ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE CIENCIAS

EVOLUCIÓN DE LA ENTROPÍA DURANTE LA TRANSICIÓN CONFORMACIONAL DEL LAZO 36 DE LA HEMAGLUTININA

TRABAJO PREVIO A LA OBTENCIÓN DEL GRADO DE MAGISTER EN FÍSICA

TESIS

KLÉVER DAVID CAJAMARCA SACTA
klever.cajamarca@epn.edu.ec

DIRECTOR: MARCO VINICIO BAYAS REA, PhD
marco.bayas@epn.edu.ec

Quito, Octubre 2017

Declaración del Autor

Yo, KLÉVER DAVID CAJAMARCA SACTA, declaro que el trabajo aquí escrito 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.

A través de la presente declaración cedo mis derechos de propiedad intelectual correspondientes a este trabajo, a la Escuela Politécnica Nacional, según lo establecido por la Ley de Propiedad Intelectual, por su reglamento y por la normatividad institucional vigente.

KLÉVER DAVID CAJAMARCA SACTA

CERTIFICACIÓN

Certifico que el presente trabajo fue desarrollado por KLÉVER DAVID CAJAMARCA SACTA, bajo mi supervisión.

_____

Marco Vinicio Bayas Rea, PhD.
Director del Proyecto

Abstract

The transition between the non-fusogenic and fusogenic conformations of loop 36 of the HA2 chain of Hemagglutinin has been studied using the trajectories of molecular dynamics simulations of the loop obtained in a previous work. The original analysis of the data was thouroughly revisited and diferent analytical techniques were used for its analysis. First, the clustering method was evaluated using the Euclidean distance between structures in Ramachandran space, instead of the RMSD, finding evidence of existence of additional intermediate conformations. Importantly, the so called quasi-harmonic approximation was used to estimate entropy. The values of entropy obtained, are in the order of $17\frac{kcal}{mol^{\circ}K}$, which are one order of magnitude higher than the results obtained in previous works. The free energy was estimated based on this improved calculation of the entropy, proving that the term $-T\Delta S$ contributes in a non-negligible way to the free energy. The internal energy term, was estimated using an ensemble average, over the equilibration MD trajectories of each intermediate conformation, as a result the height of the energy barrier for the transition was estimated in $189{,}75\frac{kcal}{mol}$.

Resumen

La transición entre las conformaciones no-fusogénica y fusogénica del lazo 36 de la cadena HA2 de la Hemaglutinina ha sido estudiada utilizando las trayectorias de simulaciones de dinámica molecular obtenidas en un trabajo previo. El análisis original de los datos se revisó exhaustivamente y se utilizaron diferente técnicas analíticas para su análisis. En primer lugar se evaluó el método de clustering utilizado utilizando como parámetro la distancia euclídea entre estructuras en el espacio de Ramachandran, en lugar del RMSD, encontrándose evidencia de la existencia de conformaciones intermedias adicionales.. De forma importante, la llamada aproximación quasi-armónica se utilizó para estimar la entropía. Los valores de entropía obtenidos, se encuentran en el orden de $17\frac{kcal}{mol^{\circ}K}$, los cuales son mayores en un orden de magnitud a los resultados obtenidos en trabajos anteriores. Se estimó la energía libre, a partir del cálculo mejorado de la entropía, comprobando que el término $-T\Delta S$ contribuye de forma no despreciable a la energía libre. El término de energía interna, se estimó a partir del promedio sobre el ensamble, aplicado sobre las trayectorias de dinámica molecular de equilibración para cada conformación intermedia, como resultado la altura de la barrera energética para la transición se estimó en $189{,}75\frac{kcal}{mol}$.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

1.1. Bases Biofísicas

1.1.1. Estructura de las Proteínas

Una proteína es un polímero, formado a partir de un conjunto de los aminoácidos. Un aminoácido es una molécula formada por un átomo de carbono central (C_α), el cual tiene enlazado un grupo amino (NH_3^+) en uno de sus lados, un grupo carboxilo (COO^-) en el otro, un átomo de Hidrógeno (H), y una cadena lateral variable que se denota mediante una R, la cual diferencia a los aminoácidos, y les confiere sus propiedades físico-químicas características (fig. 1.1). Hay 20 diferentes tipos de aminoácidos, presentes en todas las formas de vida conocidas, cada uno con su cadena lateral R característica.

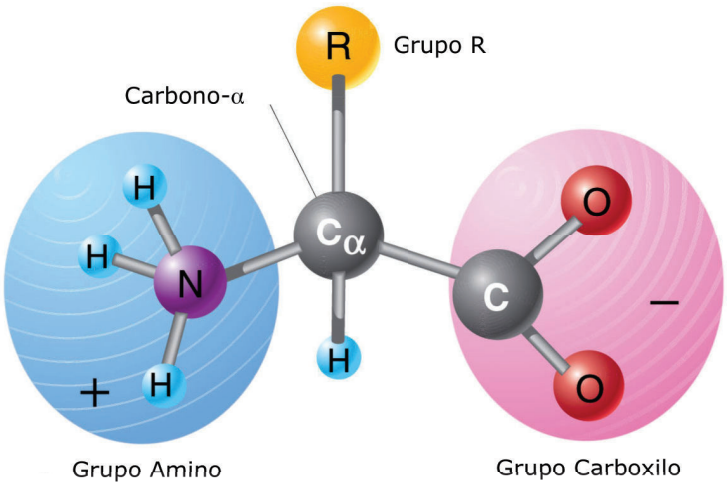

FIGURA 1.1: Estructura de un Aminoácido, imagen adaptada de [1]

Cuando los aminoácidos reaccionan para formar un polímero, el grupo amino de un aminoácido se combina con el grupo carboxilo de otro, de forma que un átomo de hidrógeno del grupo amino se separa y se combina con el grupo hidroxilo del grupo carboxilo, el cual también se separa, esto produce una molécula de agua. Al mismo tiempo se forma un enlace covalente entre el nitrógeno del grupo amino y el carbono del grupo carboxilo, este enlace es conocido como *enlace peptídico*. Como resultado de la naturaleza del enlace peptídico, los átomos de Carbono C y Oxígeno O en el primer residuo, y los átomos de Nitrógeno N e Hidrógeno H en el segundo residuo son coplanares [2, 3] (fig. 1.4).

La cadena de residuos formada, tiene una secuencia de átomos que se repiten: $-N-C_\alpha-C-$, como se indica en la figura 1.2 . Esta secuencia se conoce como el *backbone* de la proteína. Los residuos en los extremos tienen grupo que no han reaccionado, y se les conoce como terminales, existiendo el terminal amino o *terminal N* y el terminal carboxilo o *terminal C*.

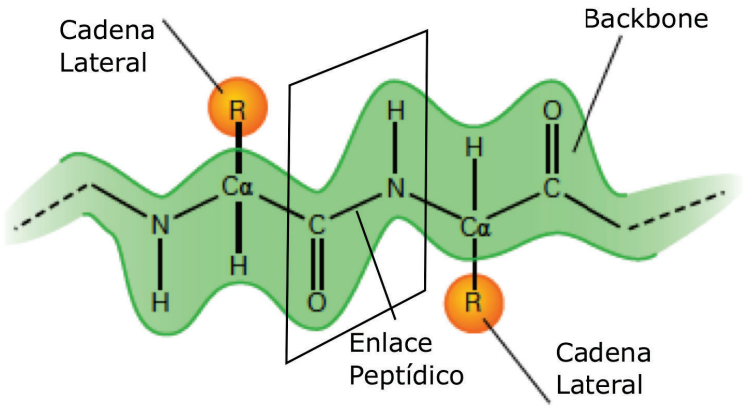

FIGURA 1.2: El *backbone* de una proteína, imagen adaptada de [4]

Una vez que el polipéptido se ha formado, sufre cambios conformacionales, hasta que adquiere una estructura particular. Uno de los problemas fundamentales de la biofísica, es como predecir la estructura, y posiblemente su función, a partir de su secuencia de aminoácidos. La importancia de este problema, se debe a que las proteínas tienen un papel central en todos los procesos en los organismos vivos. Se encargan de regular procesos biológicos, proveen soporte estructural, transportan otras moléculas, etc.

Las proteínas pueden actuar como catalizadores de reacciones bioquímicas, regulando su velocidad. Las proteína también pueden tener la función de transportar a otras moléculas, siendo el ejemplo más relevante, el de la hemoglobina, cuya función es realizar el transporte de oxígeno hacia cada célula en el organismo. De manera similar, las lipoproteínas, transportan lípidos hacia donde son necesarios, frecuentemente transportándoles a través de sitios en los cuales serían de otra manera detenidos. Las proteínas también ayudan al transporte de iones y

otras moléculas pequeñas a través de las membranas celulares. Las proteínas también tienen funciones estructurales y de movimiento, por ejemplo en la formación y contracción de los músculos.

1.1.2. Niveles de Organización

La estructura de las proteínas puede ser descrita a varios niveles de complejidad, analizándolas desde la mas elemental hasta la más compleja. Bajo este criterio, se han definido cuatro niveles de organización [3], la secuencia de la cadena lineal de aminoácidos en un polipéptido es conocida como la *estructura primaria*, la *estructura secundaria* consiste de los patrones o motivos estructurales estables que forma la cadena de aminoácidos de forma local, la *estructura terciaria* abarca todos los aspectos de la conformación espacial tridimensional de la proteína, finalmente se puede hablar de una *estructura cuaternaria* cuando la proteína consta de dos o mas subunidades. Esta organización se muestra de forma esquemática en la figura **??**:

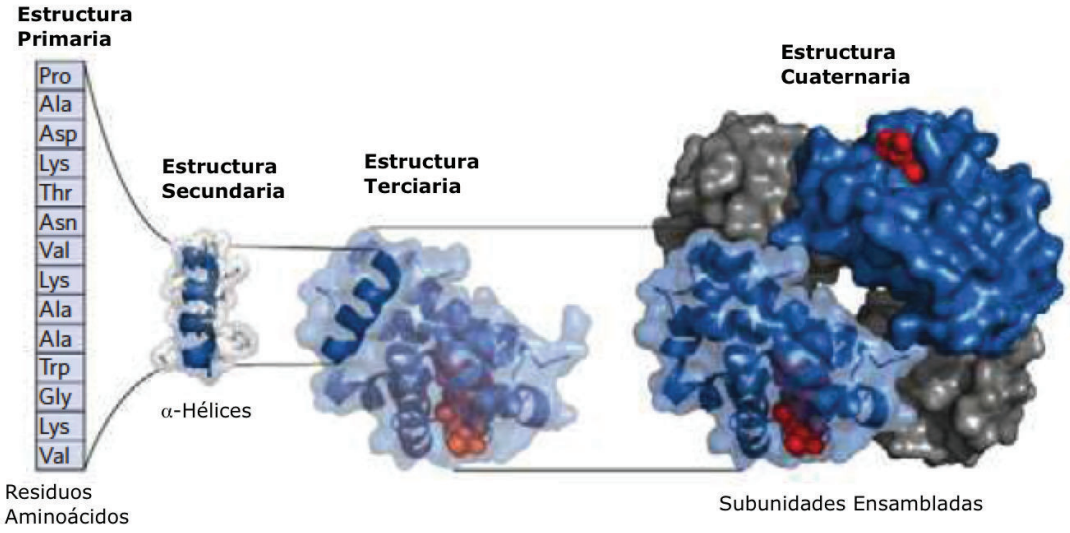

FIGURA 1.3: Niveles de Organización en Proteínas, imagen adaptada de [3]

Los grupos funcionales amino y carboxilo que constituyen un aminoácido, son fácilmente ionizados a un valor fisiológico de pH, el grupo amino por la adición de un proton, y el grupo carboxilo por la pérdida de un protón. El resultado se conoce como un *zwiterión*, una molécula ionizada que tiene carga positiva y negativa, pero una carga neta nula. En la cadena peptídica, los grupos amino y carboxilo participan en la formación de enlaces covalentes, eliminando la posibilidad de que se ionizen, excepto en los extremos de la cadena.

De los 20 aminóacidos que forman las proteínas, 5 tienen cadenas laterales R, que pueden estar ionizadas a pH fisiológico. En particular, el aspartato y el glutamato tienen carga negativa, en tanto que la histidina, lisina, y arginina tienen carga positiva. Estos aminoácidos pueden influir en la estructura proteínica de varias formas, por ejemplo pueden formar puentes salinos, es decir pueden formar un enlace iónico entre un ion positivo y negativo. Otro mecanismo es la atracción o repulsión electrostática de las cadenas laterales con grupos ionizados de la molécula.

1.1.3. Mapas de Ramachandran

Los ángulos de Ramachandran [5], representan la orientación de los aminoácidos unidos mediante enlaces peptídicos en un polipéptido o proteína. Estos ángulos permiten describir las rotaciones del *backbone o esqueleto* del polipéptido en torno a los enlaces $N - C_\alpha$ (llamado ϕ) y $C_\alpha - C$ (ψ). Como consecuencia la cadena peptídica que constituye una proteína puede considerarse como una sucesión de planos cuya orientación relativa se define con dos ángulos denotados con ϕ y ψ.

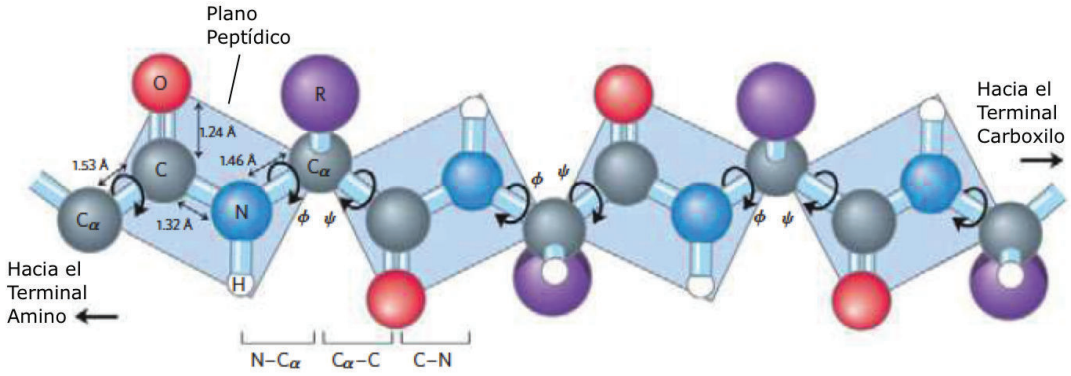

FIGURA 1.4: Ángulos de Ramachandran, imagen adaptada de [3]

La representación en un plano de ϕ vs ψ, se conoce como mapa de Ramachandran, y proporciona una manera fácil de ver la distribución de los ángulos de torsión en una estructura de la proteína. También proporciona una visión general de las regiones excluidas, que muestran que ciertas rotaciones del polipéptido no están permitidas, debido a la interacción entre los átomos de los aminoácidos. Finalmente, en el mapa de Ramachandran se pueden identificar regiones correspondientes a motivos de la estructura secundaria de una proteína, tal como las regiones correspondientes a los denominados α-hélices y hojas-β, estas regiones se muestran en la figura 1.5.

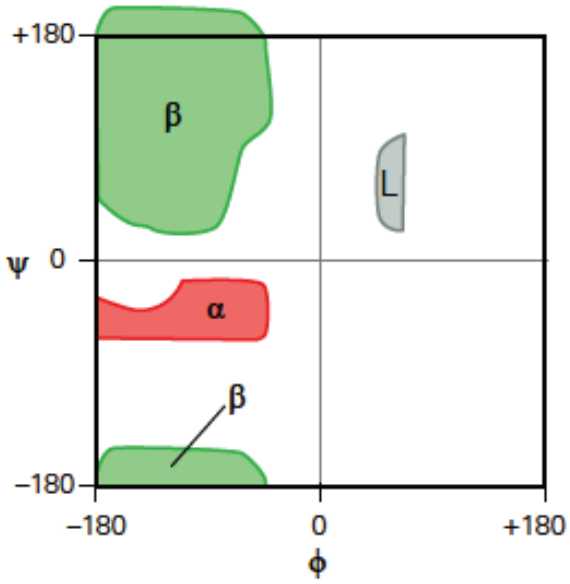

FIGURA 1.5: Mapa de Ramachandran, con las regiones correspondientes a α-hélices y hojas-β, imagen adaptada de [4]

1.1.4. α-Hélices

Las estructura de las α-hélices en proteínas, fue descubierta por Linus Pauling en 1951 [6], el cual predijo que su estructura era estable y energéticamente favorable; esto se debe a que una α-hélice es el arreglo más simple que el polipéptido puede asumir, tal que se maximizan los enlaces de hidrógeno internos. En esta estructura el *backbone* de la proteína se envuelve en torno a un eje imaginario que pasa longitudinalmente por el centro de la hélice, de tal forma que los grupos laterales R de los residuos quedan en la parte exterior de la hélice.

