ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE CIENCIAS

PROCESOS FÍSICOS EN SISTEMAS BIOMOLECULARES:

TRANSICIÓN CONFORMACIONAL DEL LAZO 36 DE LA PROTEÍNA HEMAGLUTININA

TRABAJO DE INTEGRACIÓN CURRICULAR PRESENTADO COMO REQUISITO PARA LA OBTENCIÓN DEL TÍTULO DE FÍSICO

KEVIN VINICIO CÁRDENAS CORRALES

kevinicio.cardenas@gmail.com

DIRECTOR: MARCO VINICIO BAYAS REA

marco.bayas@epn.edu.ec

DMQ, AGOSTO 2023

CERTIFICACIONES

Yo, KEVIN VINICIO CÁRDENAS CORRALES, declaro que el trabajo de integración curricular aquí descrito es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

KEVIN VINICIO CÁRDENAS CORRALES

Certifico que el presente trabajo de integración curricular fue desarrollado por KEVIN VINICIO CÁRDENAS CORRALES, bajo mi supervisión.

MARCO VINICIO BAYAS REA
DIRECTOR

DECLARACIÓN DE AUTORÍA

A través de la presente declaración, afirmamos que el trabajo de integración curricular aquí descrito, así como el(los) producto(s) resultante(s) del mismo, es(son) público(s) y estará(n) a disposición de la comunidad a través del repositorio institucional de la Escuela Politécnica Nacional; sin embargo, la titularidad de los derechos patrimoniales nos corresponde a los autores que hemos contribuido en el desarrollo del presente trabajo; observando para el efecto las disposiciones establecidas por el órgano competente en propiedad intelectual, la normativa interna y demás normas.

KEVIN VINICIO CÁRDENAS CORRALES

MARCO VINICIO BAYAS REA

RESUMEN

El lazo 36 es un polipéptido que forma parte de la cadena HA2 de la hemagluitinina, proteína que se encuentra en la superficie del virus de la influeza. La transición del polipéptido desde su estado parcialmente aleatorio hasta su estado totalmente plegado fue estudiada mediante la caracterización del cambio de su energía libre en función de una coordenada de reacción, obteniendo un potencial de fuerza media (PMF). La estrategia para el cálculo del PMF combina una simulación TMD para obtener un primer muestreo del espacio conformacional del polipéptido, análisis de grupos para obtener estructuras representativas de la transición y muestreo *Umbrella* para obtener información completa del espacio conformacional del lazo 36. El criterio utilizado para determinar la similitud entre grupos fue definir una resolución de muestreo asociada al ancho de un pico de la curva PMF obtenida de la simulación TMD. Con este protocolo, se obtuvieron 10 estructuras representativas que fueron utilizadas para crear las divisiones del muestreo *Umbrella*. Así, se obtuvo el PMF asociado a la transición natural del lazo 36, en donde se halló que para que el sistema llegue a su estado totalmente plegado, debe atravesar una barrera de potencial de una altura igual a 2.5 [kcal/mol] cuando la distancia entre centros de masa de la parte inferior y la parte superior del polipéptido es igual a 25.4 [Å] y se halló que el mínimo de energía global ocurre cuando esta distancia es igual a 35.4 [Å].

Palabras clave: Hemaglutinina, Potencial de fuerza media, Lazo 36, Muestreo Umbrella, Dinámica molecular, Analisis de grupos.

ABSTRACT

Loop 36 is a polypeptide that is part of the HA2 chain of hemagglutinin, a protein found on the surface of influenza virus. The transition of the polypeptide from its partially random state to its fully folded state was studied by characterizing the change of its free energy as a function of a reaction coordinate, obtaining a potential mean force (PMF). The strategy for calculating the PMF combines a TMD simulation to obtain a first sampling of the conformational space of the polypeptide, cluster analysis to obtain representative structures of the transition and the *Umbrella* sampling method to obtain complete information of the conformational space of the loop 36. The criteria used to determine the similarity between clusters was to define a sampling resolution associated to the width of a peak of the PMF curve obtained from the TMD simulation. With this protocol, 10 representative structures were obtained and used as sampling windows. Thus, the PMF associated with the natural transition of loop 36 was obtained, here it was found that for the system to reach its fully folded state, it must cross a potential barrier of a height equal to 2.5 [kcal/mol] when the distance between centers of mass of the lower and upper part of the polypeptide is equal to 25.4 [Å] and it was found that the global energy minimum occurs when this distance is equal to 35.4 [Å].

Keywords: Influenza hemagglutinin, Potential of mean force, Loop 36, Umbrella Sampling, Molecular dynamics, Clusters analysis.

Índice general

Índice de figuras

Capítulo 1

Descripción del componente desarrollado

1.1. Objetivo general

Entender el comportamiento estructural relacionado con la transición experimentada por el lazo 36 de la Hemaglutinina durante su transición desde su estado parcialmente aleatorio hacia su estado totalmente plegado.

1.2. Objetivos específicos

1. Identificar estructuras intermedias asociadas al cambio conformacional del lazo 36.

2. Obtener el potencial de fuerza media asociado a la transición del lazo 36.

1.3. Alcance

La Hemaglutinina (HA) es una proteína que forma parte del virus de la influenza. Durante el proceso de infección, la cadena HA2, una de las subunidades de la HA, sufre un cambio conformacional. El polipéptido conocido como *lazo 36* que forma parte de la cadena HA2 experimenta el cambio conformacional más drástico de esta subunidad. Para este tipo de sistemas biomoleculares, como el lazo 36, es de interés obtener información energética de los cambios conformacionales con el fin de

asociar posibles barreras energéticas a estructuras específicas. Sin embargo, experimentos detalladas en los que se obtenga la información energética para los cambios conformacionales a nivel atómico resultan muy limitados. Por tanto, cálculos teóricos o computacionales de energías libres son de gran importancia para comprender eventos determinantes de la dinámica de este tipo de procesos.

El presente estudio busca explorar el comportamiento estructural asociado al proceso de infección mediante el estudio computacional del las características energéticas de la transición del lazo 36. Se construirá el potencial de fuerza media (PMF) asociado a la transición del lazo 36 desde su estructura parcialmente aleatoria, hasta su estructura helicoidal, mediante simulaciones de dinámica molecular con lo cual se caracterizarán las diferentes configuraciones accesibles al polipéptido y consecuentemente su cambio en energía libre en función de una coordenada de reacción específica.

La estrategia para la obtención del PMF combina una simulación de dinámica molecular dirigida TMD en donde se utilizará un archivo de trayectoria provisto por el laboratorio de biofísica, un análisis de grupos ejecutado mediante el software de uso libre R y simulaciones de dinámica molecular realizadas mediante el software NAMD y VMD en las que se utilizará un archivo de campos de fuerzas CHARMM22 y se estudiará al sistema durante un total de 10 [ns].

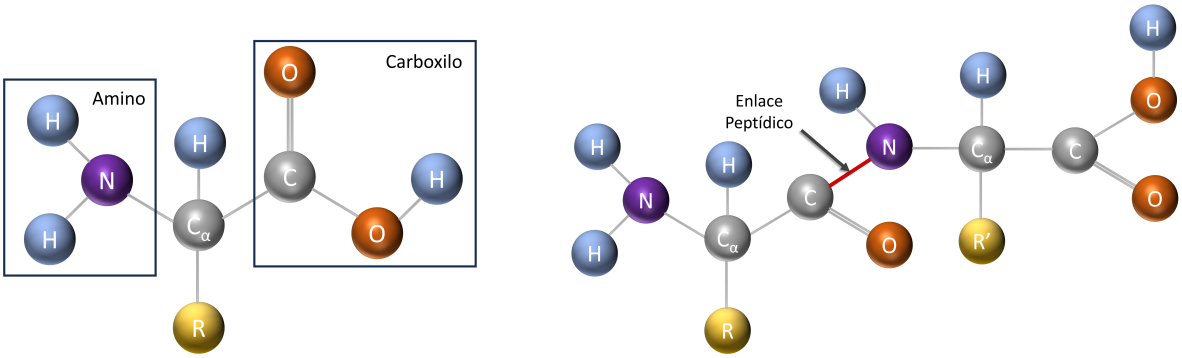
1.4. Marco teórico

1.4.1. Proteínas

Los polímeros son cadenas largas conformadas por subunidades, o monómeros, unidas por enlaces covalentes. A las cadenas conformadas por monómeros de aminoácidos se les conoce como polipéptidos y a los polipéptidos con funcionalidades biológicas se los conoce como proteínas. Las cadenas largas y sueltas de aminoácidos son inestables, por tanto en un inicio las proteínas buscan plegarse hasta obtener una estructura tridimensional estable [1].

Todos los aminoácidos poseen una estructura común que incluye un átomo de carbono central denominado *carbono* α, o simplemente C_α o CA, rodeado por: un átomo de hidrógeno, un grupo amino (α-amino), un grupo carboxilo (α-carboxilo) y un cuarto grupo denominado *cadena lateral* (R). Es habitual dividir cada aminoácido

en dos partes. La primera incluye todos los átomos no laterales de la cadena, es decir: CA, su átomo de hidrógeno, el grupo α-carboxilo y el grupo α-amino. Esta parte se denomina *backbone*, y es idéntica en todos los aminoácidos. La segunda parte consiste en la cadena lateral, (R) en la figura 1.1, que es diferente para cada aminoácido y por tanto es lo que diferencia a cada aminoácido [2]. Es usual hacer esta división para un conjunto de aminoácidos. Por tal razón, a partir de ahora se denominará *backbone* a todo el conjunto de una cadena de aminoácidos, cada uno sin su cadena lateral. Es decir, el *backbone* de un polipéptido es el conjunto de sus átomos CA, H, $\alpha-$carboxilo y $\alpha-$amino unidos por enlaces peptídicos.



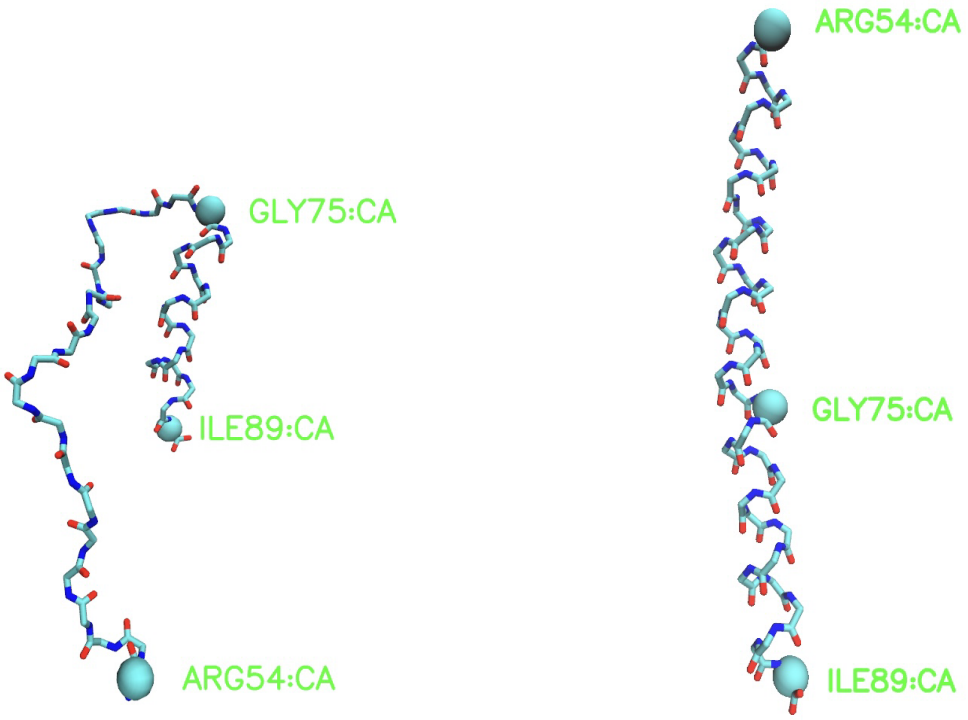
(a) (b)

Figura 1.1: (a) Estructura de un aminoácido: un carbono central denominado carbono α unido a un grupo amino (NH_2), a un grupo carboxilo (COOH), a un hidrógeno (H) y a una cadena lateral (R). (b) Esquema de un enlace peptídico (color rojo).

Los aminoácidos se encuentran unidos en péptidos y proteínas mediante un enlace amida denominado *enlace peptídico*. Estos enlaces se forman en una reacción de condensación entre el grupo carboxilo de un aminoácido y el grupo amino de otro, liberando en el proceso una molécula de agua (figura 1.1). Lo que queda del aminoácido como parte del péptido o proteína se denomina *residuo* [3].

