Biochemistry 36:13512-13523.

159. Gohlke H., Hendlich M., Klebe G. (2000). "Knowledge-based Scoring Function to Predict Protein-Ligand Interactions". J. Mol. Biol. 295:337–356.

160. Sippl M. J. (1990). "Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins". J Mol Biol 213(4):859-883.

161. Sippl MJ. (1993). "Boltzmann's principle, knowledgebased mean fields and protein folding. An approach to the computational determination of protein structures". J Comput Aided Mol Des 7(4):473-501

162. McCampbell. A., Fischbeck K. H. (2001). "Polyglutamine and CBP: fatal attraction?". Nature Med. **7**, 528–530 .

163. Karlin S., Brocchieri L., Bergman A., Mrazek J., Gentles A. J. (2002). "Amino acid runs in eukaryotic proteomes and disease associations". Proc. Natl Acad. Sci. USA **99**, 333–338.

164. Uversky V. N., Oldfield C. J., Dunker, A. K. (2008). "Intrinsically Disordered Proteins in Human Diseases: Introducing the D2 Concept". Annual Review of Biophysics (37): 215-246.

165. Dunker A.K., Brown C.J., Lawson J.D., Iakoucheva L.M., Obradovic Z. (2002). "Intrinsic disorder and protein function". Biochemistry 41:6573–82.

166. Romero PR., Zaidi S., Fang YY., Uversky VN., Radivojac P., et al. (2006)." Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms". Proc. Natl. Acad. Sci. USA 103:8390–95.

167. Xie H, Vucetic S, Iakoucheva LM, Oldfield CJ, Dunker AK, et al. (2007). "Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions". J. Proteome Res. 6:1882–98.

168. Fenwick B., Orellana l., Esteban-Martín S, Orozco M., Salvatella X. (2014). "Correlated motions are a fundamental property of β-sheets". Nature Communications 5:4070.

169. Lange OF., Grubmüller H. (2006). "Generalized Correlation for Biomolecular Dynamics". PROTEINS: Structure, Function, and Bioinformatics 62:1053–1061.

170. Zhang Y. (2008).”Progress and challenges in protein structure prediction”. Curr. Opin. Struct. Biol. 18(3):342–348.

171. Klecknera I.R., Foster M.P. (2010). "An introduction to NMR-based approaches for measuring protein dynamics".Biochim. Biophys. Acta. 1814(8):942–968.

172. Yue W.W.., Froese D.S., Brennan P.E. (2014). "The role of protein structural analysis in the next generation sequencing era". Top. Curr. Chem. 336:67-98.