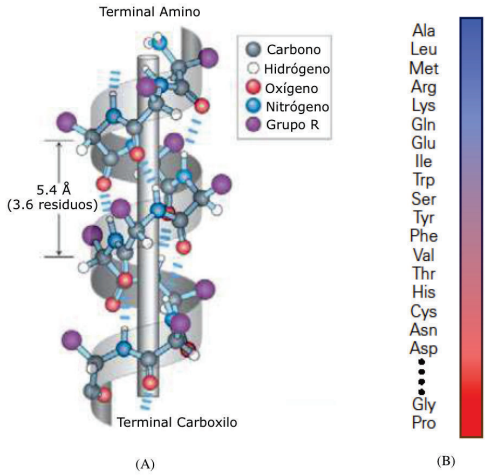

FIGURA 1.6: (A) Estructura de un α-hélice, (B) Propensión de residuos para conformar α-hélices, imagen adaptada de [3]

Las α-hélices pueden variar considerablemente en longitud, desde 4 a 5 residuos, hasta mas de 40 residuos en las proteínas globulares; la longitud promedio de un α-hélice es de alrededor de 10 residuos, correspondientes a tres vueltas de la hélice [2].

El polígono que encierra a la región correspondiente a un α-hélice en el mapa de Ramachandran esta determinada por los puntos indicados en la siguiente tabla [2]:

ϕ [grados]	ψ [grados]
-180.0	-34.9
-164.3	-42.9
-133.0	-42.9
-109.4	-32.2
-106.9	-21.4
-44.3	-21.4
-44.3	-71.1
-180.0	-71.1

TABLA 1.1: Puntos del polígono que encierra la región correspondiente a un α-hélice en el mapa de Ramachandran [2]

1.1.5. Plegado Proteínico

La forma tridimensional que adquiere la proteína en condiciones fisiológicas, se conoce como la forma *plegada*; y la caracterización del proceso de plegado, así como la predicción de la estructura plegada a partir de la cadena lineal de aminoácidos, se conoce como el problema del *plegado proteínico* [7, 8, 9].

Cuando una proteína está plegada en su forma correcta, se dice que está en su estado *nativo*, este estado puede contener un número pequeño de conformaciones que están involucradas en el funcionamiento biológico de la proteína. Cuando una proteína no está plegada, es decir no está en su estado nativo, se dice que está *denaturalizada*.

Varias proteínas requieren asistencia para plegarse, esta asistencia la proveen otras moléculas llamadas moderadores de plegado o *chaperones*, para llegar a su estado nativo. En algunos casos los chaperones catalizan un paso en el proceso de plegado, que de otra manera, ocurriría muy lentamente. En otros casos los chaperones se adhieren a la cadena peptídica y unen ciertas partes de la molécula, o proveen un entorno que favorece el estado nativo de la proteína.

1.1.6. El Entorno Energético

El entorno energético es una "hiper-superficie" que asocia la estructura o conformación de la biomolécula, con la energía, en particular con la energía libre de Gibbs. El concepto de entorno energético, esta estrechamente relacionado con las ideas de Levinthal[10] y Anfinsen[11], en relación a la termodinámica del proceso de plegado proteínico; planteándose la hipótesis que las estructuras biológicamente funcionales, corresponden a mínimos del entorno energético, y que los caminos sobre la superficie del entorno energético corresponden a trayectorias de plegado.

1.2. El Lazo 36 de la Hemaglutinina

La Hemaglutinina (HA) es una proteína presente en la superficie del virus de la influenza, siendo parte fundamental del mecanismo de adhesión y fusión del virus a la membrana celular [12, 13]. La Hemaglutinina esta formada por dos subunidades HA1 y HA2, las cuales se encuentran unidas mediante un enlace disulfuro. La estructura molecular de la Hemaglutinina ha sido caracterizada completamente de forma experimental [12, 14], y la información completa de su composición y estructura molecular se encuentran libremente disponibles en el Protein Data Bank (PDB) [15], tanto para la configuración no-fusogénica, como para la fusogénica.

De particular interés resulta la dinámica del cambio conformacional [16, 17, 18, 19] que sufre el llamado *Lazo 36*, el cual corresponde a la región comprendida entre los residuos 54 y 89 de la cadena HA2 de la Hemaglutinina. El lazo 36 de la Hemaglutinina consiste de 36 residuos (fig. 1.4), comprendiendo la region entre los residuos $54(ARG)$ y $89(ILE)$, cabe notar que la región comprendida entre los residuos $76(ARG)$ y $89(GLU)$ tiene una estructura de α-hélice en el estado no fusogénico. El residuo $75(GLY)$ corresponde a la glicina el cual es el aminoácido más simple que existe, lo cual también hace que sea el único aminoácido presente en las proteínas de naturaleza no-quiral; la glicina junto a la prolina son los aminoácidos con menor propensión a formar α-hélices [2, 3, 4]. El residuo 75 separa al lazo 36 en dos secciones, una cuya estructura es de α-hélice, y otra que forma un lazo aleatorio, comprendido entre los residuos $54(ARG)$ y $74(GLU)$ [25, 13].

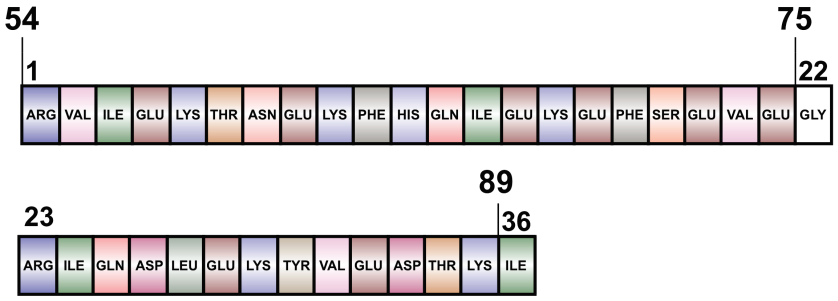

FIGURA 1.7: Secuencia del Lazo 36, la región comprendida entre los residuos 76 al 89 se encuentran en estructura de tipo α-hélice durante toda la transición conformacional

El modelo con mayor aceptación para la descripción de la transición conformacional es el conocido como *Spring-Loaded* [13], el cual plantea que la estructura nativa de la Hemaglutinina está atrapada en un estado metaestable, y que el mecanismo para liberarlo es disparado por la disminución del pH, el cual desestabiliza la estructura nativa y produce la transición. El cambio ocurre desde su conformación inicial no-fusogénica a $pH \sim 7$ hasta una conformación final fusogénica a $pH \sim 5$. Se ha planteado que el movimiento lineal producido por la transición conformacional, pueda ser aprovechado como un nanomotor lineal [20, 21, 22, 23].

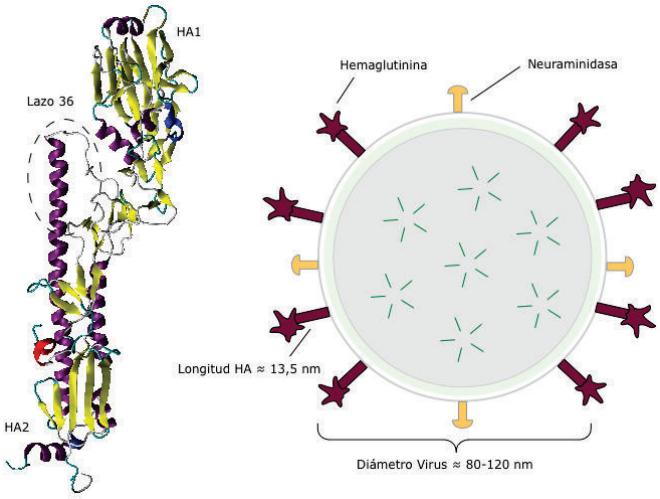

FIGURA 1.8: (A) La Hemaglutinina, con sus subunidades HA1 y HA2, el lazo 36 se encuentra en la subunidad HA2, imagen realizada utilizando VMD [24]. (B) La Hemaglutinina en el virus de la influenza

La siguiente figura muestra la estructura del lazo 36, tanto en su forma no-fusogénica, como fusogénica.

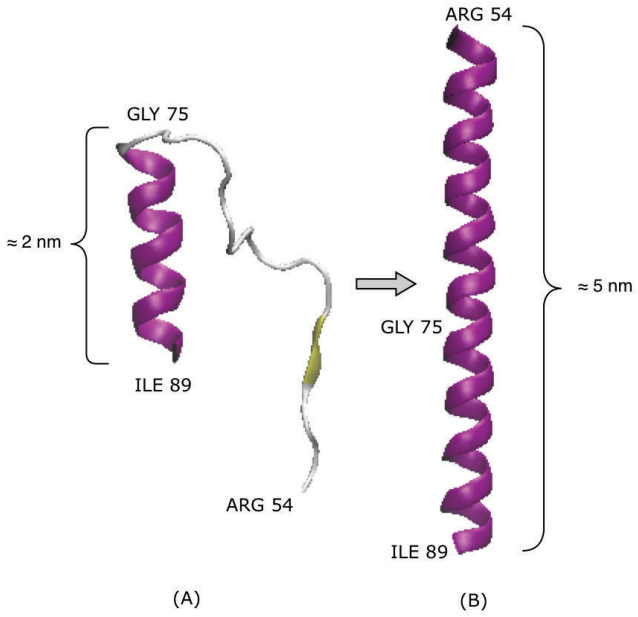

FIGURA 1.9: Estructura Secundaria del Lazo 36: (A) Conformación no-fusogénica a pH 7, imagen generada mediante el software VMD [24] a partir de la estructura 1HGF [14] del PDB [15]. (B) Conformación fusogénica a pH 5, imagen generada mediante el software VMD [15], a partir de la estructura 1HTM [12] del PDB [15]

1.3. Transición Conformacional

La información disponible para el lazo 36 en el PDB ha facilitado su estudio mediante simulaciones computacionales de dinámica molecular [19, 18, 16, 26]. Particularmente, la transición del lazo se ha estudiado utilizando simulaciones de dinámica molecular orientada (TMD: Targeted Molecular Dynamics) [26, 19]. Con este tipo de simulaciones se genera la transición entre dos estructuras mediante la aplicación de una fuerza *orientadora* proporcional a la distancia media cuadrática (RMSD) entre las posiciones de los átomos para cada instante de simulación y las correspondientes posiciones de la estructura objetivo [**TMD**]. La aplicación de simulaciones TMD al estudio de la transición del lazo 36 ha permitido identificar posibles estados intermedios y el costo energético de la misma.

En el estudio realizado por Calderon J. [26] las estructuras generadas por las simulaciones TMD fueron inicialmente agrupadas mediante la técnica de análisis de grupos [27, 28] en 11 grupos. Cada uno de estos grupos se asoció con una conformación accesible al polipéptido durante la transición desde el estado no-fusogénico al fusogénico. Posteriormente se estudió la evolución de la estructura representativa de cada conformación identificada mediante simulaciones de dinámica molecular. Las estructuras obtenidas se sometieron a un análisis de

grupos obteniéndose lo que se denominó *pseudoconformaciones*. En el trabajo citado [26] el número de pseudoconformaciones se utilizó para la estimación de la entropía, obteniéndose una contribución de la entropía a la energía libre del orden de $1,2kcal/mol$. Sin embargo, este cálculo depende del criterio utilizado para la clasificación de las estructuras asociadas a cada conformación, específicamente de la métrica utilizada para definir la distancia entre dos estructuras. Por lo tanto un cálculo independiente de la entropía se hace necesario.

En vista de que el análisis de los datos realizado anteriormente no se ha terminado de forma adecuada, se realizó un nuevo análisis de los datos. Para este nuevo análisis en primer lugar se comparó los resultados del método de agrupamiento desarrollado por Calderon J.[26] con los resultados de un agrupamiento basado en la similitud de estructuras secundarias en el espacio de Ramachandran, posteriormente se estimó la entropía del lazo 36 de la HA utilizando el método de la Aproximación Quasi-Armónica y finalmente se estimó la energía libre con el fin de probar varios candidatos a coordenadas de reacción para caracterizar la barrera energética asociada a la transición conformacional.

La transición conformacional del lazo 36 de la Hemaglutinina es un proceso que, por su simplicidad, permite estudiar el plegado proteínico. Los detalles atómicos de la transición accesibles directamente con las simulaciones de dinámica molecular, permiten el cálculo de la energía, así como de la entropía involucradas en el proceso. Esta información permitirá establecer relaciones cuantitativas entre estas dos magnitudes durante un proceso de plegado. Particularmente el entendimiento de los mecanismos atómicos asociados a la transición conformacional del lazo 36 de la Hemaglutinina tiene potenciales aplicaciones biomédicas y biotecnológicas.

Capítulo 2

Aproximación Quasi-Armónica

Dado que no existe un método exacto para calcular la entropía de una biomolécula a partir de los datos de dinámica molecular, es necesario realizar una estimación de la misma. La estrategia que su utilizó en este trabajo para estimar la entropía, es la llamada Aproximación Quasi-Armónica (QHA) [29], la cual provee un límite superior a la entropía conformacional de una biomolécula en términos de osciladores armónicos independientes asociados a sus grados de libertad, identificados con un análisis de componentes principales [30]. Para la descripción de la aproximación quasi-armónica es necesario primero introducir la técnica del análisis de componentes principales.

2.1. Análisis de Componentes Principales (PCA)

El Análisis de Componentes Principales es una técnica estadística, utilizada para analizar datos multivariantes. Los objetos prácticos de la utilización del análisis de componentes principales han sido enumerados por Jeffers [31], siendo de interés para el presente estudio, examinar la correlación entre las variables de un conjunto de datos.

En el análisis de componentes principales se parte de un conjunto de variables correlacionadas $\boldsymbol{x} = (x_1, \ldots, x_n)$, estas se expresan en un nuevo conjunto de variables no-correlacionadas $\boldsymbol{y} = (y_1, \ldots, y_n)$ conocidas como las *componentes principales*, las cuales son combinaciones lineales de las variables originales. Estas componentes principales obtenidas se encuentran ordenadas de tal forma que la primera componente principal y_1 contiene la mayor cantidad de varianza posible de los datos entre todas las combinaciones lineales de las variables de $\boldsymbol{x}$, en tanto que y_2 tiene la mayor varianza restante, y así sucesivamente. Por esto, el PCA es útil en ciertas aplicaciones para reducir el número de dimensiones básicas requeridas para describir el conjunto, de forma que se puedan reducir a un número mínimo de dimensiones significativas, tomando en cuenta las primeras componentes principales que contienen la mayor varianza de los datos, e ignorando las demás [31]. En el presente estudio, se utilizaron todas las

componentes principales, ya que para la aplicación de la aproximación quasi-armónica para el cálculo de la entropía es únicamente necesario tener variables no-correlacionadas.

2.1.1. Matriz de Datos y Matriz de Covarianza

Considerando un conjunto de datos, los cuales se pueden representar mediante $\boldsymbol{x}$, el cuál es un vector estocástico de componentes x_j con $j = 1, \ldots, n$, el cual a su vez tiene una distribución de probabilidad subyacente $P(\boldsymbol{x})$; se puede establecer una muestra $\boldsymbol{X} = \{\boldsymbol{x}_s | s = 1, \ldots, n\}$ de $P(\boldsymbol{x})$. Esta matriz $\boldsymbol{X}$ representa una matriz de datos arbitraria, cuyas filas denotan las muestras $s = 1, \ldots, n$, y cuyas columnas representan las variables (o coordenadas) de cada muestra.

Entonces se puede obtener el *primer momento* de la matriz de datos [32], como el vector promedio:

$$\langle \boldsymbol{x} \rangle = \frac{1}{n} \sum_{s=1}^{n} \boldsymbol{x}_s \tag{2.1}$$

De manera similar, se puede determinar el *segundo momento* de la matriz de datos, el cual es conocido como la *matriz de covarianza* [33]:

$$\boldsymbol{C} = \frac{1}{n} \sum_{s=1}^{n} (\boldsymbol{x}_s - \langle \boldsymbol{x} \rangle)(\boldsymbol{x}_s - \langle \boldsymbol{x} \rangle)^T \tag{2.2}$$

Esta matriz se puede expresar de forma compacta, definiendo una nueva variable $\boldsymbol{X'} = \{\boldsymbol{x}_s - \langle \boldsymbol{x} \rangle\}$:

$$\boldsymbol{C} = \frac{1}{n} \boldsymbol{X'} \boldsymbol{X'^T} \tag{2.3}$$

2.1.2. Descomposición de Valores Propios

Es conocido [34], que a partir de una matriz $\boldsymbol{A}$, se puede plantear un problema de valores propios de la forma:

$$\boldsymbol{A}\boldsymbol{u} = \lambda \boldsymbol{u} \tag{2.4}$$

$$(\boldsymbol{A} - \lambda \boldsymbol{I})\boldsymbol{u} = 0 \tag{2.5}$$

donde $\boldsymbol{u}$ es llamado vector propio o *autovector* y λ es un escalar llamado valor propio o *autovalor*, e $\boldsymbol{I}$ es la matriz identidad.