1.4.2. Lazo 36

La hemaglutinina (HA) es una proteína que forma parte del virus de la influenza. La proteína contiene dos subunidades principales: la cadena HA1 y la cadena HA2. Cuando el virus infecta a una célula objetivo, la cadena HA2 sufre un cambio conformacional el cual ha sido estudiado experimentalmente. Por tanto, las estructuras que adapta la cadena al inicio y al final del cambio conformacional son conocidas [5]. El sistema que se estudió se trata del polipéptido conocido como lazo 36 que abarca



(a) (b)

Figura 1.2: *Backbone* del lazo 36 (a) al inicio de su plegamiento y (b) en los últimos instantes del cambio conformacional. La figura fue obtenida mediante el programa VMD [4].

los 36 residuos de la cadena HA2 de la Hemaglutinina desde la Arginina 54 (ARG54) hasta la Isoleucina 89 (ILE89) [6]. En la figura 1.2 se muestra las estructuras del *backbone* del lazo 36 en su conformación inicial, parcialmente aleatoria, y final, en su conformación helicoidal. Las esferas de mayor tamaño y de color celeste representan a los carbonos alfa (CA) de los residuos ubicados en posiciones relevantes a la transición del polipéptido: ARG54, GLY75 (Glicina) y ILE89.

1.4.3. Espacio conformacional

Cuando el lazo 36 evoluciona desde su estado parcialmente aleatorio hacia su estado completamente plegado, este visita estados o estructuras intermedias. A una estructura se la describe mediante un conjunto de coordenadas espaciales de todos los átomos que la conforman. Cada estructura accesible al sistema se trata de un punto en un espacio $3-N$ dimensional asociado a las coordenadas de los N átomos del polipéptido en estudio. A este espacio se lo denominará *espacio conformacional*

y se lo notará con la letra ω.

1.4.4. Potencial de fuerza media (PMF)

El potencial de fuerza media (PMF) hace referencia a un perfil de energía libre a lo largo de una coordenada de reacción en particular. De principios de mecánica estadística se conoce que a partir de la función de partición canónica, Q, se puede calcular la energía libre, A.

El cálculo de Q para un sistema en particular puede realizarse mediante una integral sobre todo el espacio de fases, es decir, el espacio conformacional y el espacio de momentos. Si la energía potencial, V, es independiente de los momentos entonces la integral con respecto a esta variable es una constante multiplicativa de Q, que para efectos prácticos puede ignorarse. Entonces, Q se obtiene como

$$Q = \int e^{-\beta V(r)} d^{3N} q, \tag{1.1}$$

con q representando el espacio conformacional, $\beta = 1/(k_B T)$, donde k_B es la constante de Boltzmann, T la temperatura absoluta y N el número de partículas del sistema. La energía libre (Helmholtz) A está relacionada con Q mediante

$$A = -\frac{1}{\beta} \ln Q. \tag{1.2}$$

La función de partición canónica implica un número constante de partículas, un volumen constante y una temperatura constante. Si la presión, en lugar del volumen, se mantiene constante, se obtiene la energía libre de Gibbs, G. Ahora, en muchos casos se define una coordenada de reacción ξ con respecto a la cuál evolucionaría algún potencial termodinámico como G o A. A menudo, ξ se define sobre bases geométricas, como una distancia, una torsión o la diferencia entre las desviaciones cuadráticas medias de dos estados de referencia (RMSD). Con ξ definida, la distribución de probabilidad del sistema a lo largo de ξ puede calcularse integrando todos los grados de libertad a excepción de ξ,

$$Q(\xi) = \frac{\int \delta[\xi(r) - \xi] e^{-\beta V(r)} d^{3N} r}{\int e^{-\beta V(r)} d^{3N} r}. \tag{1.3}$$

Donde $Q(\xi) d\xi$ puede interpretarse como la probabilidad de encontrar el sistema

en un pequeño intervalo $d\xi$ alrededor de ξ. Esto permite calcular la energía libre a lo largo de la coordenada de reacción como $A(\xi) = -1/\beta$ ln Q(ξ). A la cantidad A(ξ) también se lo denomina potencial de fuerza media o PMF por sus siglas en inglés.

En simulaciones computacionales, las integrales directas en el espacio fase (1.3) y (1.1), son imposibles de calcular. Sin embargo, si el sistema es ergódico, es decir, si se visitan todos los puntos del espacio de fases durante la simulación, se cumple que

$$Q(\xi) = \lim_{t\to\infty} \frac{1}{t} \int_0^t \rho[\xi(t')]dt' = P(\xi). \tag{1.4}$$

Donde t denota el tiempo y ρ cuenta la ocurrencia de ξ en un intervalo dado, de anchura infinitesimal en la ecuación exacta y de anchura finita cuando se calcula un histograma.

Por tanto, el promedio en el ensamble $Q(\xi)$ es igual al promedio temporal $P(\xi)$ para un muestreo infinito en un sistema ergódico. Así, en principio, $A(\xi)$ puede obtenerse directamente a partir de simulaciones de dinámica molecular mediante el seguimiento de $P(\xi)$: la distribución del sistema a lo largo de la coordenada de reacción ξ. Aquí, el término distribución $P(\xi)$ se refiere a la frecuencia normalizada de encontrar el sistema en las proximidades de un valor dado de ξ [7].

Es decir, el PMF puede ser aproximado a partir de un histograma en un intervalo de ξ, producido en una simulación en la que no se impone ninguna restricción sobre el valor de ξ como

$$A(\xi_i) = -\frac{1}{\beta} \ln[P(\xi_i)]. \tag{1.5}$$

1.4.5. Muestreo *Umbrella*

Las simulaciones computacionales sólo se realizan durante un tiempo finito. Las regiones del espacio conformacional en las cercanías de un mínimo del potencial $V(r)$ suelen muestrearse bien, mientras que las regiones de mayor energía, o de barreras de potencial, se muestrean con poca frecuencia. En el caso de sucesos poco probables, tales como los que presentan una barrera de energía significativamente mayor que k_BT, el muestreo directo del espacio conformacional no es posible. Sin embargo, para obtener un perfil $A(\xi)$ es necesario también explorar esas regiones de alta energía, esos eventos poco probables [8].

El método conocido como Muestreo *Umbrella* es conocido por su utilidad para muestrear estos eventos menos probables y consiste en modificar la expresión de energía. En este tipo de muestreo se restringe el rango total de la coordenada de reacción mediante un *potencial de sesgo* de manera que se la "empuja" hacia un valor definido y por tanto el polipéptido explora una región específica del espacio conformacional ω, así dividiendo el rango de la coordenada de reacción en regiones más pequeñas llamadas *ventanas* [9]. Cada ventana cubre sólo una pequeña parte del rango de ξ y es muestreada individualmente. Así, es posible muestrear regiones de barreras energéticas. Finalmente, en un procesamiento posterior, los resultados de las distintas ventanas se combinan para dar lugar a un perfil global de energía libre $A(\xi)$.

Capítulo 2

Metodología

2.1. Caracterización de estructuras

Para poder estudiar la transición del lazo 36 es necesario poder diferenciar una estructura de otra. Por tanto, a continuación se mencionarán algunos parámetros que permiten caracterizar a una estructura. Al estudiar como evoluciona el backbone del lazo 36 se analiza como cambia la disposición de los átomos en diferentes configuraciones permitiendo una caracterización adecuada de una estructura en particular. Por tanto, a continuación se simplificará el estudio de la transición del lazo 36 a estudiar como evoluciona el backbone del polipéptido. En este caso, se tienen $N = 144$ átomos que conforman al backbone.

2.1.1. RMSD

Uno de estos parámetros es la raíz de la desviación cuadrática media (RMSD, por sus siglas en inglés) que se define como

$$\mathrm{RMSD}(t) = \sqrt{\frac{\sum_{i=1}^{N} ||\vec{r}_i(t) - \vec{r}_{i0}||^2}{N}}, \tag{2.1}$$

donde N es el número de átomos que son parte del backbone de la estructura, $\vec{r}_i(t)$ es el vector posición del átomo i de la estructura analizada en el tiempo t, $\vec{r}_{i0}$ el vector posición del átomo i de una estructura de referencia.

2.1.2. Radio de giro

El radio de giro cuantifica qué tan compacta es una estructura y se trata de una diferencia entre posiciones de los átomos en estudio con respecto a su centro de masa, ponderados por las masas de cada átomo. Se define como

$$R_g = \sqrt{\frac{\sum_{i=1}^{N} m_i(\vec{r}_i - \vec{r}_c)^2}{\sum_{i=1}^{N} m_i}} \tag{2.2}$$

donde $\vec{r}_i$ es el vector posición del átomo i de una estructura en particular, $\vec{r}_c$ el vector posición del centro de masa de la estructura y m_i la masa del átomo i.

2.1.3. Coordenadas espaciales

Cada átomo i de los N átomos que forman parte del backbone tienen coordenadas espaciales $(x^{3i-2}, x^{3i-1}, x^{3i})$ $\forall i = 1, 2, ..., N$. Estos $3N$ valores caracterizan a una estructura ya que son diferentes para cada instante de tiempo. En general se tiene que para la estructura k se puede definir el conjunto

$$\Xi_k = \{x^1, x^2, x^3, ..., x^{3N-2}, x^{3N-1}, x^{3N}\}_k \tag{2.3}$$

Cabe recalcar que a diferencia del RMSD o el radio de giro, que son un único valor, cuando se usan las coordenadas espaciales de las partículas se está utilizando un conjunto Ξ_k de valores , con $3N$ elementos, para caracterizar a una estructura.

2.1.4. Distancia entre dos regiones características

De la idea de las coordenadas espaciales se puede definir otro parámetro para caracterizar a una estructura. Es posible tomar solamente dos átomos, o dos regiones de átomos características, calcular la distancia entre ellos y como en general, tomando las regiones adecuadas, esta distancia es diferente para cada instante de tiempo entonces caracterizar así a una estructura.

Para el lazo 36 existen dos regiones relevantes a lo largo de la transición (ver figura 1.2). Así, se define la distancia r_{cm} como el módulo del vector que va desde el centro de masa de los átomos que conforman al residuo 89 y el centro de masa del *brazo* que se pliega al final de la transición, es decir, el conjunto de átomos que

conforman los residuos que van desde el residuo 54 hasta el residuo 75. Esta distancia se define como

$$r_{cm} = ||\vec{R}_{\mathrm{CM}_{89}} - \vec{R}_{\mathrm{CM}_{54\rightarrow 75}}||. \tag{2.4}$$

2.2. Dinámica Molecular

El proceso de plegamiento de proteínas ocurre en escalas temporales del orden de los femtosegundos. Además, las escalas de longitud están en el orden de los angstroms a nivel atómico hasta el tamaño de una proteína plegada en el orden de los nanómetros. Existen distintos métodos para describir la dinámica de sistemas biológicas en función de la longitud y la escala temporal del proceso de interés. Los métodos basados en la mecánica cuántica proporcionan los resultados más precisos. Sin embargo, son costosos computacionalmente, es decir, obtener resultados relevantes en un tiempo de simulación relativamente corto es un reto. Por tanto, mediante este método las simulaciones se limitan a escalas de tiempo cortas (picosegundos) y escalas de longitud pequeñas (nanómetros), en las que intervienen relativamente pocos átomos. Para poder simular la transición del lazo 36 en una escala de tiempo adecuada y poder caracterizar el plegamiento, es necesario acceder a escalas temporales y longitudinales grandes. Con este objetivo, se utilizan métodos basados en la mecánica clásica. Entre estos métodos, está la Dinámica Molecular en donde se simula el movimiento atómico resolviendo las ecuaciones de movimiento de Newton simultáneamente para todos los átomos del sistema [10].

2.3. Simulaciones de Dinámica Molecular

Para poder caracterizar la energía libre del sistema se obtuvo información sobre las coordenadas espaciales de las estructuras que el sistema va visitando mediante simulaciones de dinámica molecular.

NAMD (Nanoscale Molecular Dynamics) es un programa de simulaciones de dinámica molecular escrito con el modelo de programación paralela Charm++. Destaca por su eficiencia paralela y suele utilizarse para simular sistemas con un gran número de átomos [11]. VMD (Visual Molecular Dynamics) es un programa de modelización y visualización de sistemas moleculares. Fue desarrollado principalmente

como herramienta para visualizar y analizar resultados de simulaciones de dinámica molecular [4]. Estos dos programas se utilizaron en conjunto para poder estudiar la transición del polipéptido.