Por convención, se acostumbra reunir al conjunto de autovectores de $\boldsymbol{A}$ en una matriz $\boldsymbol{U}$, donde cada columna de $\boldsymbol{U}$ es un autovector de $\boldsymbol{A}$. Los autovectores se arreglan en una matriz

diagonal $\boldsymbol{\Lambda}$, de forma que se puede expresar el problema de valores propios de la forma:

$$\boldsymbol{AU} = \boldsymbol{\Lambda U} \tag{2.6}$$

Entonces a partir de esta expresión, se puede obtener la descomposición de valores propios de una matriz $\boldsymbol{A}$:

$$\boldsymbol{A} = \boldsymbol{U\Lambda U^{-1}} \tag{2.7}$$

2.1.3. Matrices Positivas Semi-definidas

Una matriz $\boldsymbol{A}$ es positiva semi-definida cuando puede ser expresada mediante el producto de una matriz por su transpuesta, es decir:

$$\boldsymbol{A} = \boldsymbol{X X^T} \tag{2.8}$$

Esta expresión implica que una matriz positiva semi-definida es siempre simétrica (es decir $\boldsymbol{A} = \boldsymbol{A^T}$). Además los autovalores de una matriz positiva semi-definida son siempre positivos o nulos [34], y sus autovectores son siempre ortogonales cuando los autovectores correspondientes son diferentes; por tanto, si se arreglan los autovectores en una matriz $\boldsymbol{U}$, la misma será ortogonal, es decir:

$$\boldsymbol{UU^T} = \boldsymbol{I} \tag{2.9}$$

Por lo tanto:

$$\boldsymbol{U^{-1}} = \boldsymbol{U^T} \tag{2.10}$$

Entonces la descomposición en valores propios de la matriz ($\boldsymbol{A} = \boldsymbol{U\Lambda U^{-1}}$) puede expresarse de la forma:

$$\boldsymbol{A} = \boldsymbol{U\Lambda U^T} \tag{2.11}$$

La matriz de covarianza es una matriz positiva semi-definida.

2.1.4. Descomposición de Valor Singular (SVD)

La descomposición de valor singular (SVD) es una generalización de la descomposición de valores propios. La SVD descompone una matriz rectangular en tres matrices: dos matrices ortogonales y una matriz diagonal. Si $\boldsymbol{A}$ es una matriz rectangular, entonces tiene una descomposición del tipo [34]:

$$\boldsymbol{A} = \boldsymbol{U\Gamma V^T} \tag{2.12}$$

donde $\boldsymbol{U}$ es la matriz de autovectores normalizados de la matriz $\boldsymbol{AA^T}$, $\boldsymbol{V}$ es la matriz de autovectores normalizados de la matriz $\boldsymbol{A^TA}$, y Γ es la matriz diagonal que contiene los *valores singulares*, de forma que: $\boldsymbol{\Gamma} = \boldsymbol{\Lambda^{\frac{1}{2}}}$ donde $\boldsymbol{\Lambda}$ es la matriz diagonal de autovalores de la matriz $\boldsymbol{AA^T}$ y de la matriz $\boldsymbol{A^TA}$, ya que los dos conjuntos de autovalores son los mismos [33].

2.1.5. PCA mediante SVD

Una matriz de datos $\boldsymbol{X}$ de dimensión $n \times m$ se puede descomponer mediante SVD de la forma $\boldsymbol{X} = \boldsymbol{U\Gamma V^T}$, por tanto la matriz de covarianza se puede expresar [30, 32]:

$$\boldsymbol{C} = \frac{1}{n}(\boldsymbol{U\Gamma V^T})(\boldsymbol{V\Gamma U^T}) = \frac{1}{n}(\boldsymbol{U\Gamma})(\boldsymbol{V^TV})(\boldsymbol{\Gamma U^T}) = \frac{1}{n}(\boldsymbol{U\Gamma})\boldsymbol{I}(\boldsymbol{\Gamma U^T}) \tag{2.13}$$

$$\boldsymbol{C} = \frac{1}{n}\boldsymbol{U\Gamma^2 U^T} \tag{2.14}$$

donde $\boldsymbol{U}$ es una matriz $n \times m$.

2.2. Componentes principales de una macromolécula derivadas de trayectorias de dinámica molecular

A partir de trayectorias de simulaciones de dinámica molecular, en las cuales se han eliminado los átomos del solvente, se puede plantear una matriz de trayectoria $\boldsymbol{R}$ la cual tiene dimensiones $3N_a \times n_f$ siendo N_a el número de átomos y n_f la cantidad de trayectorias de simulación que se consideran para el análisis, es decir:

$$\boldsymbol{R} = \begin{pmatrix} \boldsymbol{r_1} \\ \boldsymbol{r_2} \\ \boldsymbol{r_3} \\ \vdots \\ \boldsymbol{r_i} \\ \vdots \\ \boldsymbol{r_{nf}} \end{pmatrix} \tag{2.15}$$

donde $\boldsymbol{r_i}$ corresponde a la estructura i-ésima, la cual contiene las coordenadas de cada átomo, arregladas en un vector con $3Na$ componentes:

$$\boldsymbol{r_i} = (x_1^i, y_1^i, z_1^i, \ldots, x_{Na}^i, y_{Na}^i, z_{Na}^i) \tag{2.16}$$

Aplicando a la matriz de trayectoria R, una transformación de la forma [35]:

$$\boldsymbol{Y} = \boldsymbol{M^{1/2}R} \tag{2.17}$$

donde la matriz $\boldsymbol{M^{1/2}}$ de dimensiones $3N_a \times 3N_a$ viene dada por:

$$\boldsymbol{M^{1/2}} = \begin{bmatrix} \sqrt{m_1} & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & \sqrt{m_1} & & \dots & 0 & 0 & 0 \\ 0 & 0 & \sqrt{m_1} & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & 0 & 0 \\ 0 & 0 & 0 & \dots & \sqrt{m_{Na}} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sqrt{m_{Na}} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \sqrt{m_{Na}} \end{bmatrix} \tag{2.18}$$

siendo m_i la masa atómica del i-ésimo átomo [36].

Sobre esta matriz $\boldsymbol{Y}$ se aplica el análisis de componentes principales, para lo cual, en primer lugar, se obtiene la matriz de covarianza de $\boldsymbol{Y}$, cuyos elementos individuales vienen dados por:

$$\boldsymbol{C} = \frac{1}{n}\sum_{s=1}^{n}(\boldsymbol{y}^s - \langle \boldsymbol{y} \rangle)(\boldsymbol{y}^s - \langle \boldsymbol{y} \rangle)^T \tag{2.19}$$

En consecuencia la matriz de covarianza se puede representar de la forma:

$$\boldsymbol{C} = \frac{1}{n}\boldsymbol{Y}\boldsymbol{Y}^{\boldsymbol{T}} \tag{2.20}$$

Se aplica la descomposición de valor singular (SVD) a la matriz C , obteniendo un conjunto de $3N_a$ autovalores λ, los cuales debido a la ponderación realizada a través de la matriz M contienen valores de masa que son combinación de las originales [35, 37, 38], por lo cual los autovalores pueden representarse de la forma:

$$\lambda_i = m_{eff}\sigma^2_{PCA} \tag{2.21}$$

2.3. Frecuencias Asociadas a las Componentes Principales

Se puede demostrar (Apéndice A) que para un sistema con desviación estándar nula $\langle x \rangle = 0$ y varianza conocida $\langle x^2 \rangle$, el valor máximo de la entropía se presenta cuando las energías y las varianzas para cada estado son proporcionales, excepto por una constante aditiva. Esta

condición es satisfecha por el Hamiltoniano de un oscilador armónico [35, 37], es decir:

$$H = \frac{1}{2}(\frac{p^2}{m} + m\omega^2 x^2) \tag{2.22}$$

donde p es el momento canónico asociado a x.

Para este tipo de sistemas es de interés encontrar la frecuencia a partir de la varianza $\langle x^2 \rangle$. Para ello se puede utilizar la siguiente aproximación [35]:

$$m\omega^2 \langle x^2 \rangle = k_B T \tag{2.23}$$

Está expresión se cumple también en el régimen clásico, debido al teorema de equipartición.

Una generalización propuesta para el caso n-dimensional es a través de la matriz de covarianza [35], de forma que la expresión anterior se puede representar como:

$$m_{eff}\omega_n^2 \sigma_n^2 = k_B T \tag{2.24}$$

Las varianzas σ_n^2 pueden estimarse con simulaciones de dinámica molecular del sistema. Este tipo de simulaciones de dinámica molecular, se realizan usualmente en un sistema de coordenadas cartesiano, por lo cual varias de sus coordenadas estarán correlacionadas. Con el fin de obtener una transformación de coordenadas, las cuales sean independientes u ortogonales, se puede aplicar una transformación ortogonal, tal como el análisis de coordenadas principales, las cuales además están ordenadas desde la que contiene la mayor cantidad de varianza del sistema, hasta la que contiene la menor, de forma que:

$$m_{eff}\omega_i^2 \sigma_{i(PCA)}^2 = k_B T \tag{2.25}$$

Esta expresión corresponde a la hipótesis fundamental de la aproximación Quasi-Armónica, la cual considera que los autovalores obtenidos a partir del análisis de componentes principales, corresponden a un conjunto de osciladores armónicos cuánticos no-correlacionados [39, 40].

Desarrollando la ecuación anterior, para cada autovalor se puede obtener una frecuencia ω:

$$\omega_i = \sqrt{\frac{k_B T}{m_{eff}\sigma_{i(PCA)}^2}} = \sqrt{\frac{k_B T}{\lambda_i}} \tag{2.26}$$

La última ecuación asigna un valor de frecuencia a cada autovalor obtenido mediante el análisis de componentes principales.

2.4. Cálculo de la Entropía en la Aproximación Quasi-Armónica

La función de partición [41] para un oscilador armónico cuántico, incluyendo la energía del punto zero, viene dada por

$$Z = \sum_{j=0}^{\infty} e^{-\frac{\hbar\omega}{k_B T}\left(j+\frac{1}{2}\right)} = e^{-\frac{\hbar\omega}{2k_B T}} \sum_{j=0}^{\infty} \left(e^{-\frac{\hbar\omega}{k_B T}}\right)^j \tag{2.27}$$

A partir de esta función de partición se pueden obtener las funciones termodinámicas, siendo de interés la entropía de cada oscilador

$$s_{Q,i} = \frac{F-U}{T} = \frac{1}{T}\left(\frac{\hbar\omega}{e^{\frac{\hbar\omega}{k_B T}}-1}\right) - k_B \ln\left(1-e^{-\frac{\hbar\omega}{k_B T}}\right) \tag{2.28}$$

Para estimar la entropía total se considerará al sistema como conformado por una serie de osciladores armónicos cuánticos no-correlacionados [35, 37]

$$S_Q = k_B \sum_{i=1}^{3N} \left[\frac{1}{k_B T}\left(\frac{\hbar\omega_i}{e^{\frac{\hbar\omega_i}{k_B T}}-1}\right) - \ln\left(1-e^{-\frac{\hbar\omega_i}{k_B T}}\right)\right] \tag{2.29}$$

Este resultado, permite el cálculo de la entropía a partir de los valores de frecuencia, los cuales se pueden obtener a partir de la aplicación de la aproximación quasi-armónica a las trayectorias de dinámica molecular, y por tanto, este resultado se utilizará para la estimación de la entropía en el presente trabajo.

Capítulo 3

Metodología

3.1. Análisis de Grupos

El Análisis de Grupos, es una técnica del análisis estadístico, la cual intenta agrupar un conjunto de objetos, de forma que los objetos clasificados dentro de un grupo sean mas similares entre sí, que cualquier otro objeto del resto de grupos. Debe observarse que no existe una definición única de grupo para un conjunto de datos, ya que pueden utilizarse diferentes criterios para la clasificación dentro de los grupos; de manera similar, no existe un único algoritmo para realizar el análisis de grupos. Uno de los criterios más útiles para la clasificación de los grupos, consiste en el criterio de conectividad, el cual se basa en la idea de que los objetos mas *cercanos* estarán mas relacionados entre sí que los distantes[28]; los métodos de agrupamiento basados en este criterio se conocen como métodos de agrupamiento jerárquicos.

Para la aplicación de los métodos de agrupamiento jerárquicos, es fundamental la idea de distancia entre los objetos del conjunto, lo cual nos lleva al concepto de *métrica* [28, 27]; la métrica es una función $d(x, y)$ que define la distancia entre cada par de elementos de un conjunto, dicha función debe cumplir las siguientes propiedades: Debe ser no-negativa $d(x, y) \geq 0$, la distancia entre objetos indiscernibles entre sí debe ser cero $d(x, y) = 0\, si\, y\, solo\, si\, x = y$, debe cumplir la propiedad de simetría $d(x, y) = d(y, x)$, debe cumplir la desigualdad del triángulo $d(x, z) \leq d(x, y) + d(y, z)$. En la siguiente tabla se exponen algunos ejemplos de métrica utilizadas en los métodos de agrupamiento jerárquicos.

Función	Expresión Matemática
Distancia Euclídea	$d(a,b) = \sqrt{\sum_i (a_i - b_i)^2}$
Distacia Manhattan	$d(a,b) = \sum_i \lvert a_i - b_i \rvert$
Distancia Máxima	$d(a,b) = max_i \lvert a_i - b_i \rvert$

TABLA 3.1: Métricas utilizadas en los métodos de agrupamiento jerárquicos.

Una forma útil para representar los resultados de los métodos de agrupamiento jerárquicos es la representación de dendograma, el cual es un diagrama de árbol, cuya base tiene todos los elementos individuales, es decir cada elemento representa un grupo, conforme se asciende en el diagrama el número de ramas del árbol representa un grupo que contiene a los elementos de las filas inferiores, hasta la parte superior del diagrama que consiste de un solo grupo que contiene a todos los elementos.

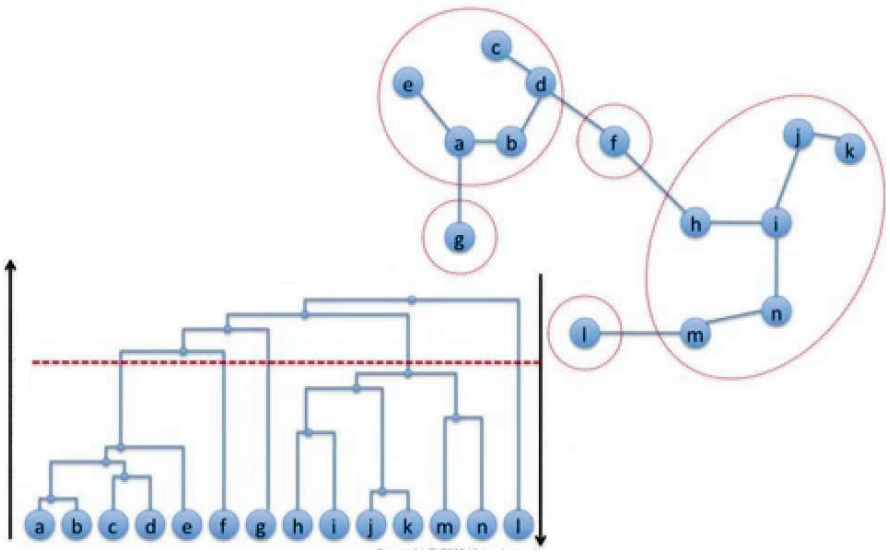

FIGURA 3.1: Esquema del análisis de grupos, y la representación de dendograma [42]

3.1.1. Algoritmo

El análisis de grupos jerárquico aglomerativo organiza los datos en forma de árboles. Cada hoja corresponde a uno de los objetos originales, y cada nodo interior representa una agrupación de objetos. Los algoritmos de agrupamiento jerárquico aglomerativo, construyen los árboles mediante una aproximación desde abajo hacia arriba, produciendo una serie de particiones de los datos, el primero consiste de n grupos con un único elemento, y el último consistente de un

único grupo con n elementos. En cada iteración, se unen elementos o conjuntos de elementos cercanos o similares, difiriendo los métodos de agrupamiento en la métrica utilizada, sin embargo el procedimiento es común a los diferentes métodos: el algoritmo comienza con n grupos formados por elementos de la forma x_i, y entonces se unen los dos grupos mas cercanos en cada iteración [43]. Esta unión se repite hasta que queda un solo cluster que contiene a todos los objetos. Debido a que en cada iteracion un par grupos se unen en uno solo, el algoritmo termina después de $n - 1$ iteraciones.

Sean n objetos $x_1, \ldots, x_n$. El grupo $C_0 = e_1, \ldots, e_n$ y las distancias $d(x_i, x_j) = d(x_i, x_j)$ para todo $i \neq j$.

Para $l = 1, \ldots, n - 1$:

1. Evaluar $(G_1, G_2) = min_{H,K \in C_{l-1}} d(H, K)$.
2. Actualizar $C_l = C_{l-1} \cup \{G_1 \cup G2\} \setminus \{G_1, G_2\}$.
3. Calcular $d(G_1 \cup G2, H)$ para todo $H \in C_l$.

Donde C_l denota el agrupamiento después de l pasos.

3.2. Trayectorias del Lazo 36 de la Hemaglutinina

3.2.1. Trayectoria de Dinámica Molecular Dirigida (TMD)

Las simulaciones de dinámica molecular dirigida (TMD: Targeted Molecular Dynamics) permiten obtener un mapeo rápido de las conformaciones de la biomolécula durante la transición entre dos estados: uno inicial y uno objetivo. De tal forma que se pueden establecer posibles conformaciones intermedias, que podrían ayudar a estudiar la dinámica de la transición, ya que podrían ser posibles estados que caracterizan a la transición, y son de interés teórico y experimental [4].

Con el fin de explorar las posibles conformaciones de la biomolécula durante la transición conformacional, Calderon J. en su trabajo [26] realizó una simulación TMD, tomando como estado inicial el correspondiente a la conformación no-fusogénica, y como estado objetivo el correspondiente a la estructura fusogénica. Como resultado de esta simulación se obtuvo una *trayectoria TMD*, consistente de 10000 estructuras, explorando uno de los posibles caminos para la transición.

Esta trayectoria TMD fue sometida por Calderon J. a un análisis de grupos aglomerativo, utilizando como parámetro la distancia euclídea entre estructuras, con una distancia de corte

de $57,1[angstrom]$. Como resultado de este análisis se encontraron 11 grupos, los cuales fueron asociados a conformaciones intermedias de la transición, tomándose la estructura representativa de cada grupo como estructura representativa de la conformación respectiva [26]. Las conformaciones intermedias identificadas se muestran en la siguiente figura.

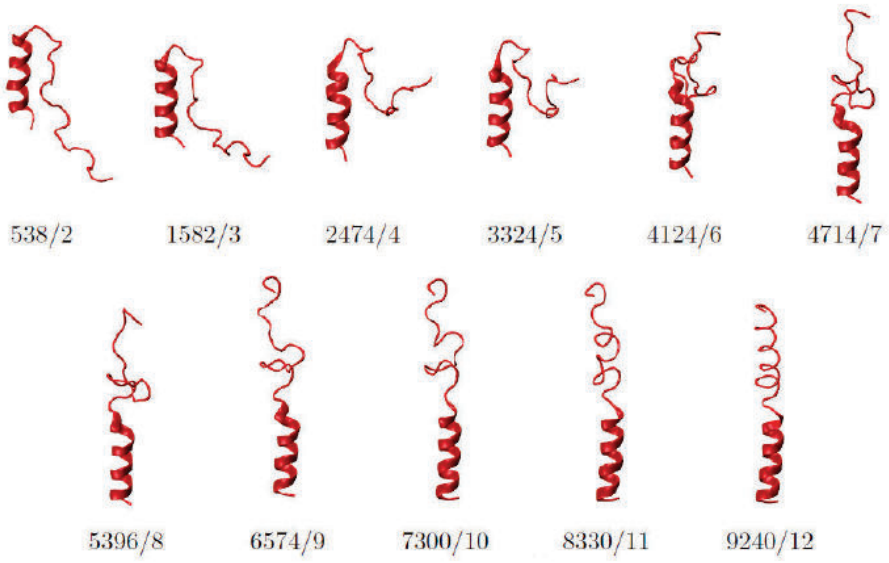

FIGURA 3.2: Las 11 estructuras representativas intermedias identificadas en [26], imagen reproducida de [26].

3.2.2. Trayectorias de Dinámica Molecular de Equilibración (TEQ)

Con el fin de explorar los estados accesibles a cada conformación, Calderon J. realizó simulaciones de dinámica molecular de equilibración, para dicha simulación se partió de las estructuras representativas de cada conformación (fig. 3.2) además de la conformación fusogénica y la no fusogénica, dejándose evolucionar cada una de ellas mediante dinámica molecular durante un tiempo total de $6ns$ con condiciones de frontera esféricas. De esta forma se obtuvieron 13 trayectorias, cada una de las cuales tiene 12000 estructuras, correspondiendo cada estructura a 20 ps de tiempo en la equilibración.

Para este trabajo se utilizará una trayectoria de dinámica molecular dirigida (TMD), y 13 trayectorias de equilibración. La trayectoria TMD se utilizó para validar el análisis de grupos, y por ende las conformaciones identificadas, en tanto que las trayectorias de equilibración se utilizaron para la estimación de la entropía y la energía libre.

3.3. Procedimiento

3.3.1. Recursos Computacionales

Los procedimientos de cálculo y visualización científica, fueron realizados en un computador *MacBook Pro* modelo A1278, con procesador *Intel Core i7* de 2.9 GHz, cuyas características mas reelevantes se enumeran en la tabla 3.2.

Dispositivo	Modelo	Características
Procesador	Intel Core i7	2.9 GHz
Gráficos	Intel HD Graphics 4000	GPU 1536 MB
Memoria	Hynix	8 GB, 1600 MHz, DDR3
Disco Duro	Kingston	SSD, 240 GB

TABLA 3.2: Características del computador MacBook Pro, utilizado para el procesamiento de datos

Los paquetes computacionales utilizados se enumeran en la tabla 3.3.

Nombre	Desarrollador	Versión	Función
VMD [24]	UIUC	1.9.3	Software para Visualización Molecular
NAMD [44]	UIUC	2.11	Software para Dinámica Molecular
R [45]	R Foundation	3.3.2	Entorno para Computación Estadística
Python [46]	Python Software Foundation	2.7.13	Lenguaje de Programación
Matplotlib[47]	Matplotlib Development Team	2.0.2	Librería de visualización científica en Python
Numpy [48]	SciPy Developers	1.8.0	Librería de computación científica en Python
Gnuplot [49]	Williams, Kelley et al.	5.0.3	Librería de visualización científica
MDTraj [50]	Stanford University y Robert McGibbon	1.8.0	Librería para el análisis de trayectorias de dinámica molecular en Python
ProtoClust [43]	Bien, J. y Tibshirani, R.	1.5	Paquete para análisis jerárquico de grupos con prototipos en R

TABLA 3.3: Paquetes computacionales utilizados

3.3.2. Evaluación del Método de Agrupamiento Inicial

Para cada estructura clasificada dentro de una conformación intermedia (grupo) en [26], se obtuvo un mapa de Ramachandran. Las conformaciones definidas en [26] se consideran adecuadas si las estructuras asociadas cumplen con las siguientes condiciones:

- La distancia euclídea entre las estructuras dentro de la conformación en el plano de Ramachandran, tiene un valor medio menor que la distancia entre diferentes conformaciones.
- Los puntos en el mapa de Ramachandran de las estructuras de una conformación ocupan un área definida.

Para evaluar estas condiciones, se define la distancia euclídea entre un par de estructuras j, k en el plano de Ramachandran de la siguiente manera:

$$\Gamma_{jk} = \frac{1}{36}\sum_{i=1}^{36}\gamma^i \tag{3.1a}$$

$$\Gamma_{jk} = \frac{1}{36}\sum_{i=1}^{36}\sqrt{(\phi_j^i - \phi_k^i)^2 + (\psi_j^i - \psi_k^i)^2} \tag{3.1b}$$

donde ϕ y ψ son los ángulos de Ramachandran, el superíndice i indica el número de residuo, los subíndices j y k denotan un par de estructuras clasificadas dentro de una conformación.

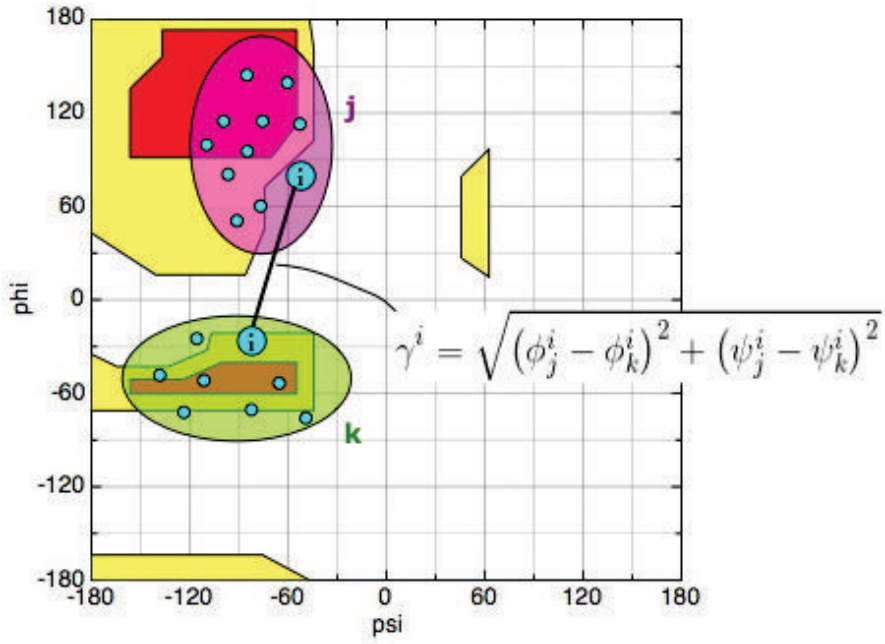

FIGURA 3.3: Distancia Euclídea en el plano de Ramachandran

El mapa de Ramachandran para cada estructura de las trayectorias de dinámica molecular orientada se obtuvo mediante el programa VMD [24], para lo cual se preparó el script correspondiente; posteriormente se determinó si los ángulos ϕ y ψ obtenidos para cada residuo se encuentran dentro de la región correspondiente al α-hélice, con esta información se calculó la helicidad para cada estructura.

3.3.3. Estimación de la Entropía

Las trayectorias disponibles, se encuentran en un formato binario (DCD), por lo cual se exportaron a coordenadas del tipo x, y, z, utilizando el software VMD [24], las coordenadas

exportadas se organizaron en una matriz, utilizando el lenguaje de programación Python [46], la estructura de la matriz es

$$\begin{bmatrix} \boldsymbol{r_{11}} & \boldsymbol{r_{12}} & \boldsymbol{r_{13}} & \cdots & \boldsymbol{r_{1n_a}} \\ \boldsymbol{r_{21}} & \boldsymbol{r_{22}} & \boldsymbol{r_{23}} & \cdots & \boldsymbol{r_{2n_a}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{r_{n_s1}} & \boldsymbol{r_{n_s2}} & \boldsymbol{r_{n_s3}} & \cdots & \boldsymbol{r_{n_sn_a}} \end{bmatrix} \tag{3.2}$$

donde $r_{ij} = (x_{ij}, y_{ij}, z_{ij})$, i indica el número de estructura y j el número de átomo, de forma que la matriz tiene dimensiones $3n_a \times n_s$.

Sobre la trayectoria de dinámica molecular de equilibración representada por esta matriz se aplicó el análisis de componentes principales (PCA), mediante un script en R [45]. A partir de los autovalores (desviaciones estándar) de las componentes principales obtenidas, se encontraron las frecuencias correspondientes a través de la Aproximación Quasi-Armónica, con estas frecuencias se estimó la entropía mediante la ecuación (sección 2.4):

$$S_Q = k_B \sum_{i=1}^{3N} \left[\frac{1}{k_B T} \left(\frac{\hbar\omega_i}{e^{\frac{\hbar\omega_i}{k_B T}} - 1} \right) - \ln\left(1 - e^{-\frac{\hbar\omega_i}{k_B T}}\right) \right] \tag{3.3}$$

3.3.4. Descripción de la Barrera Energética

Para describir la barrera energética, se buscará la coordenada generalizada, o coordenada de reacción, que mejor caracterice el comportamiento energético durante la transición, a través una gráfica de la energía interna en función de la coordenada de reacción.

Como candidato a coordenada de reacción se utilizó la primera coordenada principal resultante de la aplicación del PCA a la trayectoria TMD, ya que esta coordenada por definición contiene la mayor cantidad de varianza de los datos.

Otra posible coordenada generalizada que se utilizó es el RMSD entre las posiciones de los átomos en las estructuras calculado con respecto a la estructura inicial (fusogénica), de manera similar el RMSD calculado con respecto a la estructura final (no-fusogénica). Estos valores de RMSD se calcularon mediante la herramienta *RMSD Calculator* disponible en el software VMD [24].

De manera similar, se utilizó el *RMSD angulo*, calculado en el espacio de Ramachandran con respecto tanto a la estructura fusogénica como a la no-fusogénica.

$$RMSD_{a0} = \sqrt{\frac{1}{36}\sum_{i=1}^{36}(\phi^i_{conf} - \phi^i_0)^2 + (\psi^i_{conf} - \psi^i_0)^2} \tag{3.4}$$

$$RMSD_{af} = \sqrt{\frac{1}{36}\sum_{i=1}^{36}(\phi^i_{conf} - \phi^i_f)^2 + (\psi^i_{conf} - \psi^i_f)^2} \tag{3.5}$$

donde $\{\ \phi^i_{conf}, \psi^i_{conf}\}$ son los ángulos de Ramachandran para la estructura representativa de la conformación intermedia en la trayectoria TMD, $\{\phi^i_0, \psi^i_0\}$ son los ángulos de Ramachandran para la estructura de la conformación no-fusogénica, y $\{\phi^i_f, \psi^i_f\}$ son los ángulos de Ramachandran para la estructura de la conformación fusogénica.

Otra coordenada que se utilizó fue la *helicidad* la cual se definió como la cantidad de aminoácidos presente en un α-hélice en relación al número total de aminoácidos.

$$h_\alpha = \frac{N\alpha_R}{N_T} \tag{3.6}$$

donde $N\alpha_R$ es el número de residuos que se encuentran en el α-hélice, y N_T es el número total de residuos. El valor de $N\alpha_R$ se obtuvo a través del conteo de puntos que se encuentran en la región del espacio de Ramachandran descrita en la tabla 1.1.

Capítulo 4

Resultados y Discusión

4.1. Análisis del Agrupamiento de Estructuras

4.1.1. Agrupamiento Inicial de Estructuras TMD

El análisis de grupos original de las estructuras en la trayectoria TMD, se realizó utilizando la distancia media cuadrática (RMSD) como parámetro, con una distancia de corte de *57.1 angstroms*, y considerando únicamente 5000 estructuras, las cuales representan la mitad de todas las estructuras disponibles [26]. En primer lugar se realizó un análisis de grupos con todas las 10000 estructuras TMD disponibles y los mismos parámetros utilizados en [26]. Como resultado se encontraron 11 grupos, al igual que en [26], pero con diferente tamaño, y diferente estructura representativa.

La tabla 4.1 muestra las diferencias entre los dos análisis. Para cada configuración se muestran las características de los grupos encontrados considerando 5000 y 10000 estructuras. Se puede observar que los grupos con mayor disimilitud son el $7, 8, 9$ y 11. Por otra parte, las estructuras prototipo de los nuevos grupos encontrados, se encuentran dentro del rango de estructuras de los grupos anteriores, excepto para los grupos 8B y el 10B.

Conf.	Grupo	Estructura Inicial	Estructura Final	Rango de Estructuras	Prototipo			
					Prototipo	Helicidad (%)	$RMSD_0$	$RMSD_1$
1	1A	0	870	870	538	51.51 %	0.1942	0.0219
	1B	0	1419	1419	715	48.48 %	0.2161	
2	2A	870	2162	1292	1582	51.51 %	0.3511	0.0514
	2B	1420	2168	748	1778	48.48 %	0.4025	
3	3A	2162	2814	652	2474	48.48 %	0.5250	0.0003
	3B	2168	2803	635	2476	48.48 %	0.5253	
4	4A	2814	3700	886	3324	45.45 %	0.6520	0.0086
	4B	2804	3610	806	3301	42.42 %	0.6434	
5	5A	3700	4496	796	4124	48.48 %	0.8502	0.0288
	5B	3611	4442	831	4027	48.48 %	0.8214	
6	6A	4496	4948	452	4714	45.45 %	0.9514	0.0175
	6B	4443	5209	766	4735	51.51 %	0.9689	
7	7A	4948	5898	950	5396	45.45 %	1.0275	0.0231
	7B	5210	6632	1422	5780	48.48 %	1.0506	
8	8A	5898	6880	982	6574	54.74 %	1.1221	0.0541
	8B	6633	7284	651	6933	51.51 %	1.1762	
9	9A	6880	8120	1240	7300	42.42 %	1.2009	0.0208
	9B	7285	8219	934	7961	48.48 %	1.2217	
10	10A	8120	8472	352	8330	48.48 %	1.2187	0.0072
	10B	8220	9115	895	8474	57.57 %	1.2115	
11	11A	8472	10000	1528	9240	66.67 %	1.2188	0.0130
	11B	9116	10000	884	9416	90.90 %	1.2318	

TABLA 4.1: Comparación de las estructuras representativas obtenidas para cada conformación intermedia mediante análisis de grupos, el $RMSD_0$ se ha calculado con respecto a la primera estructura de la trayectoria TMD, el $RMSD_1$ cuantifica la diferencia de $RMSD$ entre las estructuras prototipo para cada grupo dentro de una conformación

La figura 4.1, muestra el tamaño de los grupos, y el rango de estructuras que los componen.