Para ejecutar cualquier simulación de dinámica molecular mediante el software NAMD, es necesario un conjunto de archivos:

- Un archivo *pdb* (Protein Data Bank) que almacena las coordenadas atómicas y/o las velocidades del sistema. Los archivos pdb de las estructuras del lazo 36 en su estado parcialmente aleatorio y totalmente plegado se obtendrán de la base de datos Protein Data Bank [12].
- Un archivo de Estructura Proteínica (*PSF*, por sus siglas en inglés) que almacena información estructural del polipéptido. Este archivo contiene toda la información específica del sistema necesaria para aplicar un campo de fuerzas al sistema molecular. Generar un archivo PSF requiere utilizar un archivo de topología que contiene toda la información necesaria para convertir una lista de nombres de residuos en un archivo de estructura proteínica.
- Un archivo de parámetros de campos de fuerza. Un campo de fuerza es una expresión matemática del potencial que experimentan los átomos en el sistema. En este archivo se encuentran todas las constantes numéricas necesarias para poder calcular magnitudes de fuerzas y valores de energías. El archivo de parámetros está estrechamente ligado al archivo de topología que se utilizó para generar el archivo PSF. Normalmente se distribuyen juntos y se les da nombres que coinciden.
- Un archivo de configuración, en el que se especifica todas las opciones que NAMD debe adoptar al ejecutar una simulación. El archivo de configuración le dice a NAMD con qué parámetros debe ejecutarse la simulación.

2.3.1. Dinámica molecular orientada (TMD)

Las simulaciones de dinámica molecular orientada (*TMD*, por sus siglas en inglés) son un método para observar la transición conformacional a gran escala entre dos estructuras conocidas de una sistema molecular. En las simulaciones TMD, un subconjunto de átomos es guiado hacia una estructura *objetivo* mediante fuerzas orientadoras. En cada paso temporal, t, después de alinear la estructura objetivo

con las coordenadas de la estructura en el tiempo t, se calcula el RMSD entre las coordenadas actuales y la estructura objetivo. Así, la fuerza sobre cada átomo viene dada por el gradiente del potencial virtual U_{TMD} que tiene la expresión

$$U_{TMD} = \frac{1}{2}\frac{\kappa}{N}\left(R(t) - R_0(t)\right)^2, \tag{2.5}$$

donde κ es la constante de fuerza orientadora, N es el número de átomos restringidos, $R(t)$ es la raíz de la distancia cuadrática media de la estructura simulada en el tiempo t relativa a la estructura objetivo, y $\mathrm{R}_0(t)$ es la raíz de la distancia cuadrática media de la estructura objetivo relativa a la estructura inicial que decrece linealmente en el tiempo hasta cero [13]. Para el *backbone* del lazo 36 se tiene que $N = 144$ y se utilizó $\kappa = 300\frac{kcal}{mol\text{Å}^2}$.

2.3.2. Control de presión y temperatura

En NAMD se pueden configurar las simulaciones para que el sistema evolucione con un control de presión y temperatura constante mediante la Dinámica de Langevin que introduce nuevas fuerzas como efectos amortiguador para mantener este control.

La dinámica de Langevin permite controlar la energía cinética del sistema y, por tanto, la temperatura y/o la presión del mismo [11]. El método utiliza la ecuación de Langevin para una sola partícula en donde se agregan dos términos adicionales a la fuerza ordinaria $\vec{F}_i(t)$ que experimenta la partícula i:

$$m_i\frac{d^2\vec{r}_i}{dt^2} = \vec{F}_i(t) - \gamma\dot{\vec{r}}_i(t) + \vec{R}_i(t). \tag{2.6}$$

El segundo término representa una amortiguación por fricción que se aplica a la partícula con un coeficiente de fricción γ proporcional a la velocidad y el tercer término representa fuerzas aleatorias que actúan sobre la partícula como resultado de la interacción con el disolvente. Esta fuerza aleatoria está asociada a una temperatura y añade energía al sistema, mientras que el término de fricción disipa energía. La fuerza aleatoria suele tener una distribución gaussiana con un valor medio de cero [14].

2.3.3. Esquema de las simulaciones de dinámica molecular

Para realizar una simulación de dinámica molecular utilizando NAMD y VMD es necesario seguir un protocolo específico. En particular, es necesario realizar simulaciones previas en las que se prepara al sistema para poder realizar un muestreo adecuado del espacio conformacional del polipéptido. Los diferentes códigos fueron escritos en tcl y adaptados para ser ejecutados por VMD y NAMD (Anexo A).

Alineación

Para poder obtener valores de parámetros consistentes a lo largo de una simulación es necesario que el sistema esté alineado a lo largo de un mismo eje. En este caso, se ha escogido el eje z del sistema de coordenadas definido por VMD. El sistema fue primero desplazado hacia el centro de masas y luego alineado hacia el eje z mediante una matriz de rotación. Aquí, se definió un vector $\vec{\mathcal{R}}$ que va desde el átomo CA del residuo 75 hacia el átomo CA del residuo 89 (ver figura 1.2) ya que esta sección del polipéptido no se mueve significativamente a lo largo de la transición. En este caso, se generó un nuevo archivo pdb pero esta vez con nuevas coordenadas de los átomos de manera que $\vec{\mathcal{R}}$ esté alineado con el eje z.

Generación del archivo PSF

Una vez el sistema se encuentra alineado, se utilizó como entrada el archivo pdb de la estructura alineada para generar el archivo psf. En principio, como el archivo psf está asociado con la información estructural de la proteína sería posible obtener el archivo psf directamente con el archivo pdb antes de que se le haya aplicado la rotación. En cualquier caso, para mantener la organización que se sigue en [11] se obtuvo el archivo psf después de la rotación. Se generó el archivo psf mediante el paquete provisto por NAMD llamado *psfgen* y el archivo de topología *top_all27_prot_na.rtf* detallado por el campo de fuerzas CHARMM22.

Solvatación

Ahora, para poder simular el ambiente celular es necesario solvatar al sistema, es decir, colocarlo en agua. Para el presente estudio, se realizará una solvatación utilizando un cubo de agua con condiciones de borde periódicas. El uso de este

tipo de condiciones de borde implica rodear al sistema con celdas unitarias virtuales idénticas, así, los átomos de los sistemas virtuales circundantes interactúan con los átomos del sistema real. Además, el cubo debe ser lo suficientemente grande como para que la proteína no interactúe con su respectiva imagen en las celdas aledañas y para que el agua siga sumergiendo significativamente a la proteína cuando se encuentre totalmente extendida.

La solvatación se realizó mediante el paquete *solvate* provisto por VMD. Las dimensiones del cubo de agua fueron tales que haya una capa de agua de 15Å en cada dirección desde el átomo con la coordenada espacial más grande en cada dirección, x, y y z.

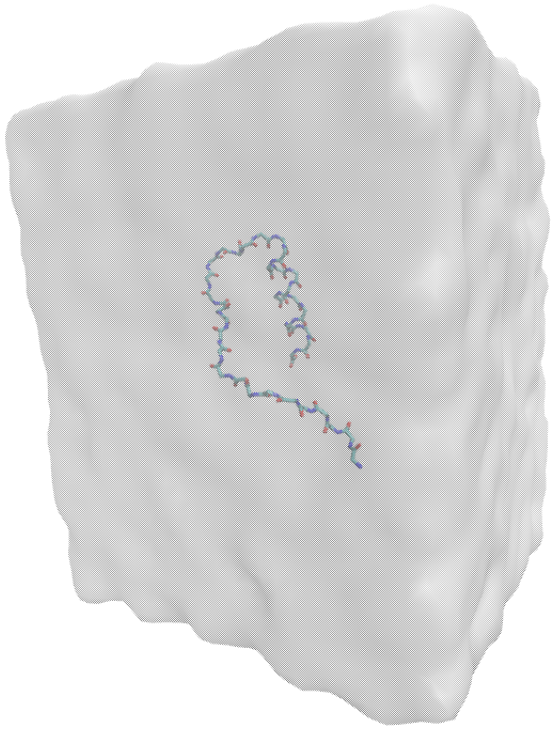


Figura 2.1: Solvatación del Lazo 36 utilizando una caja de agua con condiciones periódicas.

Minimización de la energía

Ahora que la proteína está disuelta en un cubo de agua, es necesario realizar una minimización de la energía del sistema con el fin de eliminar cualquier mal contacto

entre los átomos que se han añadido. Como por ejemplo, átomos de hidrógeno que se han añadido en una posición que está demasiado cerca de otro átomo de la proteína. Esto implica calcular la fuerza sobre todos los átomos y utilizarla para empujarlos hacia configuraciones estables de energía mínima, es decir, un lugar en el que la molécula esté relajada, variando sistemáticamente las posiciones de los átomos y calculando la energía.

Aquí, es necesario obtener constantes del campo de fuerzas asociadas a cada residuo. Para este trabajo se utilizó el archivo *par_all27_prot_na.prm* detallado por el campo de fuerza CHARM22 que se utiliza para sistemas de proteinas (*prot*) y ácidos nucléicos (*na*).

La simulación de minimización se realizó a una temperatura $T = 0$, durante 0.01 [ns], con una frecuencia de recolección de datos igual a 1000 pasos y fijando dos átomos de manera que no se muevan a lo largo de la simulación: el átomo CA del residuo 89 y el átomo CA del residuo 75 (ver figura 1.2). Se mantienen estos átomos fijos para que el sistema explore el espacio conformacional en una vecindad de la estructura con la que se inicia la simulación.

Calentamiento del sistema

La estructura de energía mínima es una aproximación de la estructura que adoptarían las moléculas en condiciones cercanas del cero absoluto. Si se intentase ejecutar inmediatamente la dinámica molecular a temperatura ambiente o corporal en este sistema, las moléculas explotarían. Para evitarlo, se calienta suavemente el sistema asignando un valor de temperatura con cierta frecuencia. En este caso, se utilizó una incremento de asignación igual a $30K$ con una frecuencia de 1000 pasos hasta alcanzar una temperatura constante $T = 300K$, en cada paso se ejecuta el control de temperatura de Langevin descrito anteriormente. Además, se aplicó un potencial armónico para restringir el movimiento en los ejes x y y a los dos átomos mencionados en la anterior simulación.

Equilibración del sistema

El siguiente paso consiste en equilibrar el sistema mediante una simulación de dinámica molecular isotérmica-isobárica (NPT). Aquí, la idea es que la ecuación de movimiento dada por la segunda ley de Newton se resuelve para cada átomo del sistema para dictar su trayectoria. Esto ajustará el tamaño del cubo, asegurando

que la densidad del agua en el cubo periódico sea la correcta para la temperatura y presión definidas.

Esta simulación fue ejecutada para una temperatura constante $T = 300$[K], una presión constante $P = 1$[atm] , con un tiempo de simulación de 0.1[ns] y con una restricción en los ejes x y y para el átomo CA del residuo 89.

Equilibración de dinámica libre

Una vez se tenga la densidad del cubo de agua adecuada, el paso final es una simulación real de dinámica molecular. Los parámetros y restricciones para esta simulación fueron los mismos que para la de equilibración antes mencionados pero esta vez se ejecutó por 1 [ns].

2.4. Análisis de grupos

El análisis de clusters o grupos comprende un conjunto de métodos para clasificar datos multivariantes en subgrupos. Al organizar los datos multivariantes en dichos subgrupos, la agrupación puede ayudar a revelar características existentes en los datos [15]. En el presente estudio este tipo de análisis es de utilidad para poder identificar estructuras visitadas por el lazo 36 a lo largo de su transición que sean similares entre sí.

Entre los métodos más conocidos está la *clasificación jerárquica aglomerativa*. Aquí, un conjunto de n datos se divide en un número determinado de clases o grupos siguiendo una serie de particiones. Las clasificaciones jerárquicas pueden representarse mediante un diagrama bidimensional denominado *dendrograma*, que ilustra las fusiones o divisiones realizadas en cada etapa del análisis. Cada nodo interior del dendrograma corresponde a una fusión de dos grupos o datos. En la figura 2.2 se muestra un ejemplo de este tipo de diagrama.

Los procedimientos aglomerativos producen una serie de particiones de los datos. Como se muestra en la figura 2.2, desde abajo hacia arriba: la primera partición consiste en n grupos de un solo miembro y la última consiste en un único grupo que contiene todos los n individuos. En cada etapa, el método se encarga de fusionar a los individuos o grupos de individuos más cercanos (o más similares).