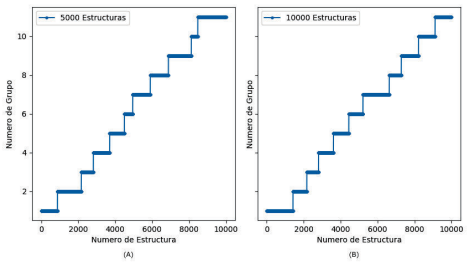

FIGURA 4.1: Distribución de las estructuras en grupos como resultado del análisis hecho con (A) 5000 [26] y (B) 10000 estructuras (este trabajo)

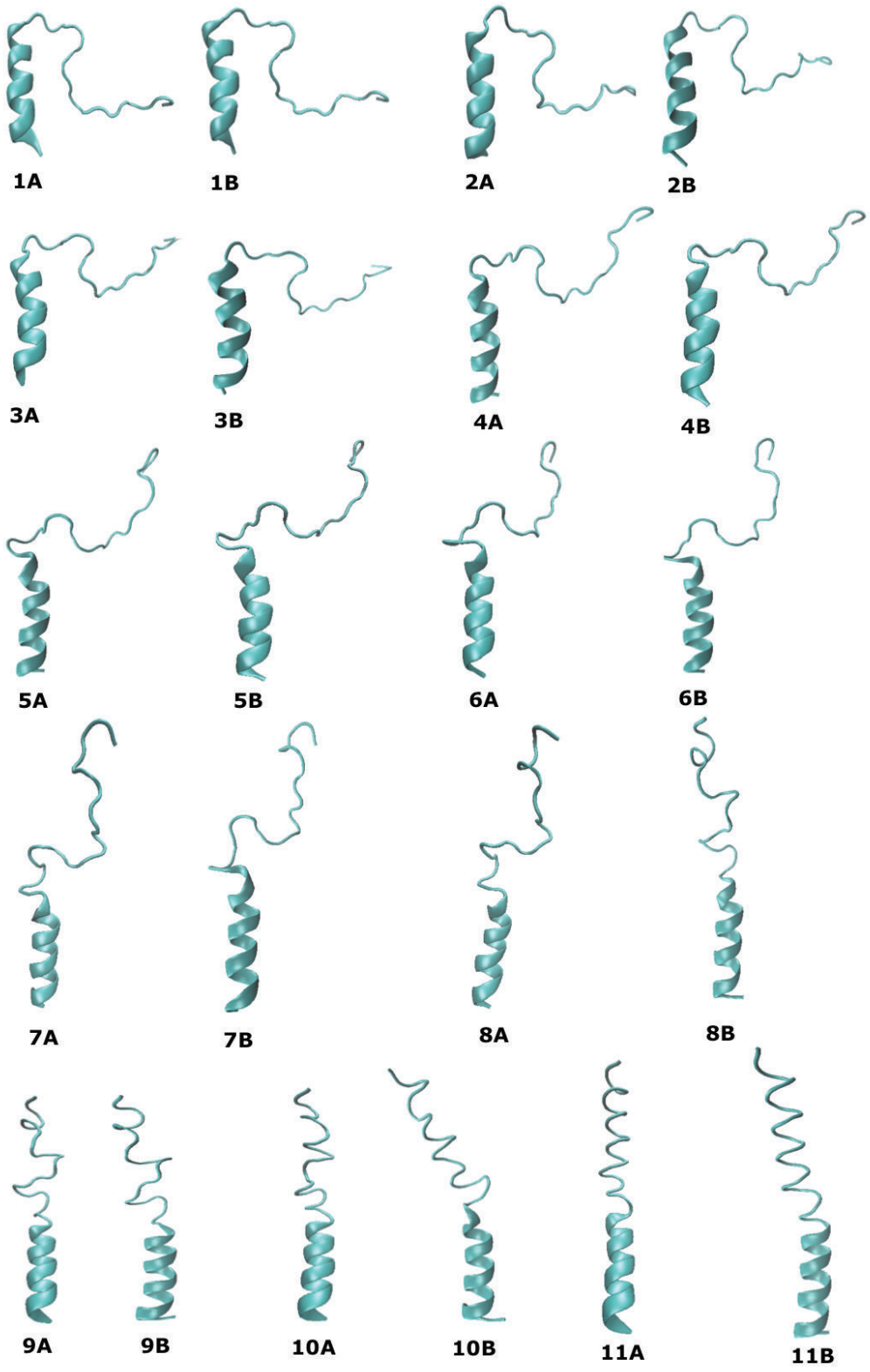

FIGURA 4.2: Estructura representativa para cada grupo

4.1.2. Mapa de Ramachandran para la trayectoria de Dinámica Molecular Dirigida

Se generó un histograma para los ángulos de las 10000 estructuras TMD en el mapa de Ramachandran, con el fin de comprobar las regiones de mayor ocupación, el cual se muestra en la figura 4.3.

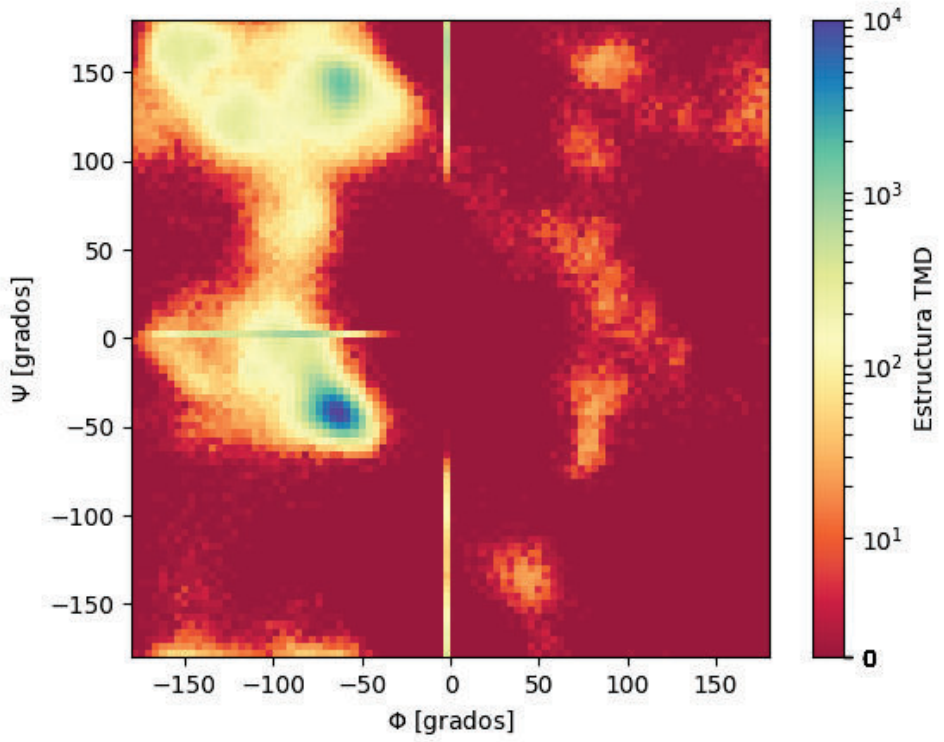

FIGURA 4.3: Histograma del Mapa de Ramachandran para las estructuras TMD

Se puede observar que la región de mayor ocupación, en el plano de Ramachandran, es la región correspondiente a estructuras secundarias de tipo α-hélice. La región correspondiente a las estructuras tipo hojas-β también es ocupada con mayor frecuencia que el resto de regiones del plano.

Mapa de Ramachandran para cada Residuo

Se generó un mapa de Ramachandran para cada residuo, durante la trayectoria TMD, con el fin de observar su movimiento en el espacio de Ramachandran, cuyos resultados se muestran en la figura 4.4.

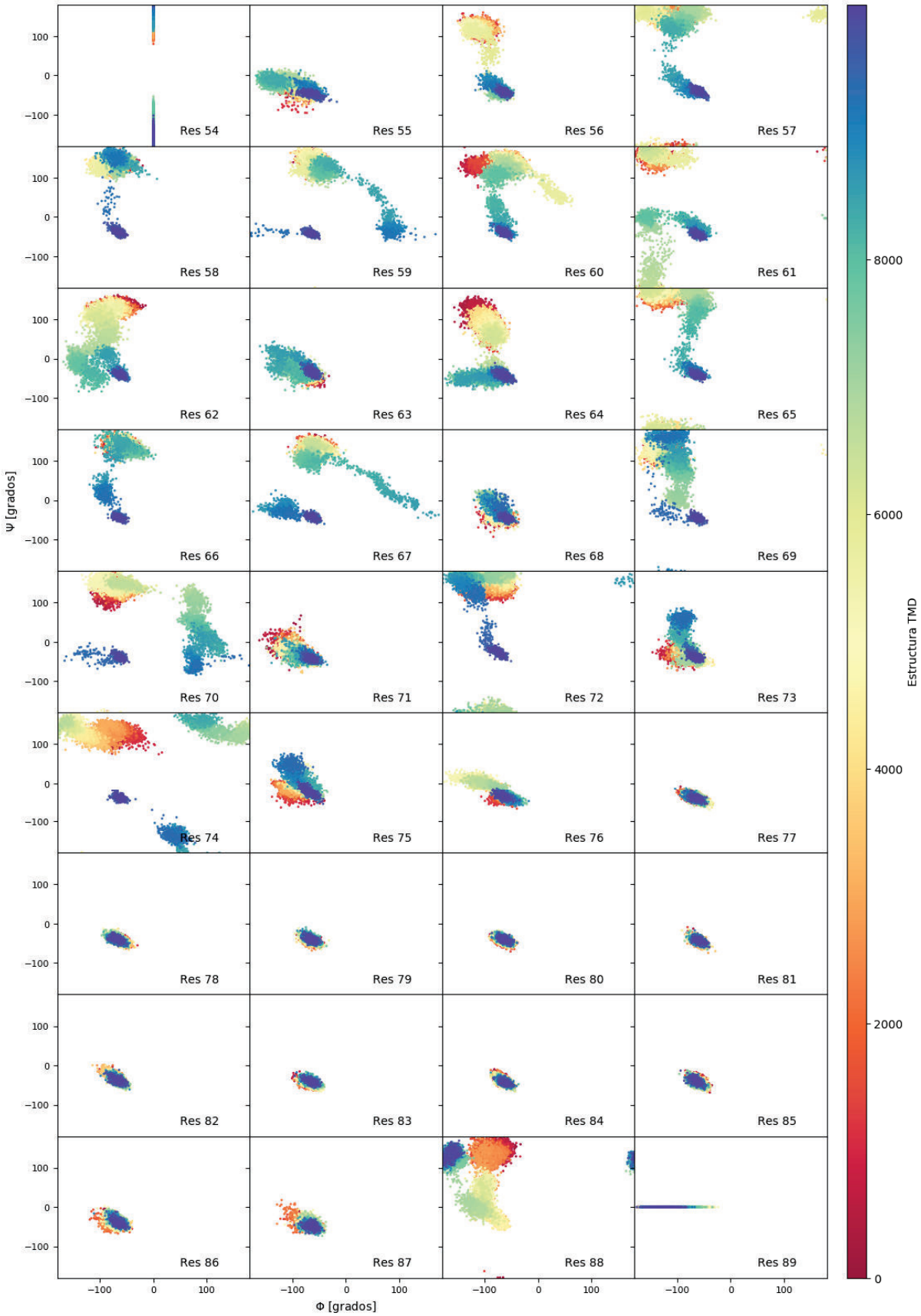

FIGURA 4.4: Mapa de Ramachandran para cada residuo

Debido a que los ángulos están definidos en torno al enlace formado por dos residuos, no a uno solo, los mapas correspondientes a los residuos de los extremos se ubican sobre los ejes. El residuo 54 corresponde al terminal amino del polipéptido, y su mapa de Ramachandran tiene valores de $\phi = 0$, es decir que se encuentra sobre el eje vertical. De manera similar, el residuo 89 corresponde al terminal carboxilo, y su angulo de torsión $\psi = 0$, por lo que se ubica sobre el eje horizontal del mapa de Ramachandran.

Como se conoce a partir de la información de la estructura del lazo 36 [12, 14], los residuos desde el 76 hasta el 89 forman una estructura de tipo α-hélice, como se puede comprobar también en la figura 4.4, excepto para el residuo 88. Por ello el mapa de Ramachandran del residuo 88 es anómalo, y se puede considerar como un artefacto de la simulación causado por las restricciones impuestas.

Considerando estos hechos, en todos los análisis posteriores, no se tomaran en cuenta los datos de los ángulos de torsión de los residuos 54, 88 y 89. De este modo se eliminaran datos que no aportan información relevante.

Mapas de Ramachandran para cada grupo

Mapa de Ramachandran, en el cual se representaron los ángulos de torsión ϕ y ψ para todas las estructuras clasificadas dentro de un grupo. Este análisis permite apreciar de forma gráfica la dispersión de las estructuras dentro de un grupo en el plano de Ramachandran.

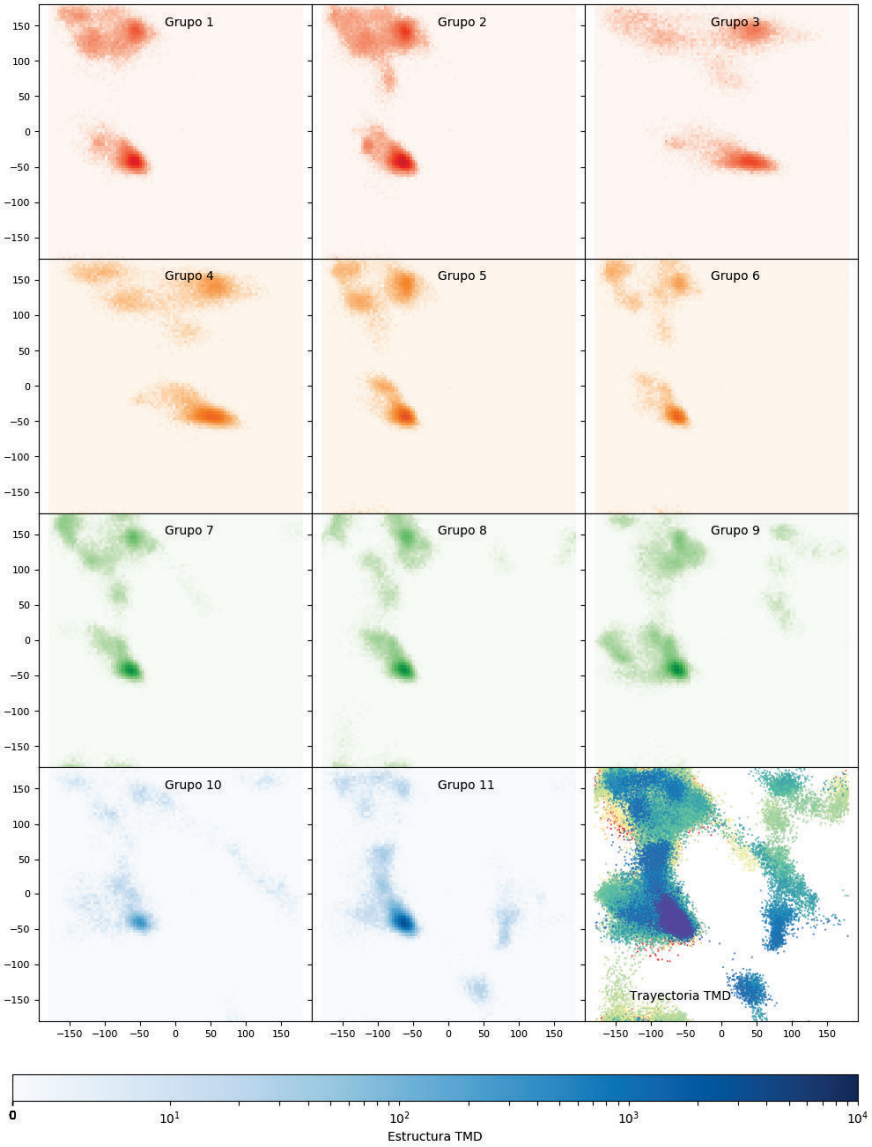

FIGURA 4.5: Mapas de Ramachandran para cada grupo

Se puede observar que, inicialmente los ángulos de torsión de los residuos están dispersos en el mapa de Ramachandran, particularmente en los cuadrantes 3 y 4 del mapa, visitando las regiones usualmente ocupadas por las estructuras tipo hojas-β.

Los mapas de Ramachandran de los grupos desde el 1 hasta el 6, presentan dos aglomerados, uno en torno a la región correspondiente a α-hélices, y otra en la vecindad de la región correspondiente a hojas-β. El grupo 7 presenta 3 aglomerados, en tanto que el grupo 8 se encuentra fragmentado en 7 aglomerados. El grupo 9 presenta esencialmente 3 aglomerados, en tanto que el grupo 10 presenta 4 aglomerados, los cuales están muy dispersos, y a la vez próximos. Finalmente el grupo 11 presenta 4 aglomerados, los cuales están bien definidos.

4.1.3. Distancias en el Plano de Ramachandran

Se realizó un histograma de la distancia euclídea de los ángulos de torsión entre las estructuras de cada grupo, para explorar la consistencia de los grupos definidos en [26]. La estructura que se tomó como referencia fue la estructura inicial, correspondiente a la conformación no-fusogénica.