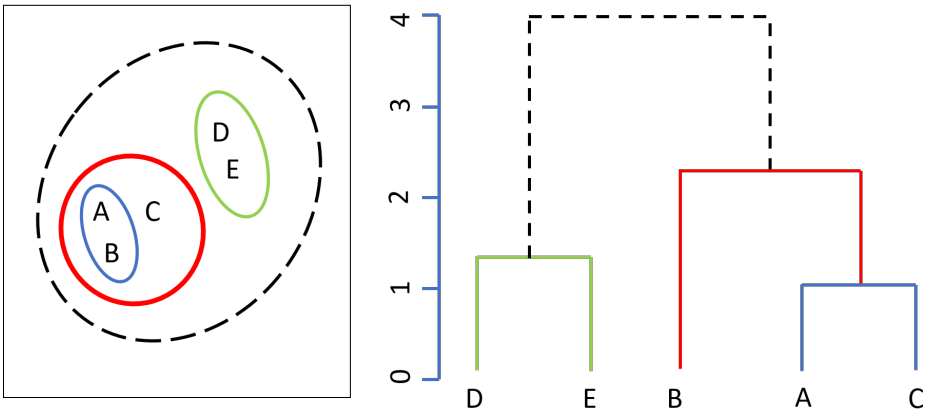


Figura 2.2: Diagrama representativo de la secuencia de grupos producido por la clasificación jerárquica aglomerativa para $n = 5$, 5 datos.

2.5. Cálculo del Potencial de fuerza media (PMF)

A partir del muestreo del espacio conformacional en el que se obtiene información sobre las coordenadas de los átomos que conforman al polipéptido, se pueden calcular diferentes parámetros que caractericen a una estructura en particular, como aquellos definidos en 2.1. Estos parámetros pueden ser utilizados como coordenada de reacción para el cálculo del potencial de fuerza media mediante la relación (1.5). Así, si se realiza un muestreo por todo el espacio conformacional en el que el polipéptido visite diferentes estructuras hasta la estructura totalmente plegada, entonces se obtendría el PMF de la transición.

2.5.1. Muestreo del espacio conformacional

La forma en la que se realiza este muestreo debe tomar en cuenta algunos aspectos. Por un lado, se puede aprovechar alguna característica del sistema para realizar un muestreo amplio pero al mismo tiempo con pocos recursos computacionales y en un tiempo relativamente corto. Por otro lado, se debe considerar que a lo largo de la transición es posible que existan barreras de energía que, en el proceso natural, el polipéptido debe vencer para llegar al estado final de la transición. Es decir, el uso de simulaciones de dinámica molecular estaría limitado ya que el muestreo se detendría cuando el sistema se encuentre con una barrera de energía. Así, obteniendo un muestreo incompleto el espacio conformacional y por tanto una curva PMF incompleta también.

2.5.2. Combinación de simulaciones TMD, muestreos *Umbrella* y análisis de grupos para la obtención del PMF

Una estrategia para poder obtener una primera noción del espacio conformacional del polipéptido es realizar una simulación de dinámica molecular orientada, TMD. Así, se obtendría un primer muestreo a partir del cuál es posible construir un primer PMF, mediante el seguimiento de la distribución de una coordenada de reacción en particular, como se mencionó en 1.4.4. Aquí, el polipéptido vence cualquier barrera de energía mediante las fuerzas que orientan al sistema hacia la estructura final. A pesar de que mediante esta simulación es posible obtener una primera curva PMF, esta no es del todo *confiable* ya que, como se mencionó en 2.3.1, el potencial mediante el cual el sistema evoluciona desde la estructura inicial hacia la final se trata de uno *virtual* entonces es posible que el polipéptido visite estructuras no accesibles de manera natural. Es decir, el PMF asociado a esta simulación no representa de manera adecuada el proceso natural de la transición.

Para que el PMF sea más confiable y apegado a la transición, es necesario dejar que, a partir de una estructura inicial, el sistema evolucione sin ser guiado a una estructura específica, es decir, dejar que el sistema evolucione sin restricciones. Sin embargo, nuevamente, el problema sería que el muestreo de ω se detendría cuando el sistema se encuentre con una barrera de potencial.

La estrategia que se utilizó para realizar un muestreo completo de ω utiliza el muestreo *Umbrella*. Aquí, se divide el dominio de la coordenada de reacción en intervalos o ventanas. Luego, para cada ventana se realizan simulaciones de dinámica molecular, según el esquema mencionado en 2.3.3, con estructuras iniciales asociadas al valor de coordenada de reacción de la respectiva ventana. Al realizar esta división por ventanas es posible muestrear regiones en donde existan barreras de energía ya que la región de ω que explora el sistema es menor y por tanto es posible muestrear estos eventos menos probables.

Ahora, siguiendo un procedimiento similar al que se menciona en [9], se escogieron las estructuras iniciales para los muestreos *Umbrella* a partir de la simulación TMD. El cambio más importante y novedoso que se aplicó en este estudio fue la manera en la que se seleccionan estas estructuras. Mediante un análisis de grupos se evaluó la similitud entre las coordenadas espaciales de las estructuras que va visitando el sistema a lo largo de la simulación TMD. Aquí, se agrupan a las estructuras de manera que a cada grupo le corresponde una *estructura representativa*. Es decir, se

obtendrá un conjunto de estructuras representativas a la transición. Este conjunto de estructuras representativas, y por tanto conjunto de coordenadas de reacción, son las que se utilizarán para realizar los muestreos *Umbrella*.

2.5.3. Muestreo inicial de ω

A partir de la simulación TMD se realizó un primer muestreo de ω. De este muestreo se obtuvieron datos de la distancia entre el centro de masa del residuo 89 y el centro de masa del brazo que abarca los residuos del 54 al 75, r_{cm}, el RMSD con respecto a la estructura inicial del lazo 36, es decir aquella que representa su estado parcialmente aleatorio y el radio de giro, rg.

La evolución temporal de estos parámetros se muestra en la figura 2.3. Aquí, se puede observar que existen regiones temporales en las que se puede considerar que el valor de estos parámetros permanecen constantes. Este comportamiento está relacionado con la forma en la que el sistema evoluciona hacia el estado final. En esta evolución, el polipéptido visita estructuras similares y por tanto estas presentan características similares entre si.

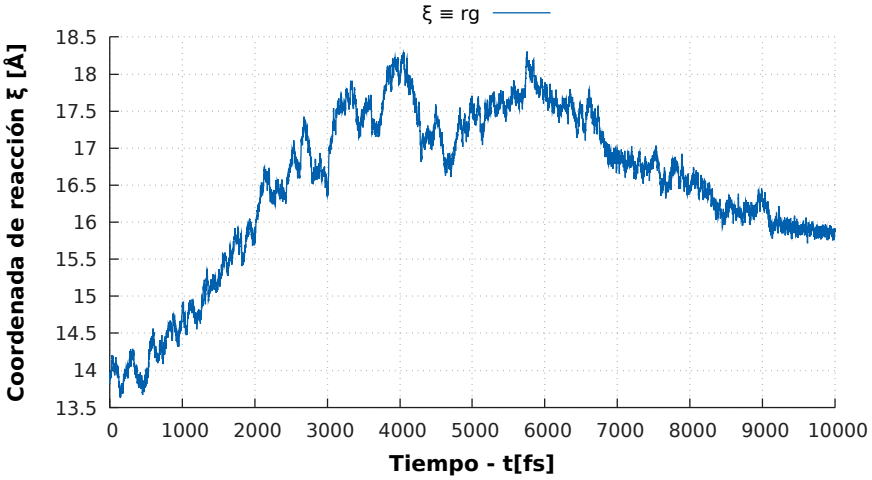
Además, con estos datos se construyó la curva del PMF, esta serviría para obtener una primera idea de qué ocurre con la evolución de la energía del sistema. El gráfico de este PMF utilizando como coordenada de reacción a r_{cm} se muestra en la figura 2.4.

2.5.4. Criterio para la distancia de corte

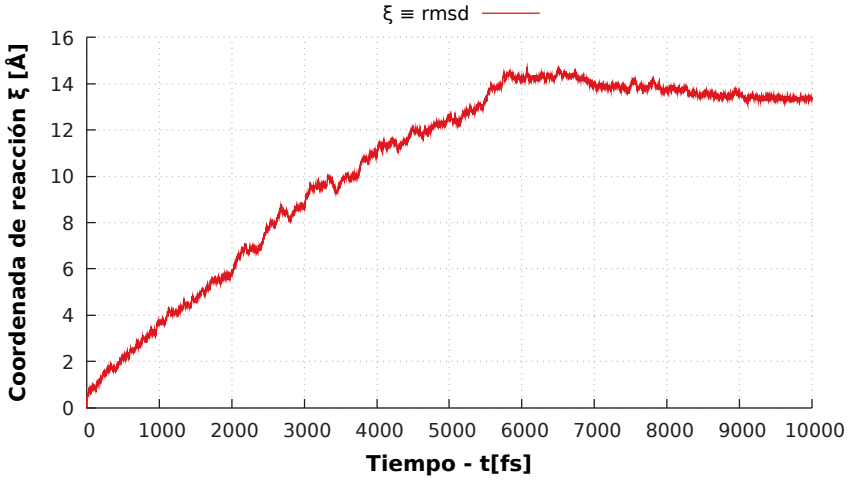
Cuando se realiza un análisis de grupos mediante el paquete *protoclost* provisto por el software R [16], es necesario definir una distancia de corte h que actúa como una métrica para definir qué tan diferente es un grupo de otro. Siendo así, es necesario definir cierto criterio asociado al objetivo de la clasificación en grupos para poder establecer un valor concreto de h.

El criterio que se usó toma en cuenta que al momento de realizar un gráfico del PMF a lo largo de cierta coordenada de reacción mediante los datos obtenidas de la simulación TMD, se observarán ciertas regiones de interés en las que sería conveniente realizar un muestreo extensivo. Entonces, definiendo una *resolución de muestreo*, $\Delta\xi$, es posible calcular el valor de h.

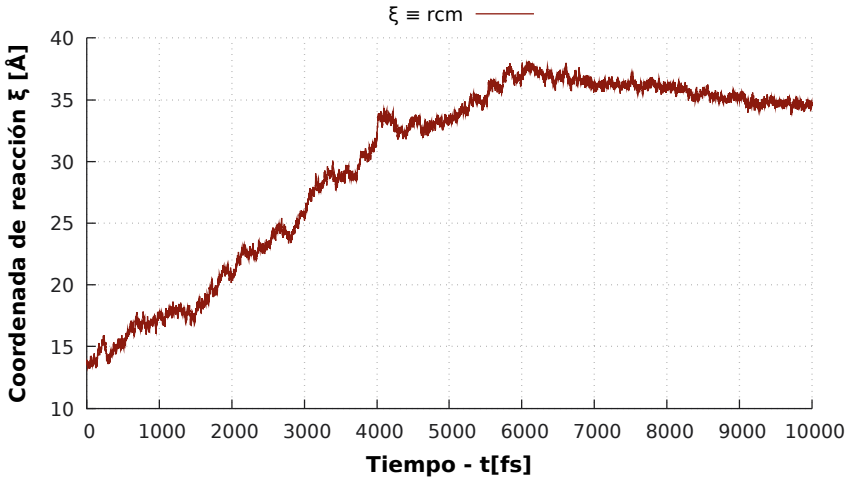
A cada estructura representativa le corresponde una coordenada de reacción ξ_i



(a) Radio de giro, rg, en función del tiempo.



(b) RMSD con respecto a su estructura inicial parcialmente aleatoria en función del tiempo.



(c) Distancia entre los centros de masa del brazo y el residuo 89, r_{cm}, en función del tiempo.

Figura 2.3: Evolución temporal de tres parámetros obtenidos de la simulación TMD.

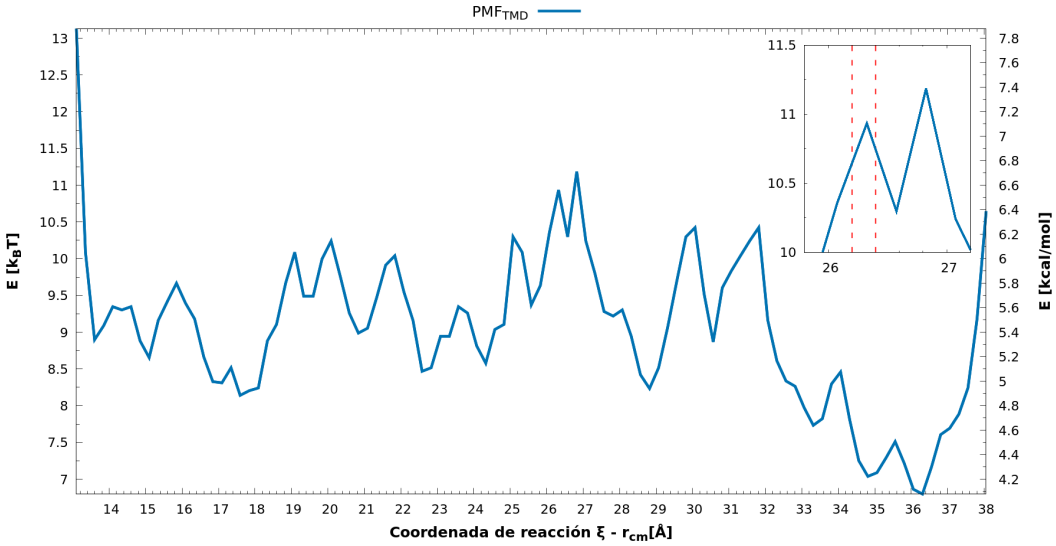


Figura 2.4: Curva PMF obtenida a partir de datos de la simulación TMD a lo largo de la coordenada de reacción r_{rcm} y como inset la región de interés definida como el pico más angosto con una anchura aproximadamente igual a 0.21Å.

que determina un límite superior o inferior de una *ventana* en la que se dejara que el sistema evolucione de manera natural. Esto quiere decir que buscamos un h tal que la mínima diferencia entre un par de coordenadas de reacción que definen una ventana de simulación sea igual a la resolución $\Delta\xi$. De esta manera se asegura que el sistema evolucione en un rango en el que se pueda obtener información sobre una región de interés como serían los picos que se observan en los primeros PMF.

Es decir, se inicia proponiendo un valor inicial de h arbitrario. Con este valor de h se obtiene cierto número de grupos N_g, a cada grupo le corresponde una estructura representativa y por tanto es posible calcular un valor de coordenada de reacción de cada una de estas estructuras ξ_i. Luego, para cada par de estructuras representativas consecutivas se halla la diferencia entre sus coordenadas de reacción $\Delta\xi_i$ donde $i = 1, 2, ..., N_g - 1$. Finalmente, de todas las diferencias $\Delta\xi_i$, se obtiene el valor mínimo $\Delta\xi_m$. Si el valor del mínimo de estas diferencias es aproximadamente igual a la resolución definida entonces se trabaja con ese valor de h y ese número de grupos N_g. Caso contrario, se prueba con otro valor de h hasta que se cumpla que $\Delta\xi_m = \text{mín}\{\Delta\xi_1, \Delta\xi_2, \Delta\xi_3, ..., \Delta\xi_{N_g-1}\} \approx \Delta\xi_{\text{ref}}$.

La clasificación se hará en función de las coordenadas espaciales $\{x, y, z\}$ de los 144 átomos que conforman al *backbone* del lazo 36 de cada una de las estructuras que visita el polipéptido a lo largo de la simulación TMD, en total 2 millones de

estructuras. Así, se evaluará la distancia euclidiana entre estructuras, δ, para definir qué tan *diferente* es una estructura de otra. Las estructuras que cumplan que $\delta < h$ se considerarán similares entre sí y por tanto pertenecerán a un mismo grupo. El resultado será una lista de grupos, cada uno con una estructura representativa y por tanto con un valor de coordenada de reacción representativo ξ_i.

Definición de la región de interés

Se definió como *región de interés* a uno de los picos que se observan en la curva PMF obtenida a partir de la simulación TMD (ver figura 2.4). Siendo así, se tomó como *resolución de muestreo* un valor de $\Delta\xi_{\text{ref}} = 0.2[\text{Å}]$ el cual está asociado al ancho de los picos menos angostos que se pueden visualizar en la figura 2.4.

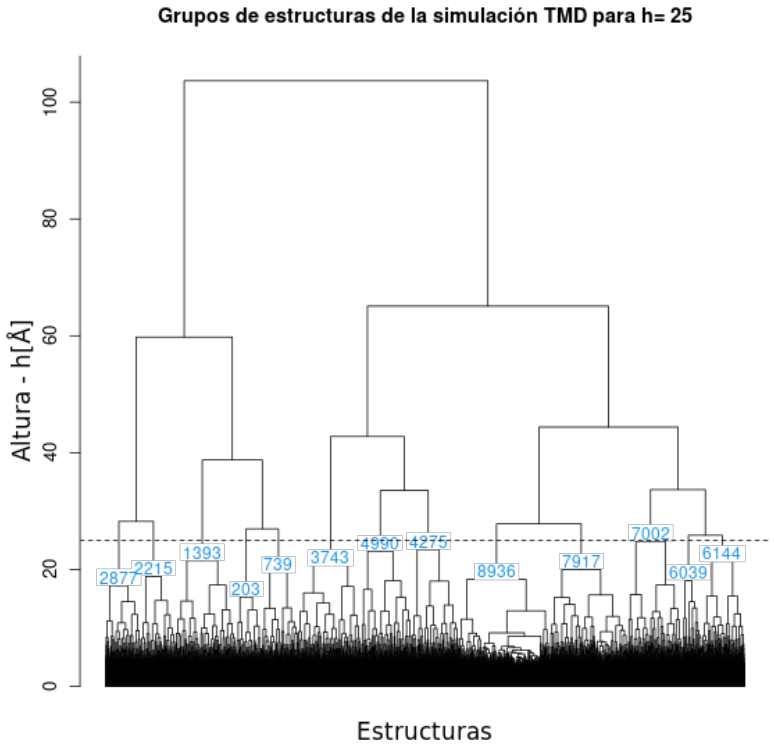


Figura 2.5: Dendrograma para las estructuras obtenidas de la simulación TMD y una distancia de corte $h = 25$.

Así, para una distancia de corte de $h = 25[\text{Å}]$ se obtuvieron 13 grupos y una diferencia entre coordenadas de reacción mínima igual a $\Delta\xi_m = 0.21[\text{Å}]$. El dendrograma de esta clasificación se muestra en la figura 2.5 en donde se muestra el índice

asociado a la estructura representativa de cada grupo, este índice coincide con el valor temporal de la estructura que visita el polipéptido a lo largo de la simulación.

En la tabla 2.1 se muestra la estructura representativa con su respectivo valor temporal t_{TMD} y el valor de la coordenada de reacción asociada a esa estructura en particular.

Estructura representativa (e)	**Valor temporal - t_{TMD} [ps]**	**Coordenada de reacción ξ - rcm [Å]**	**Diferencia entre valores de ξ - $\lvert\Delta\xi\rvert$ [Å]**
e_1	203	15.41	1.86
e_2	739	17.26	2.34
e_3	1393	19.6	5.12
e_4	2215	24.72	2.44
e_5	2877	27.16	5.2
e_6	3743	32.36	3.59
e_7	4275	35.95	**0.21**
e_8	4990	36.15	4.46
e_9	6039	40.61	0.4
e_{10}	6144	41.01	1.35
e_{11}	7002	39.66	0.53
e_{12}	7917	39.13	0.63
e_{13}	8936	38.5	-

Cuadro 2.1: Tabla con la lista los valores de la coordenada de reacción r_{cm} de las estructuras representativas resultantes del análisis de grupos obtenidas a partir de los datos de la simulación TMD.

Capítulo 3

Resultados, conclusiones y recomendaciones

3.1. Resultados y Discusión

3.1.1. Exploración de ω

Cuando se realiza una simulación de dinámica molecular a partir de una de las estructuras de la lista 2.1, el lazo 36 explorará cierto rango de la coordenada de reacción r_{cm}. Ahora, se quiere evitar que el rango que explore el polipéptido a partir de alguna de las estructuras mencionadas se solape con el rango obtenido de otra estructura ya que la información sobre el muestreo de ω sería redundante, es decir, no se estaría explorando todo el espacio conformacional lo que implica un muestreo incompleto.

Como en un principio no se conoce el rango de r_{cm} que explorará el polipéptido a partir de una estructura en particular, entonces para evitar el problema mencionado se comenzó por realizar simulaciones de dinámica molecular para las estructuras inicial, $t_{TMD} = 0$, y final, $t_{TMD} = 9999$, del lazo 36. El rango de r_{cm} que explora el sistema a partir de estas dos estructuras fue $\Delta r_{cm}^{0} = [11.38\text{Å}, 15.66\text{Å}]$ y $\Delta r_{cm}^{9999} = [36\text{Å}, 39.15\text{Å}]$. Los histogramas para las dos simulaciones a partir de estas estructuras se muestran en las figura 3.1, aquí se puede observar que el rango de r_{cm} que explora el polipéptido está confinado a cercanías del valor de coordenada de reacción de la estructura representativa. Esto fue obtenido sin la necesidad de restringir la coordenada de reacción representativa de una ventana en particular, es

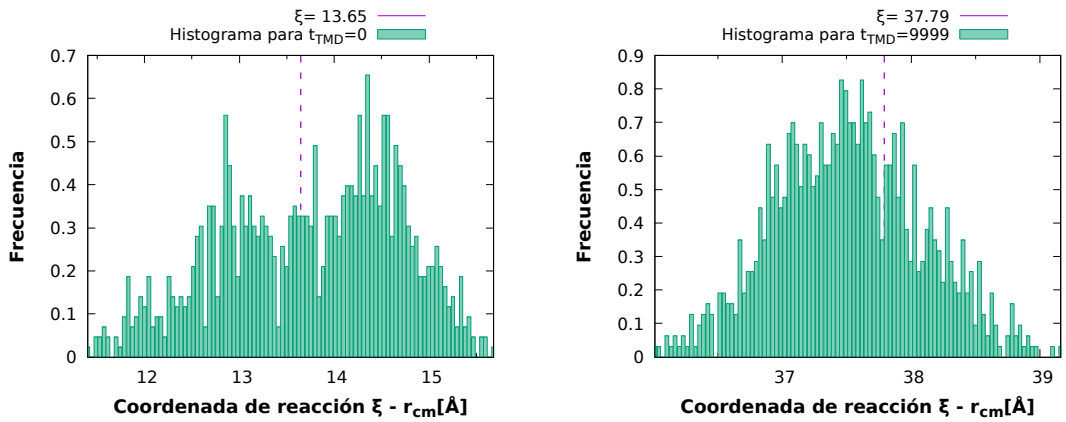


Figura 3.1: Histogramas de r_{cm} obtenidos a partir de las estructuras inicial y final de la transición.

decir, sin utilizar un *potencial de sesgo* como se suele hacer en muestreos *Umbrella* tradicionales. Este resultado muestra que existen barreras de energía naturales a lo largo de la transición del polipéptido que permiten que el rango de r_{cm} esté confinado a esos valores y por tanto quita la necesidad del potencial de sesgo.

Para continuar con el muestreo de ω, y por tanto de todo el dominio de la coordenada de reacción r_{cm}, a continuación se escogen estructuras con valores de coordenada de reacción que estén fuera de los rangos de los dos histogramas ya obtenidos, figura 3.1, para así minimizar el solapamiento entre los rangos de exploración de cada estructura. Es decir, se modificará la lista previa de estructuras 2.1.

Estructura representativa (e)	**Valor temporal - t_{TMD} [ps]**	**Coordenada de reacción ξ - rcm [Å]**
e_1	0	13.65
e_2	216	15.78
e_3	739	17.26
e_4	1586	20.37
e_5	2215	24.72
e_6	3078	30.27
e_7	3602	32.04
e_8	3938	34.43
e_9	4990	36.15
e_{10}	9999	37.79

Cuadro 3.1: Tabla con la lista los valores de la coordenada de reacción r_{cm} de las estructuras utilizadas para la exploración de ω.

Por tanto, el análisis de grupos nos permite obtener una *lista preliminar* de las estructuras a partir de las cuales se realizarán los muestreos por ventanas, ya que en principio no se conoce qué rango de ω va a explorar cada estructura representativa. Después de adaptar este esquema de selección de estructuras se obtuvo una lista de 10 estructuras, entre ellas se encuentran 3 estructuras de la lista inicial, 2.1. La nueva lista se muestra en la tabla 3.1. Este resultado es comparable con lo reportado en [17], en donde se escogieron 10 potenciales armónicos para 10 ventanas o regiones equidistantes representando las divisiones del dominio de la coordenada de reacción. Es decir, el número de estructuras representativas, y por tanto el número de ventanas, halladas mediante la estrategia empleada en este escrito está en buena aproximación a los valores típicos que se usan en otras referencias. El aspecto en el que difieren estas ventanas es que las definidas aquí no son equidistantes, aquí se tienen ventanas más amplias que otras.