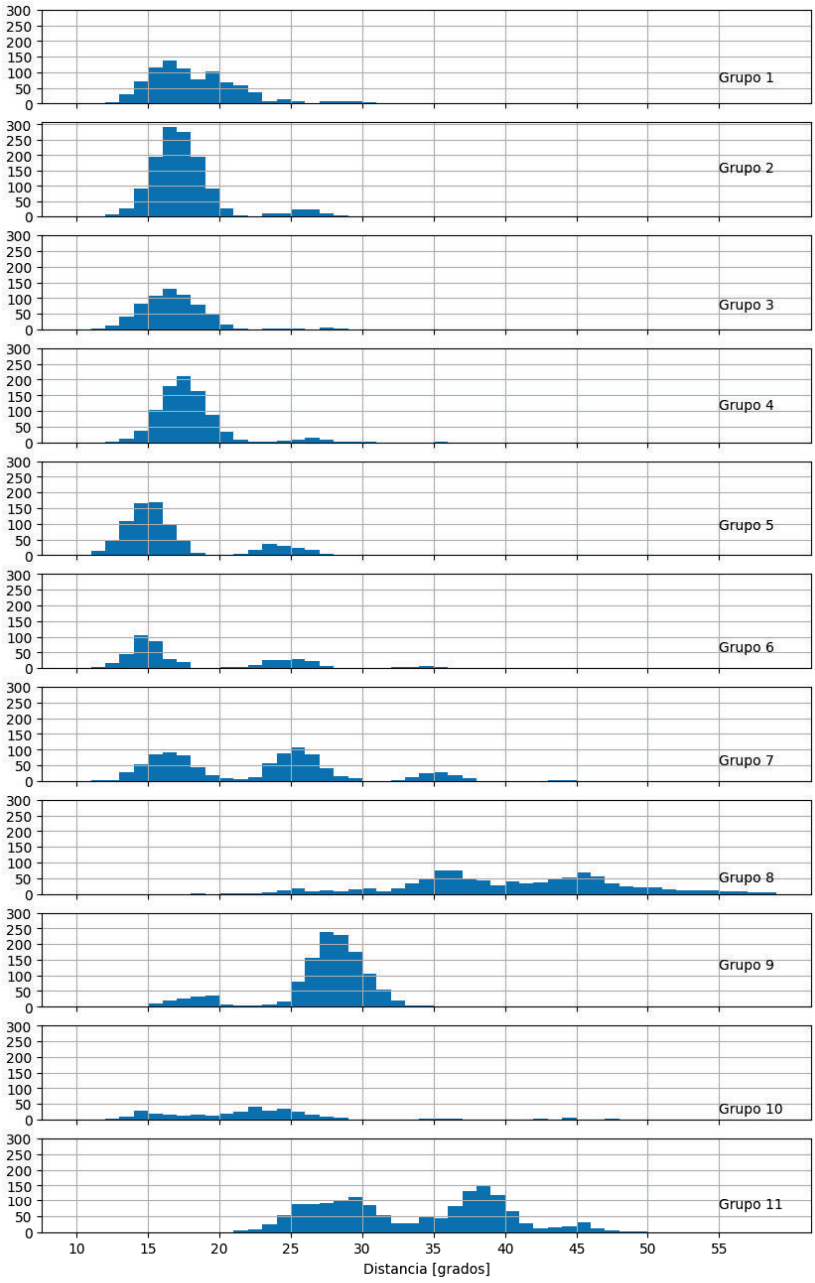

FIGURA 4.6: Histogramas para las distancias entre las estructuras de cada grupo en la Trayectoria TMD, se señalan con numerales la ubicación de los máximos

Un máximo en el histograma representa un conjunto de estructuras, cuya distancia euclídea en el espacio de Ramachandran calculada con respecto a la estructura fusogénica, es de magnitud similar; lo cual a su vez indica similitud entre dichas estructuras, y por tanto podrían considerarse como posibles subgrupos. Los valores máximos determinados en el histograma, y la distancia entre ellos se presentan en la siguiente tabla.

Grupo	No. de Máximos	Máximo	Distancia entre máximos
1	2	I = 16 , II = 19	$d_{I-II} = 3$
2	1	I = 16	N/A
3	1	I = 16	N/A
4	1	I = 17	N/A
5	2	I = 15 , II = 23	$d_{I-II} = 8$
6	2	I = 14 , II = 25	$d_{I-II} = 11$
7	3	I = 16 , II = 25, III = 35	$d_{I-II} = 9$, $d_{II-III} = 10$
8	3	I = 30 , II = 35 , III = 40	$d_{I-II} = 5$, $d_{II-III} = 5$
9	2	I = 19 , II = 27	$d_{I-II} = 8$
10	4	I = 14 , II = 18 , III = 22 , IV = 24	$d_{I-II} = 4$, $d_{II-III} = 4$, $d_{II-III} = 2$
11	4	I = 25 , II = 29 , III = 38 , IV = 45	$d_{I-II} = 4$, $d_{II-III} = 9$, $d_{II-III} = 7$

TABLA 4.2: Máximos en el histograma de cada grupo, y distancia entre ellos.

A partir de la tabla anterior, se puede observar que los grupos 7 y 8, tienen 3 máximos definidos. En tanto que los grupos 9 y 10 tienen 4 máximos definidos. Esto sugiere la existencia de un número mayor de conformaciones, las cuales se manifiestan a través de las estructuras acumuladas en el valor máximo, y en su entorno.

De manera similar, se realizaron histogramas acumulativos para las distancias entre las estructuras en el espacio de Ramachandran.

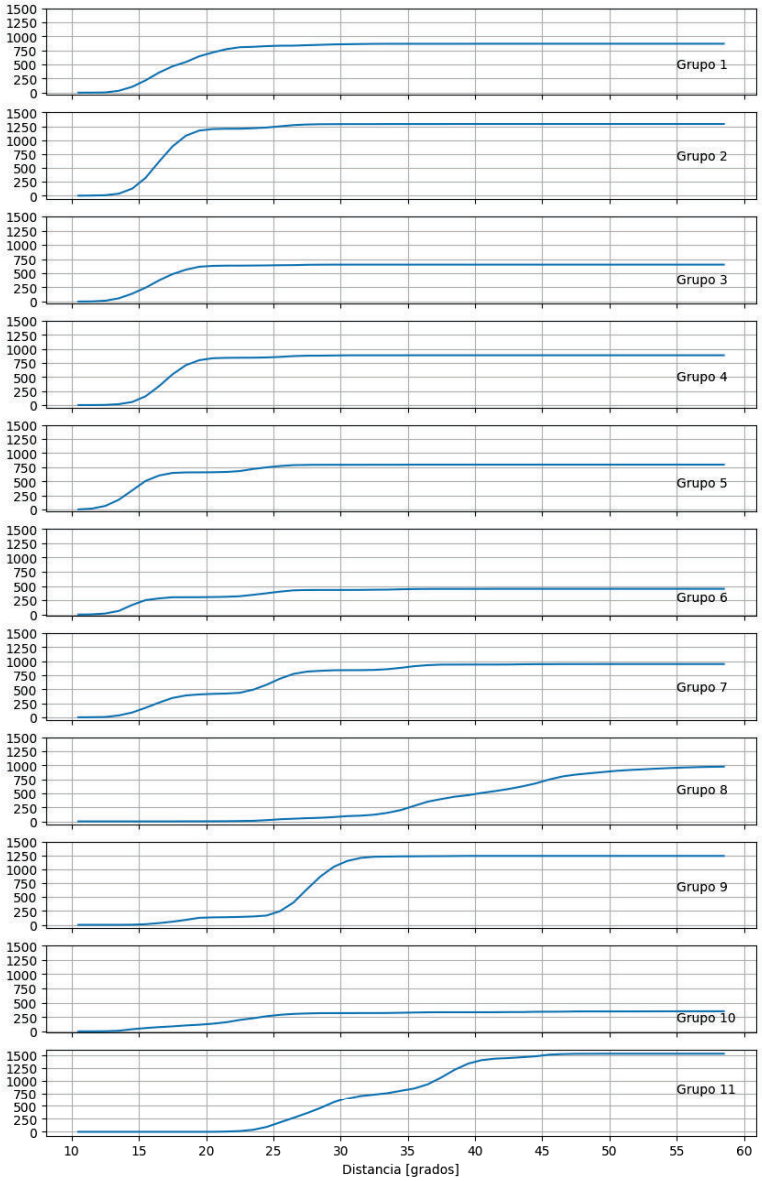

FIGURA 4.7: Histogramas acumulativos para las distancias entre las estructuras de cada grupo en la Trayectoria TMD

4.1.4. Análisis de Grupos en el Espacio de Ramachandran

Para explorar la posibilidad de la existencia de grupos adicionales, se realizó un nuevo análisis de grupos sobre cada conjunto de estructuras de la trayectoria TMD clasificadas dentro de un grupo de acuerdo al criterio del RMSD, utilizando como parámetro la distancia euclídea de los ángulos de torsión entre las estructuras del grupo. De forma que se obtuvieron 11 dendogramas, correspondientes a cada grupo, los cuales se muestran en la siguiente figura.

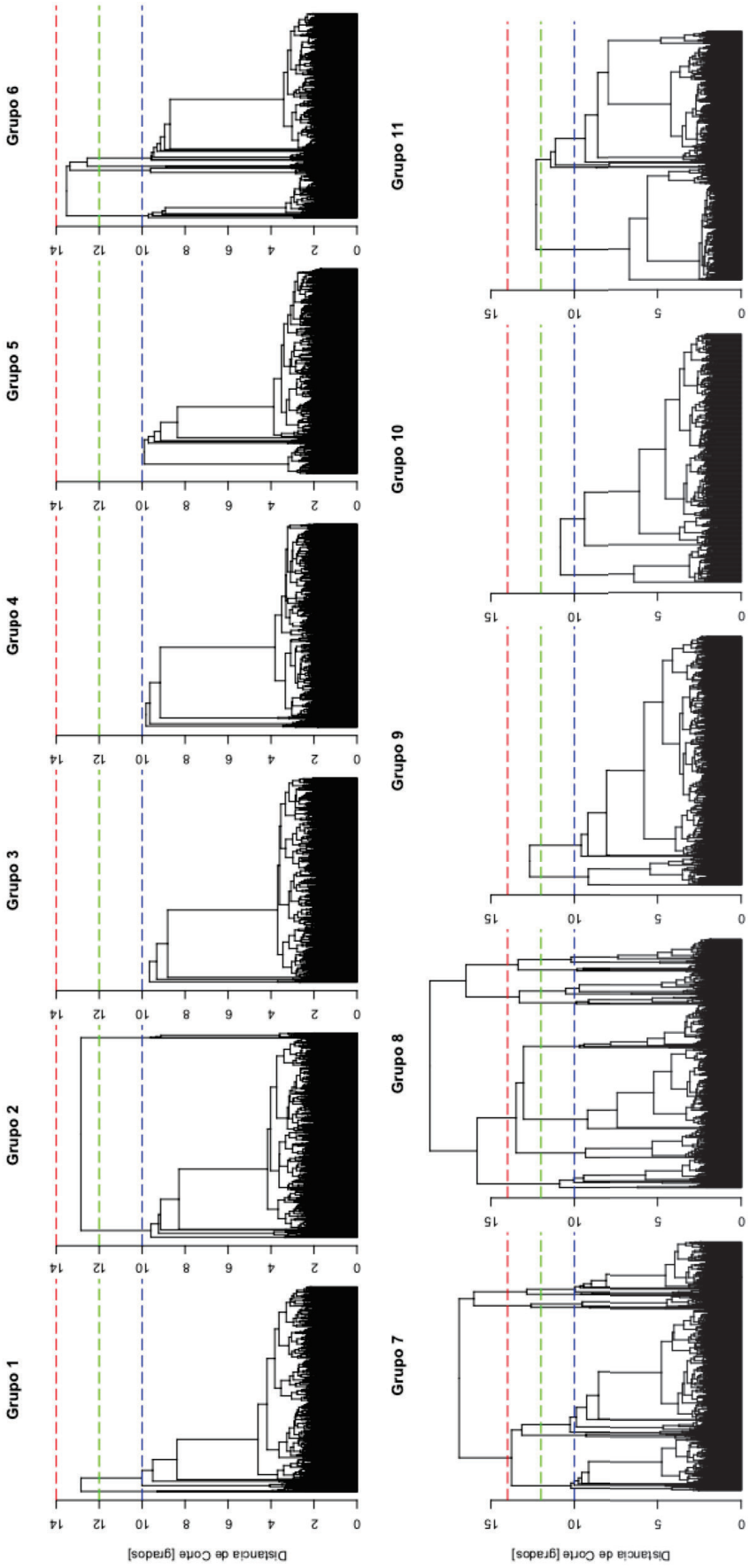

FIGURA 4.8: Dendogramas para cada grupo

A distancias de corte representativas, tales como 10, 12 y 14 grados, se obtiene un número de grupos comparable para cada dendograma. Los resultados de este análisis se reportan en la siguiente tabla.

Grupo	Distancia de Corte		
	14	12	10
	Número de Grupos		
1	1	2	2
2	1	2	2
3	1	1	1
4	1	1	1
5	1	1	1
6	1	5	5
7	3	7	9
8	4	4	11
9	1	2	2
10	1	1	2
11	1	2	5

TABLA 4.3: Numero de grupos para distancias de corte específicas

Se puede comprobar que los grupos $3, 4$ y 5 no tienen subgrupos para las distancias de corte consideradas, por ello se puede afirmar que están bien definidos de acuerdo a los dos criterios utilizados (RMSD y distancia en el espacio de Ramachandran). Por otra parte los grupos 7 y 8 son los que mas subgrupos presentan para las distancias de corte consideradas.

4.2. Entropía Conformacional

4.2.1. Frecuencias de los osciladores armónicos

En la figura 4.9 , se muestran las frecuencias de los osciladores armónicos en la aproximación Quasi-Armónica, correspondientes a las componentes principales, para cada una de las conformaciones intermedias.

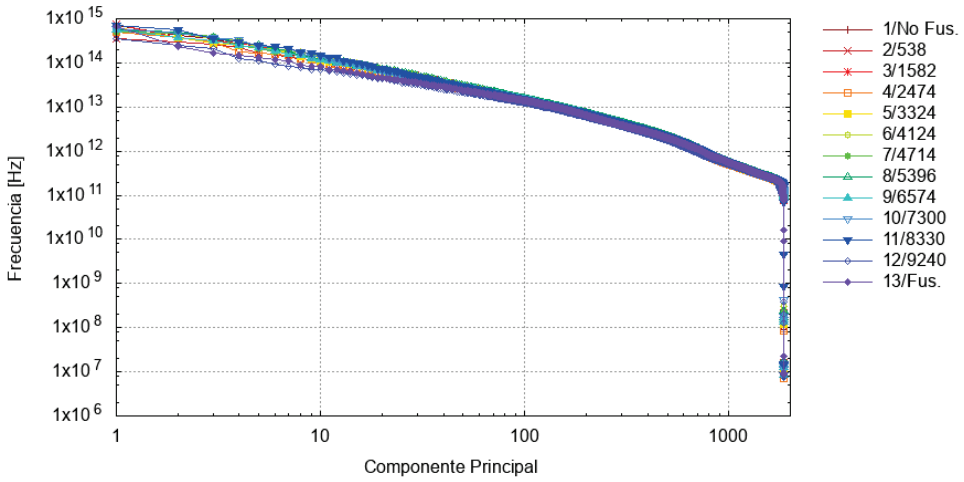

FIGURA 4.9: Frecuencias de los osciladores armónicos de la aproximación quasi-armónica, correspondientes a las componentes principales

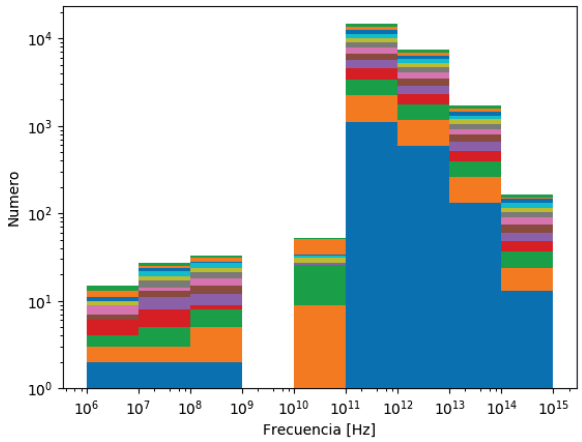

FIGURA 4.10: Histograma de las frecuencias de los osciladores armónicos correspondientes a las componentes principales

A partir de los histogramas, se puede observar que las frecuencias de los osciladores armónicos, se encuentran mayoritariamente en el rango entre $10^{11}[Hz]$ y $10^{14}[Hz]$. Considerando que la frecuencia de oscilación de las vibraciones moleculares se encuentran en el orden de entre $10^{13}[Hz]$ y $10^{14}[Hz]$ [51, 52], se puede observar que las frecuencias obtenidas son cercanas a este rango.

Por otra parte, en los histogramas se puede apreciar la presencia de frecuencias en un rango bajo, las cuales corresponden a las últimas frecuencias obtenidas mediante la aproximación quasi-armónica, como se puede observar en la figura 4.11.