Así, a partir de las coordenadas espaciales obtenidas en las simulaciones de dinámica molecular, se graficaron (ver Figura 3.2) los histogramas del seguimiento de r_{cm} para las estructuras de la tabla 3.1.

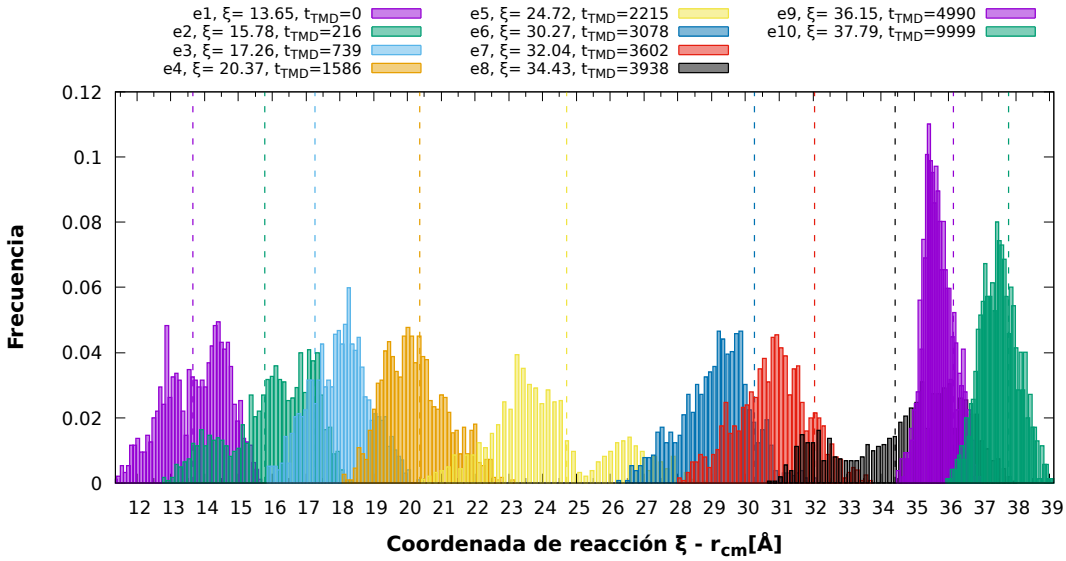


Figura 3.2: Histogramas de la evolución del lazo 36 para diferentes regiones del dominio de r_{cm} muestreadas mediante 10 ventanas sin un potencial de sesgo.

Aquí, se puede observar que el muestreo se realiza en las cercanías del valor de coordenada de reacción asociado a una ventana en particular para todas las estructuras. Además, la interpretación de las regiones en dónde se observa la frecuencia más alta se asocia con un valor de energía mínimo, pues el sistema explora estas regiones sin un esfuerzo alguno y por tanto se mantiene en las cercanías de ese

valor de coordenada de reacción un mayor tiempo. Por otro lado, las regiones con valores menores de frecuencia se asocian con valores de energía muy altos, pues el sistema requiere de un mayor esfuerzo para estar en las cercanías de esos valores de coordenada de reacción y por tanto explora esa región con una frecuencia menor.

3.1.2. PMF a lo largo de r_{cm}

A continuación, a partir de los histogramas 3.2 se construyó el PMF a lo largo de r_{cm}, este gráfico se muestra en la figura 3.3 en donde se ha graficado también el PMF obtenido mediante la simulación TMD con notación PMF_{TMD}, y para la curva obtenida mediante los muestreos *Umbrella* y dinámica molecular, la notación PMF_{FD}, en dónde se ha hecho énfasis en que esta curva del PMF ha sido obtenida prácticamente, como se menciona en el protocolo 2.3.3, mediante simulaciones de dinámica libre (*Free Dynamics*).

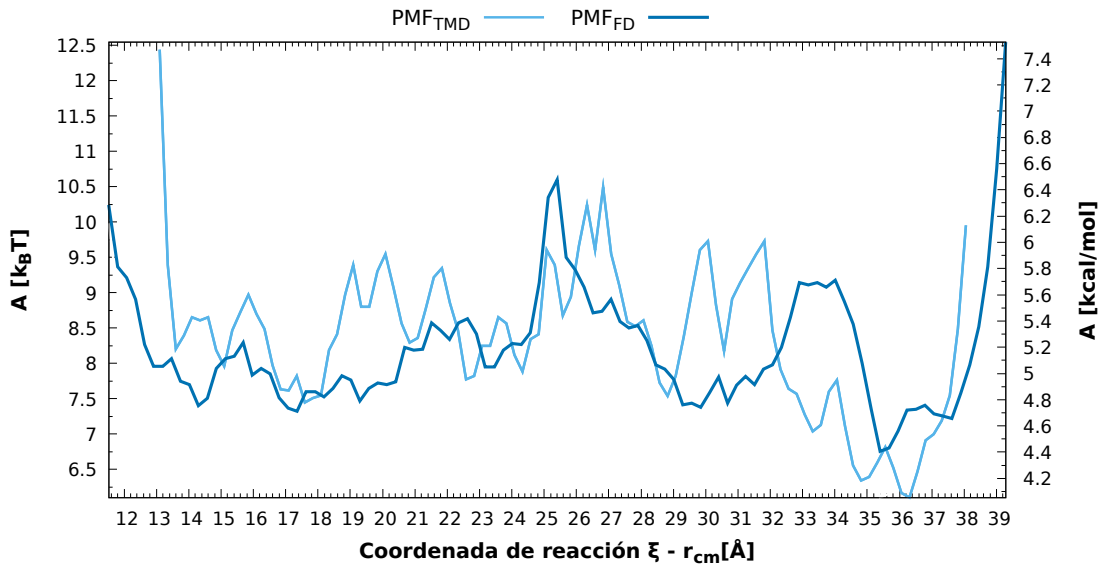
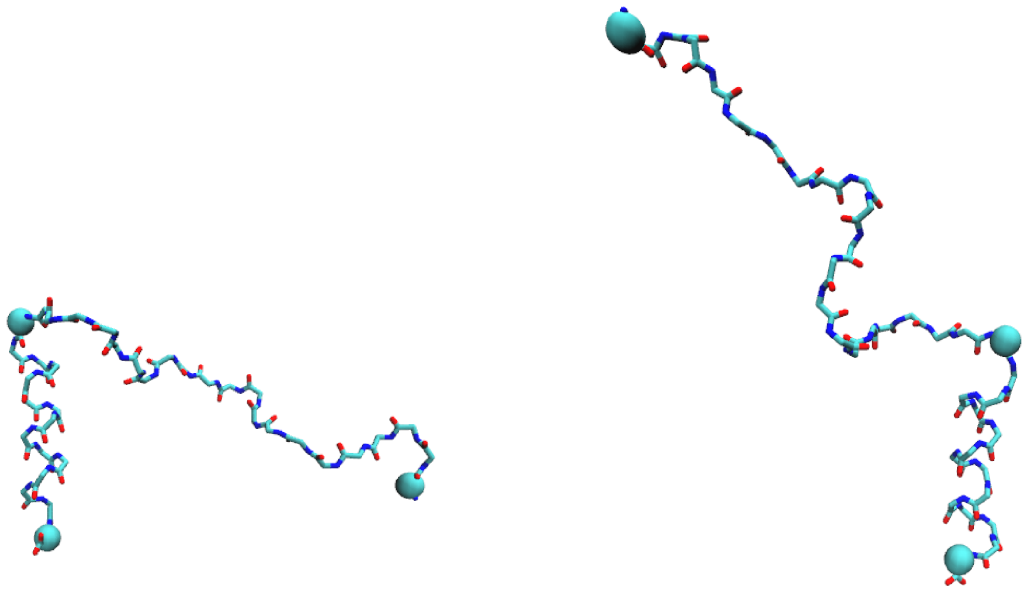


Figura 3.3: PMF a lo largo de r_{cm} obtenido a partir de la simulación TMD y obtenido a partir de las simulaciones de dinámica libre (FD, *Free Dynamics*).

En esta figura se observa que para las dos curvas se observa que entre $r_{cm} = [24, 27]$Å existe una región de alta energía y que en general el comportamiento de las dos curvas coincide. Este resultado muestra que la elección como coordenada de reacción al parámetro r_{cm} permite una caracterización confiable de la transición del lazo 36. Por otro lado, existen regiones en las que se observa un aparente pico en la curva PMF_{TMD} pero no en PMF_{FD}, como por ejemplo cerca de $r_{cm} = \{19, 20, 30, 32\}$Å.

Esto quiere decir que en la curva PMF_{TMD} se presentaron picos *falsos* y por tanto la curva PMF_{TMD} en un principio no es del todo confiable, es necesario realizar un muestreo con más detalle para obtener información que describa el proceso natral de la transición del polipéptido. Así, entonces se puede mencionar que existe una barrera de potencial cuando $r_{cm} = 25.4$[Å] con un ancho de aproximadamente 1.6 [Å] y una altura de 2.49[kcal/mol] con respecto al valor mínimo de energía de 3.98[kcal/mol] que se encuentra cuando $r_{cm} = 35.4$[Å].

Las estructuras asociadas a la región de la barrera de potencial y la región de energía mínima se muestran en la figura 3.4. Aquí se puede observar que cerca de la barrera de potencial el lazo 36 no se ha plegado todavía y el *brazo* está extendido casi por completo. Además, cerca del mínimo de energía, el brazo esta extendido y el plegamiento ya ha comenzado.

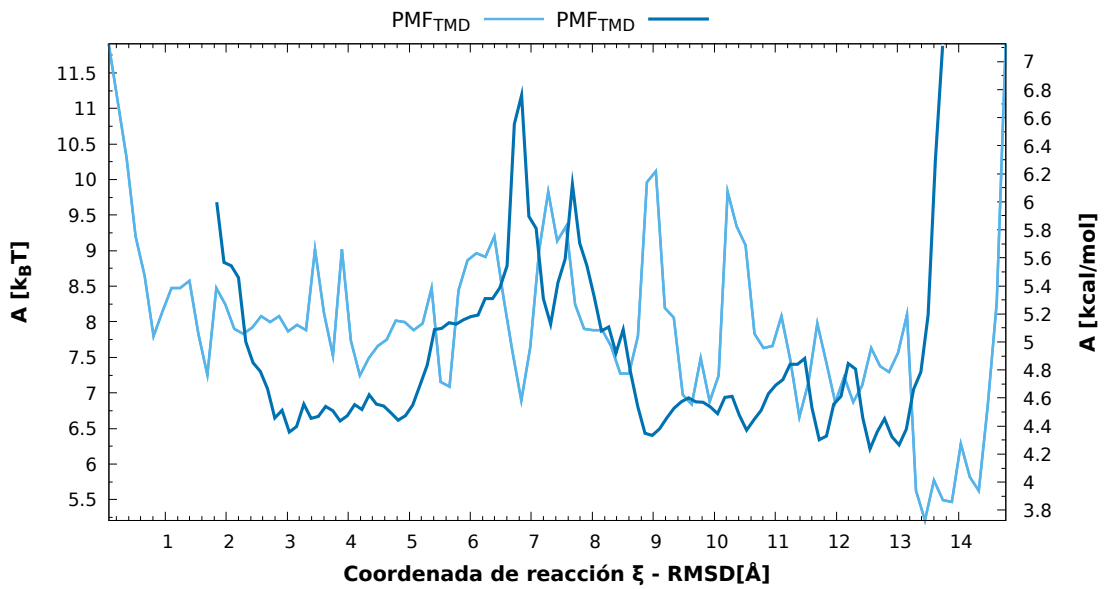


(a) (b)

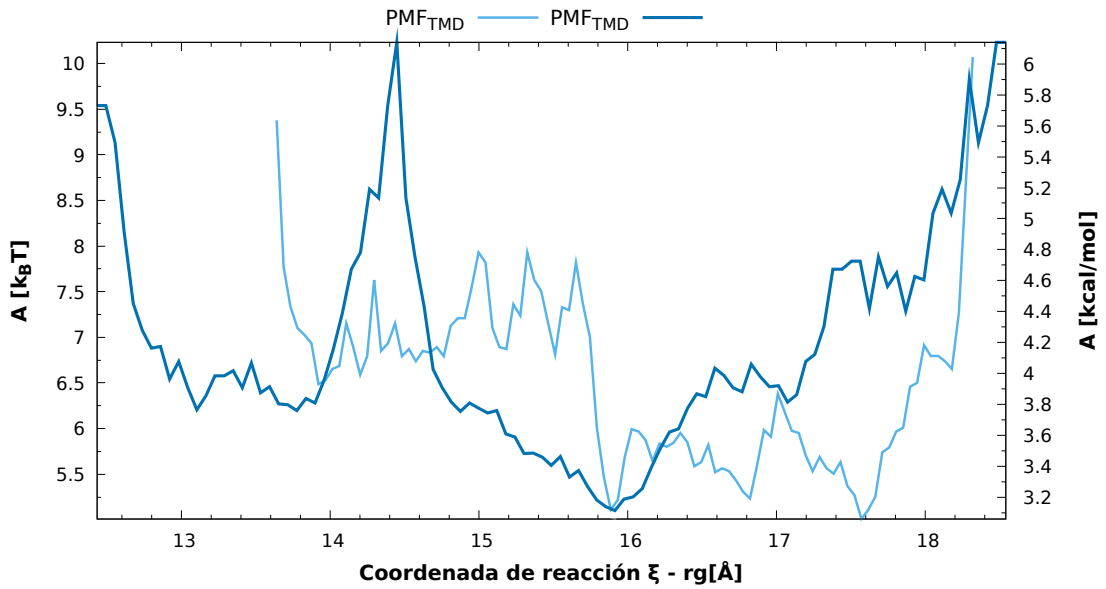
Figura 3.4: *Backbone* del lazo 36 (a) cerca de la región de la barrera de potencial y (b) en las proximidades del mínimo de energía.

3.1.3. Comparación de coordenadas de reacción

Como se menciona en [9], para obtener una curva PMF confiable es necesario seleccionar una coordenada de transición válida. Entonces para poder comprobar la validez de la coordenada de reacción r_{cm} se comparó el comportamiento general



(a) PMF a lo largo del $RMSD$ obtenido a partir de la simulación TMD y obtenido a partir de las simulaciones de dinámica libre.



(b) PMF a lo largo del rg obtenido a partir de la simulación TMD y obtenido a partir de las simulaciones de dinámica libre.

Figura 3.5: Curvas PMF a lo largo del radio de giro y el RMSD obtenidas a partir de la simulación TMD y simulaciones de dinámica libre.

de las curvas PMF con respecto a otras coordenadas de reacción. Esto permitió adicionalmente comprobar el comportamiento de las regiones de alta y baja energía del PMF_{FD} de la figura 3.3. Siendo así, se construyó la curva PMF a lo largo de las coordenadas de reacción: $RMSD$ y radio de giro, rg. El procedimiento es similar al realizado en la anterior sección pero esta vez se a partir de las coordenadas espaciales de los átomos del *backbone* del lazo 36, se calculan los parámetros mencionados. Estos gráficos se muestran en la figura 3.5.

Al comparar estas tres curvas se observa que el orden de magnitud del rango de energía libre es consistente entre los PMF, entre 3 [kcal/mol] y 7 [kcal/mol]. Además, la barrera de potencial se observa independientemente de la coordenada de reacción que se utilice. Estas barreras no son simétricas y tienen una altura semejante, en particular para el rg la altura es 3.08[kcal/mol] cuando $rg = 14.45$[Å] y para el RMSD la altura es 2.99 [kcal/mol] cuando $RMSD = 6.84$[Å]. Adicionalmente, en 3.5 se puede observar que tanto para el RMSD como para el rg, la curva PMF_{TMD} y la curva PMF_{FD} no tienen un comportamiento tan similar como es el caso del PMF a lo largo de r_{cm}. De esta manera, se comprueba también que el PMF a lo largo de r_{cm} describe de manera más confiable la transición del lazo 36 ya que a pesar de que en la simulación TMD el sistema visita estados no accesibles de manera natural, estas discrepancias entre el PMF_{FD} y PMF_{TMD} son mínimas con respecto a cuando se grafican las curvas PMF a lo largo de las otras coordenadas, RMSD y rg.

Por otro lado, al comparar el orden de magnitud de los valores de energía libre en otras curvas PMF con referencias como [9, 17] se encuentran valores consistentes con los calculados en este trabajo.

En la curva PMF de la transición del subdominio de 35 residuos de la región superior de la proteína *Villina* (HP35) calculado en [9], se encontraron valores de energía libre del orden de 1 [kcal/mol], estos valores bajos relativos a los valores hallados en las figuras 2.4, 3.5 podrían ser debidos a la estructura de los sistemas estudiados. El lazo 36 tiene una estructura más compleja que la del sistema HP35 por tanto, de manera general se esperaría que los valores de energía libre involucradas sean mayores para el lazo 36. Además, en la curva PMF del estiramiento de la decalanina calculada en [17], se encontraron valores del orden de 20 [kcal/mol]. Estos valores relativamente altos, se explicarían con una diferencia encontrada en la metodología utilizada en el artículo mencionado. Los autores realizaron simulaciones en un ambiente de vacío, es decir, el sistema no se encontraba dentro de un medio de agua lo que resultaría en un crecimiento en los valores de energía libre altos.

3.2. Conclusiones y recomendaciones

El lazo 36 debe superar una barrera de potencial no simétrica de aproximadamente 2.5 [kcal/mol] de altura y un ancho de 1.6 [Å] para poder pasar de su estado parcialmente aleatorio hacia su estado helicoidal. Esta única barrera de potencial del PMF es consistente con las curvas PMF halladas en otras referencias para sistemas similares, como en [9] en donde se estudio el PMF a lo largo del radio de giro para el proceso de transición del sistema HP35. Por tanto encontrar una única barrera de potencial podría ser un aspecto característico de procesos de transición moleculares.

La combinación de una simulación TMD, muestreos *Umbrella* y análisis de grupos permitió obtener un PMF a lo largo de r_{cm} fiel a la transición del polipéptido ya que existió una coincidencia en el comportamiento general de esta curva, como las regiones de máximos y mínimos de energía, obtenida por dos tipos de simulaciones TMD y FD. Esto a la vez, permite concluir que la elección de una coordenada de reacción que describa de manera adecuada un proceso físico, es importante ya que así, se facilita la observación de regiones energéticas relevantes al proceso.

Por otro lado, los potenciales de sesgo utilizados en el muestreo *Umbrella* tradicional solamente son necesarios para sistemas con estructuras que exploran un rango extensivo de su espacio conformacional y por tanto que no permitan muestrear eventos pocos probables. El lazo 36 presentó estructuras visitadas que por su naturaleza, exploraban un rango de la coordenada de reacción en estudio con un solapamiento mínimo y por tanto muestreando una sección de ω restringida permitiendo el cálculo de la curva PMF sin la necesidad de correcciones debido a potenciales de sesgo.

El criterio de selección a través de una resolución de muestreo para definir una distancia de corte en el análisis de grupos permitió obtener 13 ventanas *iniciales*, el orden de magnitud de este número de ventanas es consistente con los valores que se suele usar en otras referencias como [9] en donde se divide al dominio de la coordenada de reacción en 10 regiones equidistantes. Sin embargo, debido a la naturaleza del sistema, analizar la similitud entre la estructuras que visita el lazo 36 durante una simulación TMD no es suficiente ya que obtienen estructuras con valores de coordenada de reacción cercanos lo que resulta en solapamientos del rango de coordenada de reacción y por tanto evita que se pueda realizar un muestreo completo de ω. Es decir, el criterio para definir el valor de la distancia de corte en el Análisis de Grupos puede ser mejorado.

La estrategia para muestrear la totalidad del espacio conformacional, fue realizar muestreos parciales en donde se escoge la siguiente ventana o estructura representativa a simular a partir del rango que exploraban otras estructuras en simulaciones previas. Mediante esta estrategia fue posible encontrar 10 ventanas con anchos desiguales con solapamientos del rango de coordenada de reacción mínimos y que por tanto permitieron obtener un muestreo amplio de ω. La diferencia relevante con otras referencias, como [9], es que las ventanas definidas para el muestreo *Umbrella* son equidistantes entre sí. Por tanto, dependiendo del sistema, al realizar una exploración del espacio conformacional para obtener una curva PMF mediante muestreos *Umbrella*, no es necesario realizar un muestreo con ventanas equidistantes. Además, es posible obtener un PMF fiel al cambio conformacional del lazo 36 mediante la información asociada a estructuras representativas a la transición.

Capítulo A

Anexos

A.1. Script para orientar la macromolécula con respecto al eje z

```
### Script to rotate the molecule and align it with the z-axis
### run as: vmd -dispdev text -e 0.rotation.tcl

set n 0
set dir ".."
set molname ${dir}/hgfh_${n}/hgfh_${n}

#Load hgfh_n_.pdb
mol load pdb ${molname}_.pdb

set com [atomselect top all]
set o [measure center $com weight mass]
$com moveby [vecscale 1.0 $o]

set sel89 [atomselect top "protein and resid 89 and name CA"]
set com89 [measure center $sel89 weight mass]

set sel75 [atomselect top "protein and resid 75 and name CA"]
set com75 [measure center $sel75 weight mass]

set v [vecsub $com75 $com89]
set uv [vecnorm $v]
```

```
set th [expr acos([lindex $uv 2])]
set ss [veccross $v {0 0 1}]
set rot [trans axis $ss $th rad]
$com move $rot

#$com writepsf ${molname}r.psf
$com writepdb ${molname}r.pdb

mol delete top
exit
```

A.2. Script para la generación del archivo PSF

```
#!/bin/sh

# Run the psfgen program, taking everything until "ENDMOL" as
    input.
# run as: sh ./1.psfgenerator

psfgen << ENDMOL

# Read in the topology definitions for the residues we will create.
# This must match the parameter file used for the simulation as
    well.

topology /home/Kevin/Documents/Sim2/2.Scripts/top_all27_prot_na.rtf

segment HA {
  pdb /home/Kevin/Documents/Sim2/hgfh_0/hgfh_0r.pdb
}

coordpdb /home/Kevin/Documents/Sim2/hgfh_0/hgfh_0r.pdb HA

writepsf /home/Kevin/Documents/Sim2/hgfh_0/hgfh_0.psf

# This ends the matching coordinate pdb file.  The psf and pdb
    files
# are a matched set with identical atom ordering as needed by NAMD.

writepdb /home/Kevin/Documents/Sim2/hgfh_0/hgfh_0.pdb

ENDMOL
```

A.3. Script para colocar al sistema en una caja de agua

```
### Script to immerse the molecule in a water box
#run as: vmd -dispdev text -e 2.waterbox.vmd

#Load hgfh_n.pdb
set n 0
set dir ".."
set molname ${dir}/hgfh_${n}/hgfh_${n}

package require solvate
solvate ${molname}.psf ${molname}.pdb -t 15 -o ${molname}_wbn

set dim [measure minmax [atomselect top all]]
set rmin [lindex $dim 0]
set x1 [lindex $rmin 0]
set y1 [lindex $rmin 1]
set z1 [lindex $rmin 2]
set rmax [lindex $dim 1]
set x2 [lindex $rmax 0]
set y2 [lindex $rmax 1]
set z2 [lindex $rmax 2]

puts "CENTER OF MASS: [measure center [atomselect top all] weight
   mass]"
puts "DIMENSIONS: [expr $x2-$x1] [expr $y2-$y1] [expr $z2-$z1]"
mol delete top

package require autoionize
autoionize -psf ${molname}_wbn.psf -pdb ${molname}_wbn.pdb -sc 0.1
   -o ${molname}_wb
mol delete top
exit
```

A.4. Archivo de configuración para la simulación de minimizacion de la energia

```
# MD configuration file for Energy minimization
## structure: lazo 36 de la hemaglutinina in a water box.
```

```
3
4  # Load the PDB file
5  set n 0
6  set dir ".."
7  set molname ${dir}/hgfh_${n}/hgfh_${n}
8
9  # molecular system
10 structure     ${molname}_wb.psf
11 coordinates     ${molname}_wb.pdb
12
13 temperature       0
14
15 # Input
16 paratypecharmm    on
17 parameters      par_all27_prot_na.prm
18
19 # Force-Field Parameters
20 exclude           scaled1-4
21 1-4scaling        1.0
22 switching         on
23 cutoff            12
24 switchdist        10
25 pairlistdist      13.5
26
27 #output
28 outputname    ${molname}_min
29 dcdfreq     1000
30 outputEnergies    50
31 outputPressure    50
32
33 # Fixed Atoms Constraint (set PDB beta-column to 1)
34 fixedAtoms         on
35 fixedAtomsCol     B
36 #fixedAtomsFile   ${molname}_wb.pdb
37
38 minimize    10000
```