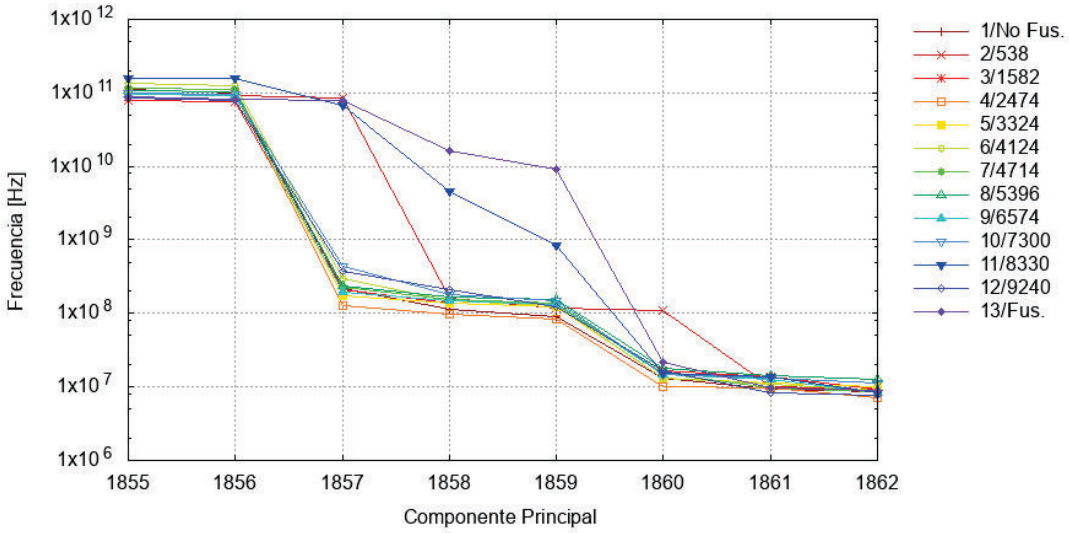

FIGURA 4.11: Frecuencias de los últimos osciladores armónicos, correspondientes a las últimas componentes principales, los 6 últimos valores presentan una variación significativa con respecto a los anteriores, para todos los grupos

Para explicar la presencia de estas frecuencias, en primera instancia se analizó la posibilidad de existencia de oscilaciones del centro de masa de la molécula, para lo cual se calcularon las posiciones del centro de masa a lo largo de la trayectoria TMD, las cuales se muestran figura 4.12.

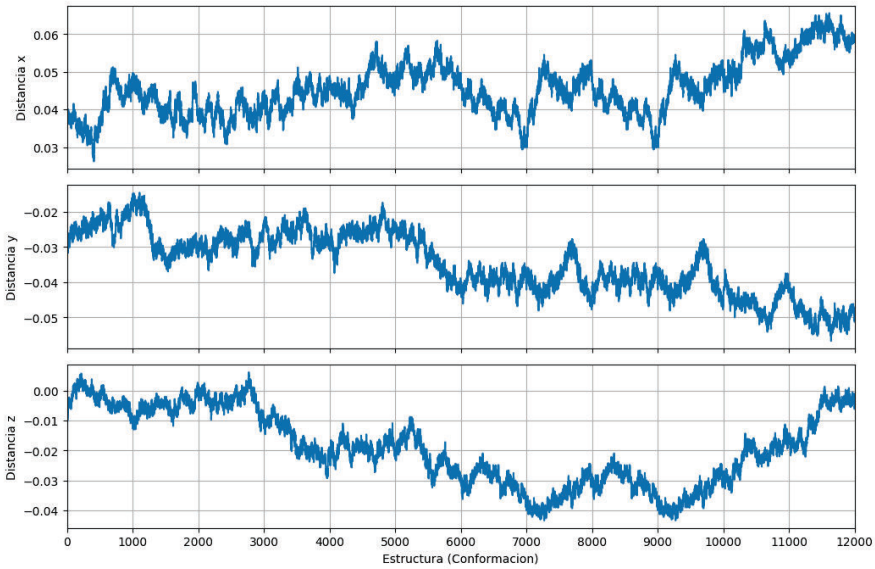

FIGURA 4.12: Posición del centro de masa para la trayectoria de equilibración de la conformación no-Fusogénica

Para caracterizar estas oscilaciones del centro de Masa, se procedio a obtener la transformada

de Fourier mediante el método FFT, considerando ademas que dado que se tiene en total 12000 estructuras correspondientes a $6[ns]$ de tiempo de simulación, el tiempo de muestreo sería de $500[ps]$, lo cual nos permite obtener el rango de frecuencias para la FFT. Los resultados de este análisis se muestran en la figura 4.13.

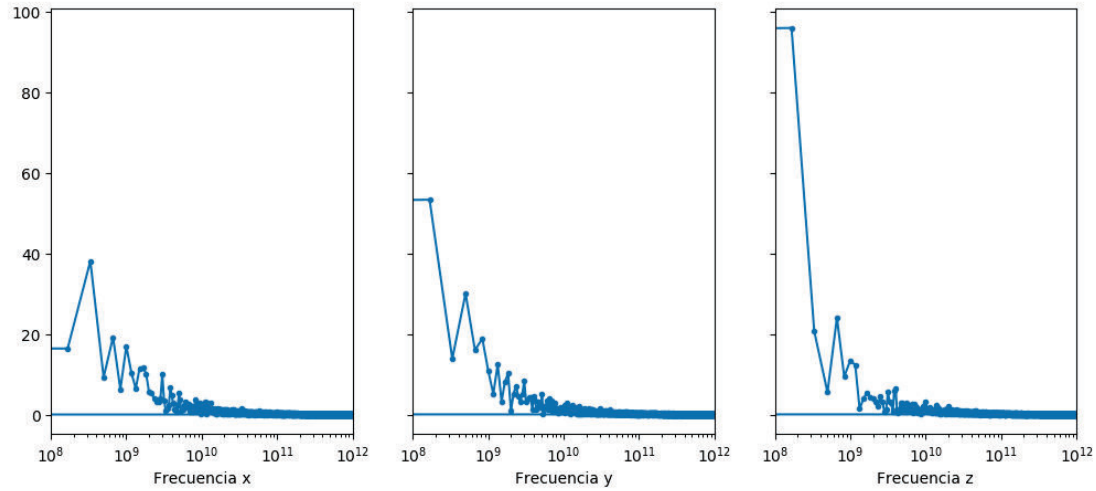

FIGURA 4.13: Transformada de Fourier de las coordenadas del centro de masa para la trayectoria de equilibración de la conformación no-Fusogénica

A partir de la figura anterior, se puede observar que la mayor componente de frecuencia se encuentre en el rango entre $10^8[Hz]$ y $10^9[Hz]$, para cada coordenada analizada, lo cual permite explicar 3 de los 6 valores de la figura 4.11.

4.2.2. Entropía Absoluta

Para estimar la entropía se utilizó la ecuación *2.29* con las frecuencias de los osciladores armónicos obtenidos de la aproximación quasi-armónica. En primera instancia, el PCA aplicado sobre trayectorias sin alinear produjo 2 coordenadas próximas a cero, esto ocasiona que las frecuencias estimadas a partir de la QHA diverjan (ecuación *2.26*) y por ende el valor de la entropía también será divergente. Dado que dichas coordenadas principales representarían coordenadas ligadas, es decir restricciones, no se tomarán en cuenta para el cálculo de la entropía.

Con estas consideraciones se estimó la entropía para cada trayectoria correspondiente a una conformación intermedia, obteniéndose en total 13 valores de entropía, los cuales se encuentran en el orden de 17 $[kcal/molK]$.

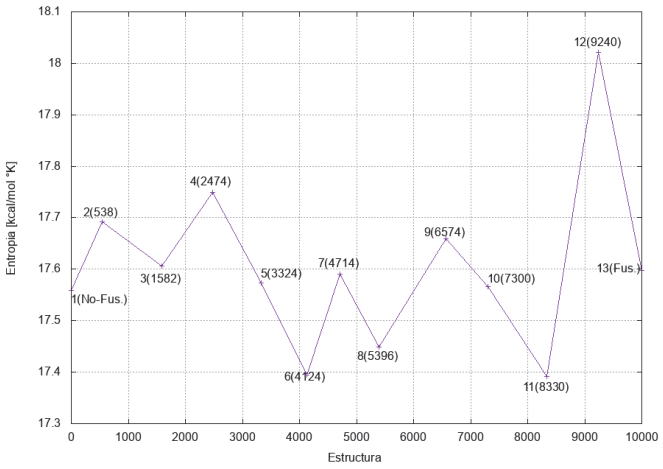

FIGURA 4.14: Entropía absoluta obtenida mediante la Aproximación Quasi-Armónica, el valor máximo de entropía $18{,}021\frac{kcal}{mol^{\circ}K}$ ocurre para la conformación 12 cuya estructura representativa es la 9240, en tanto que disminuye para la conformación fusogénica, aproximándose al valor $17{,}597\frac{kcal}{mol^{\circ}K}$ semejante al de la conformación inicial (no-fusogénica)

Se puede comprobar que la entropía aumenta considerablemente hasta $18{,}021\frac{kcal}{mol^{\circ}K}$ para la conformación inmediatamente anterior a la fusogénica, disminuyendo para el caso de la conformación fusogénica, hasta un valor de $17{,}597\frac{kcal}{mol^{\circ}K}$.

4.2.3. Cambio de Entropía

A partir de los valores obtenidos de entropía absoluta, se puede determinar el cambio de entropía entre conformaciones, tomando como valor de referencia el primer valor de entropía, el cual corresponde a la conformación no-fusogénica.

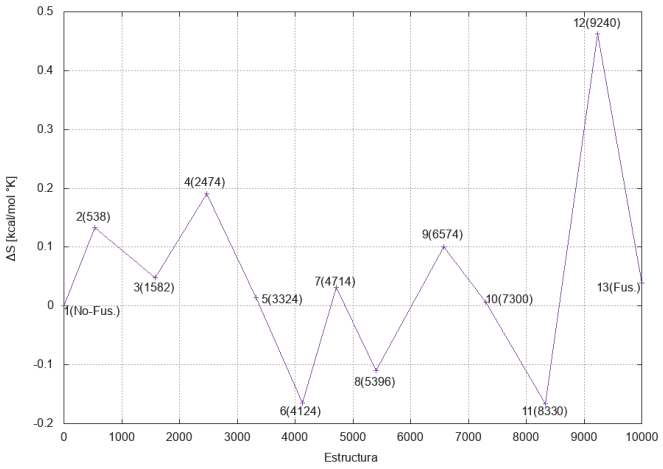

FIGURA 4.15: Diferencia de entropía entre conformaciones, tomando como referencia la entropía de la conformación inicial (no-fusogénica), el valor máximo de la diferencia de entropía es de $0{,}4622 \frac{kcal}{mol^{\circ} K}$

4.3. Energía Libre

Para estimar la energía libre de Gibbs ΔG, se requiere de dos componentes, la diferencia de energía interna ΔU, y la contribución de la entropía $-T\Delta S$.

A partir de la estimación de la diferencia de entropía, cuyo resultado se presentó en la sección anterior, se puede determinar el término de contribución de la entropía a la energía libre $-T\Delta S$, dada la temperatura $T = 300K$ a la que se realizó la dinámica molecular.

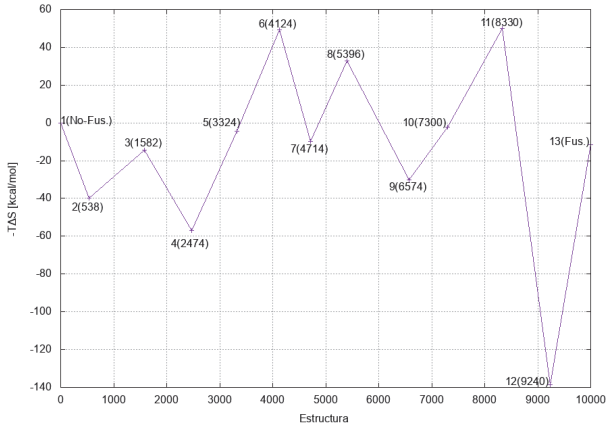

FIGURA 4.16: Contribución de la entropía a la energía libre

4.3.1. Energía Interna en la Trayectoria TMD

La energía interna se puede estimar utilizando el plugin *Namd-Energy* incluido en el software VMD [24], sobre la trayectoria TMD al igual que en [26]. Como resultado se obtienen 10000 valores de energía, correspondientes a cada estructura de la trayectoria, como se muestra en la siguiente figura:

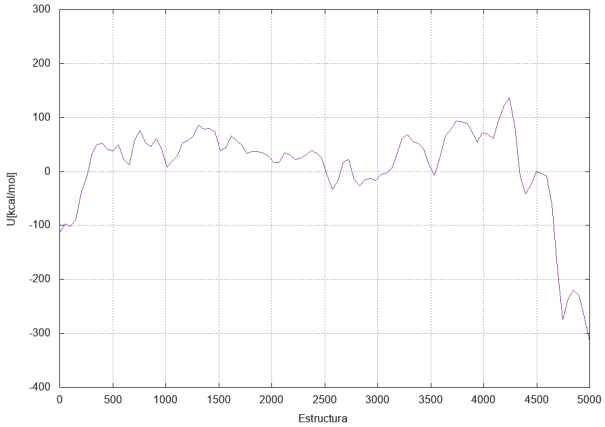

FIGURA 4.17: Energía interna U, estimada a partir de la trayectoria TMD

Para el análisis se tomó el valor de la energía interna para la estructura representativa de cada grupo, añadiéndole la contribución de la entropía, para así obtener la energía libre de Gibbs; los resultados de esta aproximación se muestran en la siguiente figura:

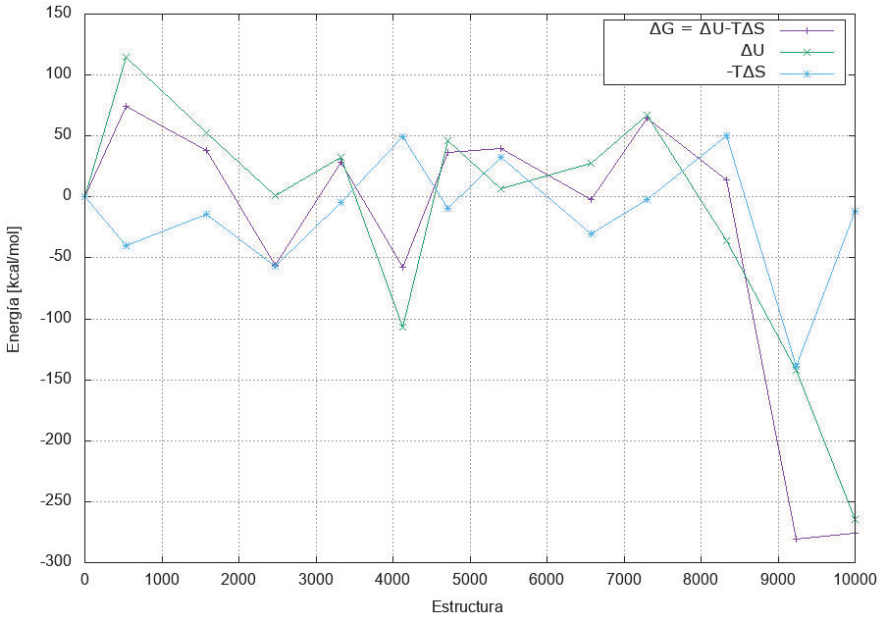

FIGURA 4.18: Energía libre de Gibbs, estimada utilizando la energía interna de la estructura representativa, obtenida a partir de la trayectoria TMD

Se puede observar, que el valor máximo de energía libre $74[kcal/mol]$ se presenta para el grupo $1(538)$, y se mantiene relativamente estable, después se reduce drásticamente, hasta el mínimo $-280{,}77[kcal/mol]$ en la estructura correspondiente al grupo $11(9240)$, después incrementa levemente hasta la estructura correspondiente a la conformación fusogénica.

4.3.2. Estimación utilizando las Trayectorias de Equilibración para cada Conformación Intermedia

Para estimar la energía interna, se utilizaron las trayectorias de equilibración para cada grupo, suponiendo que todas las estructuras son equivalentes y por tanto pertenecientes a una conformación, de forma que cada trayectoria se puede considerar como un ensamble, y por tanto la energía interna se puede estimar mediante la siguiente expresión: [53]:

$$\langle E \rangle = \frac{1}{M} \sum_{n=1}^{M} E^n \tag{4.1}$$

donde M es el total de estructuras en el ensamble (trayectoria), y n es la n-ésima estructura.

Para el análisis, resulta conveniente tomar un subconjunto de estructuras de la trayectoria de equilibración, las que tengan un valor de RMSD estable, para lo cual resulta conveniente analizar el RMSD de la trayectoria de equilibración de la estructura no-Fusogénica, la cual se presenta en la siguiente figura:

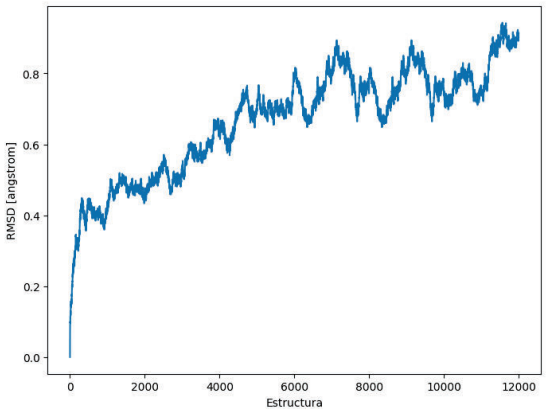

FIGURA 4.19: RMSD para la trayectoria de equilibración correspondiente a la estructura no-Fusogénica

Por tanto, se decidió tomar las últimas 2000 estructuras de las trayectorias de equilibración para cada grupo para el cálculo de la energía interna, considerando además que $U = \langle E \rangle$.