A.5. Archivo de configuración para la simulación de calentamiento

```
1  # MD configuration file for Heating
2  ## structure: lazo 36 de la hemaglutinina in a water box.
```

```
3
4  # Load the PDB file
5  set n 0
6  set dir ".."
7  #set dir "/home/Kevin/Documents/Sim2"
8  set molname ${dir}/hgfh_${n}/hgfh_${n}
9
10 puts $molname
11
12 ## MOLECULAR SYSTEM
13 structure               ${molname}_wb.psf
14 coordinates             ${molname}_min.pdb
15 bincoordinates          ${molname}_min.coor
16
17 set temperature   0
18
19 ## SIMULATION PARAMETERS
20
21 # Input
22 paraTypeCharmm      on
23 parameters          par_all27_prot_na.prm
24
25 # Force-Field Parameters
26 exclude             scaled1-4
27 1-4scaling          1.0
28 switching           on
29 cutoff              12.
30 switchdist          10.
31 pairlistdist        13.5
32
33 # Integrator Parameters
34 timestep              1.0
35 rigidbonds    all
36 nonbondedFreq         1
37 fullElectFrequency    2
38 stepspercycle         10
39
40 #Periodic Boundary Conditions
41 cellBasisVector1  57.3 0 0
42 cellBasisVector2  0 62.101 0
43 cellBasisVector3  0 0 74.672
44 cellOrigin    -3.01 1.773 -8.35
45 wrapAll               on
```

```
46
47 #Constant temperature control
48 langevin    on
49 langevinTemp          $temperature
50 langevinDamping         5
51 langevinHydrogen      off    ;# don't couple langevin bath to
      hydrogens
52
53
54 # PME (for full-system periodic electrostatics)
55 PME                 yes
56 PMEGridSizeX        60
57 PMEGridSizeY        64
58 PMEGridSizeZ        75
59
60 # output
61 set outputname    ${molname}_heat
62 outputname          $outputname
63 dcdfreq           1000
64 outputenergies          100
65 outputPressure          100
66
67 # Constraints
68 constraints     on
69 consexp     2
70 consref     ${molname}_min.pdb
71 conskfile    ${molname}_min.pdb
72 conskcol    B
73
74 selectconstraints   on
75 selectconstrX   on
76 selectconstrY   on
77 selectconstrZ   off
78
79 # Heating protocol
80
81 temperature   $temperature
82 reassignfreq    1000
83 reassignincr    30 ; # = reassignfreq*reassignhold/pasos
84 reassignhold    300
85
86 run 10000 ; # = pasos
```

A.6. Archivo de configuración para la simulación de equilibracion

```
# MD configuration file for Equilibration
## structure: lazo 36 de la hemaglutinina in a water box.

# Load the PDB file
set n 0
set dir /home/Kevin
set molname ${dir}/hgfh_${n}/hgfh_${n}

set temperature    300
set outputname      ${molname}_eq01 ;#ha36_203_eq01

## MOLECULAR SYSTEM
structure               ${molname}_wb.psf
coordinates             ${molname}_min.pdb
bincoordinates          ${molname}_heat.coor
binvelocities           ${molname}_heat.vel

firsttimestep       0

## SIMULATION PARAMETERS

# Input
paraTypeCharmm        on
parameters            par_all27_prot_na.prm

# Force-Field Parameters
exclude               scaled1-4
1-4scaling            1.0
switching             on
cutoff                12.
switchdist            10.
pairlistdist          13.5

# Integrator Parameters
timestep    1.0
rigidbonds    all
nonbondedFreq         1
fullElectFrequency    2
```

```
stepspercycle          10

#Periodic Boundary Conditions

extendedSystem           ${molname}_heat.xsc
wrapall      on

#Constant temperature control
langevin     on
langevinTemp           $temperature
langevinDamping          5
langevinHydrogen       off     ;# don't couple langevin bath to
   hydrogens

# Constant Pressure Control (variable volume)
usegrouppressure  yes
useflexiblecell   no
useconstantarea   no

langevinPiston           on
langevinPistonTarget     1.01325 ;#  in bar -> 1 atm
langevinPistonPeriod     200.0
langevinPistonDecay      100.0
langevinPistonTemp       $temperature

# PME (for full-system periodic electrostatics)
PME                 yes
PMEGridSizeX        60
PMEGridSizeY        64
PMEGridSizeZ        75

#output
outputname           $outputname
dcdfreq           5000
outputenergies           100
outputPressure           100
#patchdim     4

# Monitor performance
outputTiming    5000
#twoAwayX yes
#twoAwayY yes
#twoAwayZ yes
```

```
81
82 # Constraints
83 constraints     on
84 consexp     2
85 consref     ${molname}_min.pdb
86 conskfile   ${molname}_min.pdb
87 conskcol    B
88
89 selectconstraints   on
90 selectconstrX   on
91 selectconstrY   on
92 selectconstrZ   on
93
94 run 500000
```

A.7. Archivo de configuración para la simulación de dinamica libre

```
1 # MD configuration file for Free dynamics
2 ## structure: lazo 36 de la hemaglutinina in a water box.
3
4 # Load the PDB file
5 set n 0
6 set dir /home/Kevin
7 set molname ${dir}/hgfh_${n}/hgfh_${n}
8
9 puts $molname
10
11 set temperature   300
12 set outputname      ${molname}_fd01
13
14 ## MOLECULAR SYSTEM
15 structure               ${molname}_wb.psf
16 coordinates             ${molname}_eq.pdb
17 bincoordinates          ${molname}_eq01.coor
18 binvelocities           ${molname}_eq01.vel
19
20 firsttimestep       0
21
22 ## SIMULATION PARAMETERS
23
24 # Input
```

```
paraTypeCharmm          on
parameters              par_all27_prot_na.prm

# Force-Field Parameters
exclude                 scaled1-4
1-4scaling              1.0
switching               on
cutoff                  12.
switchdist              10.
pairlistdist            13.5

# Integrator Parameters
timestep    1.0
rigidbonds    all
nonbondedFreq           1
fullElectFrequency      2
stepspercycle           10

#Periodic Boundary Conditions
extendedsystem    ${molname}_eq01.xsc
wrapAll                   on

#Constant temperature control
langevin    on
langevinTemp            $temperature
langevinDamping           5
langevinHydrogen        off    ;# don't couple langevin bath to
   hydrogens

# Constant Pressure Control (variable volume)
usegrouppressure  yes
useflexiblecell   no
useconstantarea   no

langevinPiston            on
langevinPistonTarget    1.01325 ;#  in bar -> 1 atm
langevinPistonPeriod    200.0
langevinPistonDecay     100.0
langevinPistonTemp      $temperature


# PME (for full-system periodic electrostatics)
PME                  yes
```

```
PMEGridSizeX        60
PMEGridSizeY        64
PMEGridSizeZ        75

#output
outputname          $outputname
dcdfreq          1000
outputenergies          100
outputPressure          100

# Constraints
constraints     on
consexp     2
consref     ${molname}_eq.pdb
conskfile   ${molname}_eq.pdb
conskcol    B

run 1000000
```

A.8. Script para el análisis de grupos implementado en R

```
#!/usr/bin/env Rscript

library(protoclust)

file <- "hgfh_tmd_2M_k300.str"
df <- read.table(file, header=FALSE)

d <- dist(df,method="euclidean")
hc <- protoclust(d)
h_ <- 15
cut <- protocut(hc, h = h_)
plot(hc, imerge = cut$imerge, col=4, cex=0.8, hang=-1, main=paste("
    Clusters of TMD structures for h=", as.character(h_)),xlab="
    Structure", sub="")
abline(h=h_, lty=2)
pr <- cut$protos[cut$cl]
print(h_)
print(ncol(pr))
print(table(pr))

```

```r
#### PCA analysis ####
library(tidyverse)
dfm <- mutate(df,cluster=pr)
hgfh_tmd.pca <- prcomp(df)
library(factoextra)

fviz_pca_ind(hgfh_tmd.pca, geom.ind = "point", pointshape = 21,
   pointsize = 0.5, habillage= dfm$cluster, addEllipses=TRUE,
   ellipse.level=0.9, repel=TRUE)+theme(text = element_text(size =
   15),axis.title = element_text(size = 20),axis.text =
   element_text(size = 20),panel.border=element_rect(fill=NA),
   plot.margin = margin(2, 2, 2, 2, "cm"),)
```

Referencias bibliográficas

[1] Thomas Andrew Waigh. *Applied Biophysics*. John Wiley & Sons, 9 2007.

[2] Amit Kessel and Nir Ben-Tal. *Introduction to Proteins*. 3 2018.

[3] Ulo Langel, Benjamin F. Cravatt, Astrid Graslund, N.G.H. Von Heijne, Matjaz Zorko, Tiit Land, and Sherry Niessen. *Introduction to Peptides and Proteins*. CRC Press, 11 2009.

[4] William Humphrey, Andrew Dalke, and Klaus Schulten. Vmd: Visual molecular dynamics. *Journal of Molecular Graphics*, 14(1):33–38, 1996.

[5] Chavela M. Carr, Charu Chaudhry, and Peter S. Kim. Influenza hemagglutinin is spring-loaded by a metastable native conformation. *Proceedings of the National Academy of Sciences of the United States of America*, 94(26):14306–14313, 12 1997.

[6] Chavela M. Carr and Peter S. Kim. A spring-loaded mechanism for the conformational change of influenza hemagglutinin. *Cell*, 73(4):823–832, 5 1993.

[7] Jan Hermans and Barry Lentz. *Equilibria and Kinetics of Biological Macromolecules*. John Wiley & Sons, 12 2013.

[8] Johannes Kästner. Umbrella sampling. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 1(6):932–942, 5 2011.

[9] Qing Wang, Tuo Xue, Chunnian Song, Yan Wang, and Guangju Chen. Study on the application of the combination of tmd simulation and umbrella sampling in pmf calculation for molecular conformational transitions. *International Journal of Molecular Sciences*, 17:692, 5 2016.

[10] Susanna Hug. *Classical Molecular Dynamics in a Nutshell*, pages 127–152. Humana Press, Totowa, NJ, 2013.

[11] James G. Phillips, David J. Hardy, Julio D.C. Maia, John H. Stone, João Ribeiro, Rafael C. Bernardi, Ronak Buch, Giacomo Fiorin, Jérôme Hénin, Wei Jiang, Ryan McGreevy, Marcelo C. R. Melo, Brian K. Radak, Robert D. Skeel, Abhishek Singharoy, Yi Wang, Benoît Roux, Aleksei Aksimentiev, Zaida Luthey-Schulten, Laxmikant V. Kale, Klaus Schulten, Christophe Chipot, and Emad Tajkhorshid. Scalable molecular dynamics on CPU and GPU architectures with NAMD. *Journal of Chemical Physics*, 153(4):044130, 7 2020.

[12] Helen M. Berman, Tammy Battistuz, Talapady N. Bhat, Wolfgang F. Bluhm, Philip E. Bourne, Kyle Burkhardt, Zukang Feng, Gary L. Gilliland, Lisa Iype, Shri Mohan Jain, Phoebe Fagan, Jessica Marvin, David Padilla, Ravichandran Veerasamy, Bohdan Schneider, Narmada Thanki, Helge Weissig, John D. Westbrook, and Christine Zardecki. The Protein Data Bank. *Acta Crystallographica Section D-biological Crystallography*, 58(6):899–907, 5 2002.

[13] Jürgen Schlitter, Marc Engels, and Peter Krüger. Targeted molecular dynamics: A new approach for searching pathways of conformational transitions. *Journal of Molecular Graphics*, 12(2):84–89, 6 1994.

[14] Frank Jensen. *Introduction to Computational Chemistry*. Wiley, 11 2006.

[15] Brian S. Everitt, Sabine Landau, Morven Leese, and Daniel Stahl. *Cluster Analysis*. Wiley, 2 2011.

[16] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022.

[17] Sang-Hyun Park, Fatemeh Khalili-Araghi, Emad Tajkhorshid, and Klaus Schulten. Free energy calculation from steered molecular dynamics simulations using Jarzynski's equality. *Journal of Chemical Physics*, 119(6):3559–3566, 8 2003.