En la siguiente figura, se muestra la energía libre de Gibbs ΔG, además de las componentes que la constituyen: la energía interna ΔU, y la contribución de la entropía $-T\Delta S$.

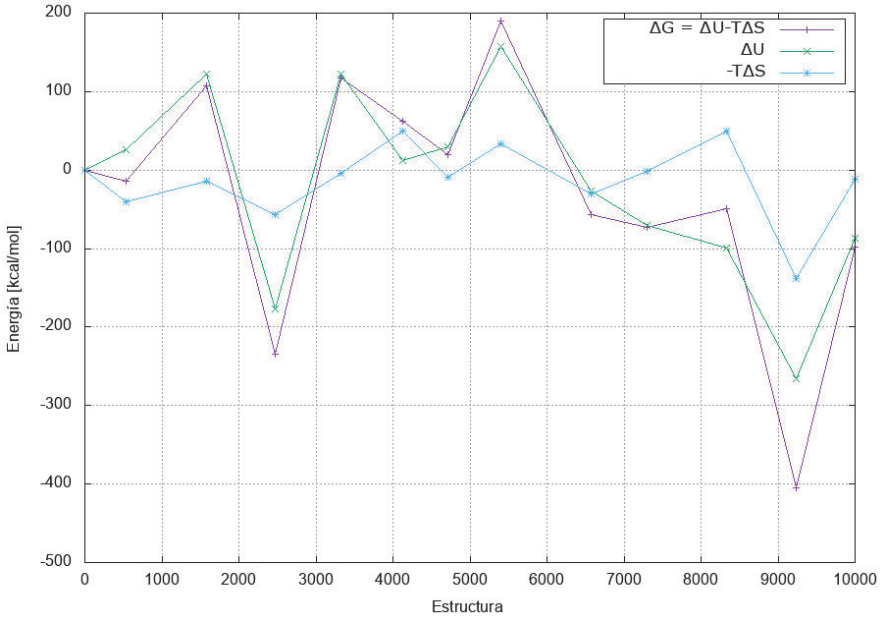

FIGURA 4.20: Energía libre de Gibbs, utilizando el promedio obtenido a partir de las trayectorias de equilibración para la energía interna

A partir de la figura, se puede comprobar que el valor máximo de energía libre $189{,}75[kcal/mol]$ se presenta para el grupo $6(5396)$, y se tienen dos valores mínimos, uno de $-234{,}55[kcal/mol]$ correspondiente al grupo $3(2474)$, y un mínimo global de $-404{,}72[kcal/mol]$ correspondiente al grupo $11(9240)$, el cual por lo tanto corresponde a la conformación intermedia más estable.

4.3.3. Comparación de los valores de Energía Libre obtenidos mediante los dos métodos

La energía libre obtenida directamente a partir del valor de la energía libre de la estructura representativa en la trayectoria TMD, fluctúa considerablemente en comparación con la obtenida a partir del promedio sobre el ensamble generado a partir de las trayectorias de equilibración.

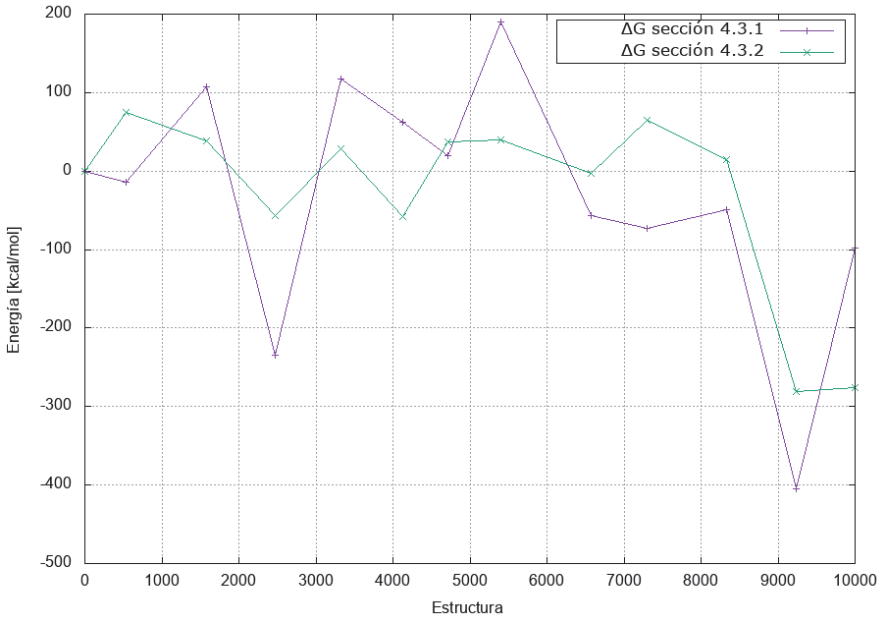

FIGURA 4.21: Comparación de la energía libre de Gibbs, obtenida mediante los dos métodos anteriores.

En las dos aproximaciones el valor de energía libre de la penúltima estructura representativa corresponde a un mínimo, sin embargo se puede ver que en la aproximación mediante el promedio sobre el ensamble, la energía del grupo correspondiente a la conformación fusogénica es comparable a este mínimo, sin embargo en la otra aproximación, el valor de energía es mayor que este valor mínimo; por tanto se puede ver que la aproximación mediante el promedio sobre el ensamble, provee un valor que se aproxima más a la realidad, ya que se ha comprobado de forma experimental [12] que el estado fusogénico es estable.

4.4. Descripción de la Barrera Energética

Con los valores de energía obtenidos en la sección anterior, se puede comprobar la viabilidad de posibles coordenadas de reacción, para lo cual se probaron los siguientes candidatos: La primera coordenada principal del PCA, $RMSD_0$, $RMSD_f$, Helicidad, $RMSD_{a0}$ y $RMSD_{af}$.

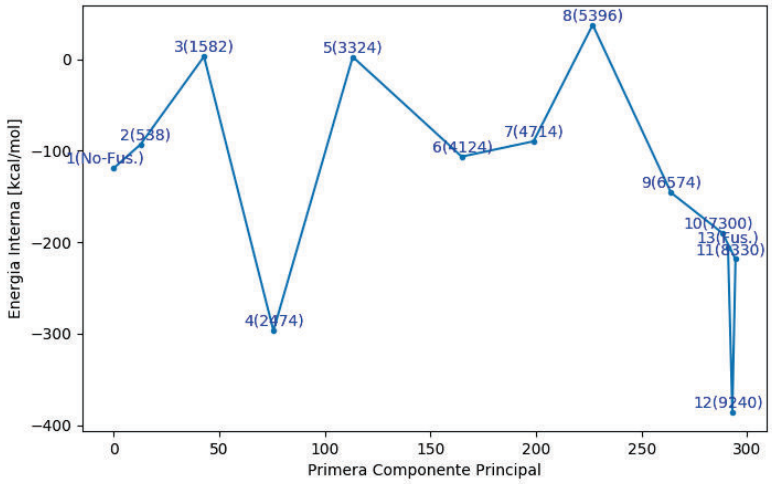

FIGURA 4.22: PCA, obtenido de la aplicación del análisis de componentes principales a las estructuras de la trayectoria TMD, se ha aplicado la transformación $PCA' = max(PCA) - PCA$ de forma que se muestre el incremento de la variable

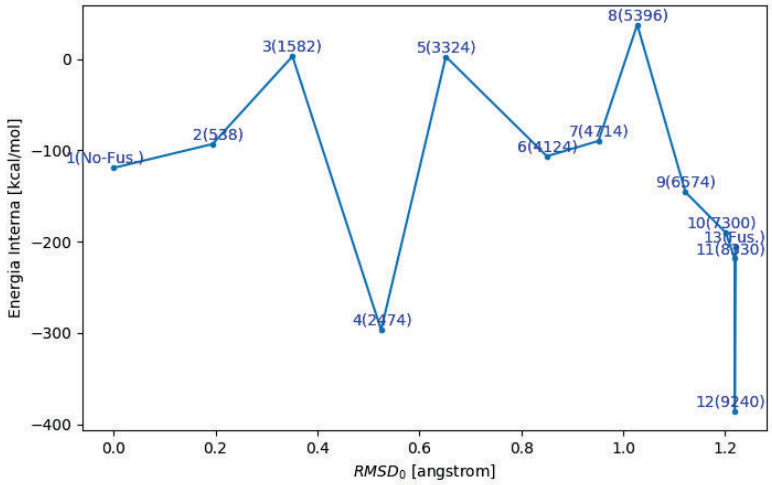

FIGURA 4.23: $RMSD_0$, este parámetro fue calculado con respecto a la estructura no-fusogénica

Tanto para el PCA como para el $RMSD_0$ como coordenadas de reacción, se puede ver el mínimo local de energía libre correspondiente a la cuarta conformación intermedia, en tanto que las conformaciones 10, 11, 12 y 13 se encuentran muy próximas en el eje correspondiente tanto al PCA como al $RMSD_0$, por lo cual se puede ver una correlación entre las dos posibles coordenadas de reacción.

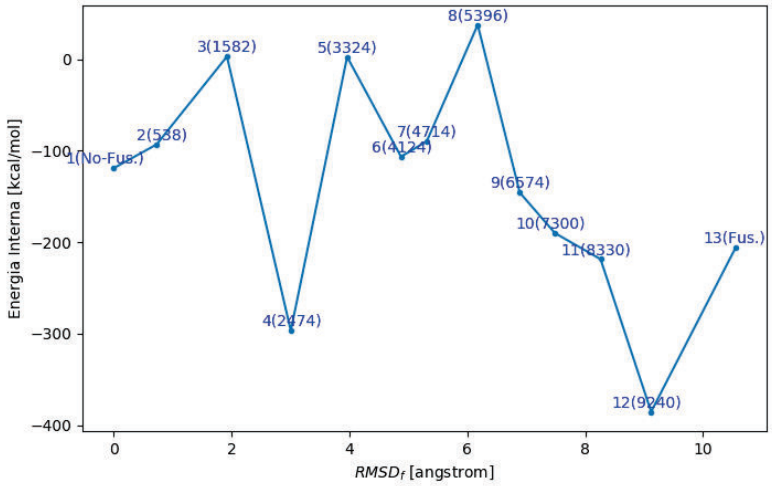

FIGURA 4.24: $RMSD_f$, calculado con respecto a la estructura fusogénica, se ha aplicado la transformación: $RMSD'_f = max(RMSD_f) - RMSD_f$, de forma que se muestre el incremento de la variable

Por otra parte para el $RMSD_f$ se puede observar que las conformaciones 10, 11, 12 y 13 están mejor espaciadas, y se manifiesta una posible transición natural entre las conformaciones 8, 9, 10, 11, 12.

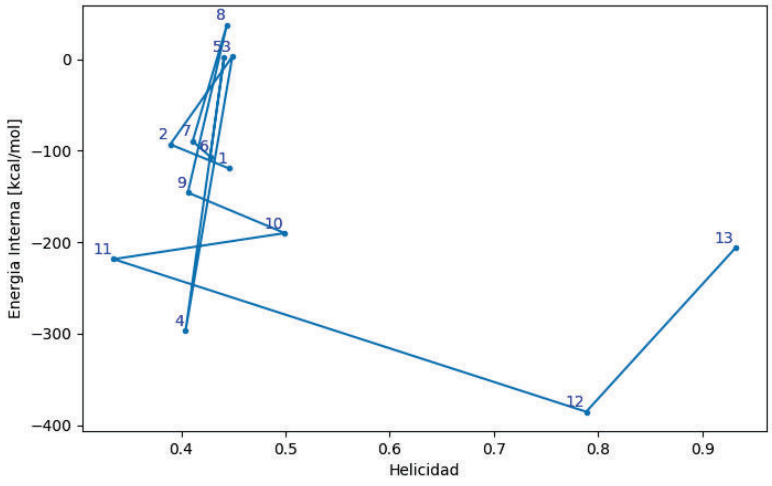

FIGURA 4.25: $Helicidad$

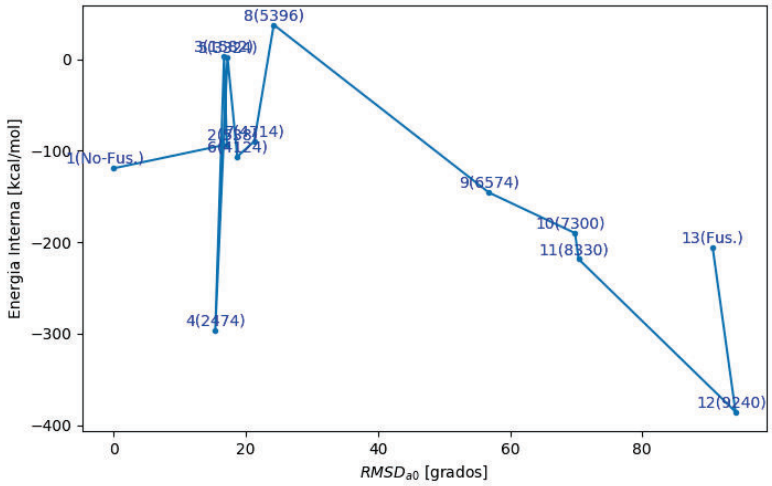

FIGURA 4.26: $RMSD_{a0}$, este parámetro fue calculado con respecto a la estructura no-fusogénica

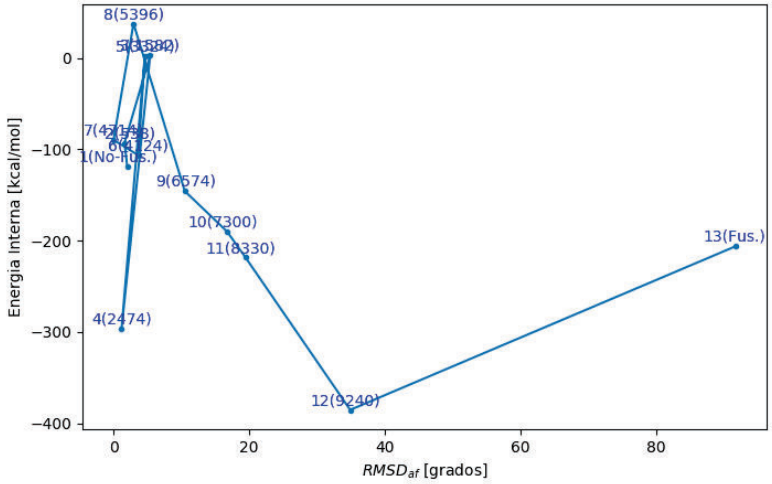

FIGURA 4.27: $RMSD_{af}$, calculado con respecto a la estructura fusogénica, se ha aplicado la transformación: $RMSD'_{af} = max(RMSD_{af}) - RMSD_{af}$, de forma que se muestre el incremento de la variable

En el análisis de las coordenadas de reacción en el espacio de Ramachandran (ϕ, ψ), la helicidad oscila en torno a 0.4 para las primeras 11 conformaciones intermedias, y se incrementa drásticamente para la conformación 12. Por otra parte el análisis del $RMSD$ en el espacio de Ramachandran calculado tanto con respecto a la conformación no-fusogénica como a la conformación fusogéncia, muestra que las conformaciones $2hasta7$ tienen distancias muy próximas, en tanto que a partir de la conformación 8 la distancia se incrementa. Es notable

también que en tanto que el $RMSD_{a0}$ para las conformaciones 12 y 13 tiene valores muy próximos, para el $RMSD_{af}$ se puede observar la transición de forma mas marcada, por lo cual esta coordenada describe mejor la transición en sus últimas etapas.

Se puede ver que de las 6 coordenadas probadas, la coordenada que mejor describe el sistema es $RMSD_f$, ya que permite apreciar la barrera energética durante todas las etapas de la transición.

Capítulo 5

Conclusiones

En el presente trabajo se realizó un análisis alternativo de las trayectorias de dinámica molecular correspondientes a la transición conformacional del lazo 36 de la Hemaglutinina obtenidas en un estudio anterior [26]. El análisis realizado en este trabajo abordó las limitaciones del estudio anterior, particularmente en lo que respecta al cálculo de la entropía y a los criterios para el agrupamiento de las estructuras. Las conformaciones definidas en el trabajo previo se examinaron teniendo en cuenta los parámetros de las estructuras secundarias en el plano de Ramachandran y el cálculo de la entropía se realizó utilizando la aproximación quasi-armónica [29].

En el trabajo anterior, se utilizaron la mitad de las estructuras de dinámica molecular disponibles, citando como dificultad el costo computacional [26] requerido para el análisis. En el presente trabajo, además de reproducir los resultados del agrupamiento anterior, se utilizaron todas las 10000 estructuras disponibles para el análisis de grupos. Comparando el tamaño de los grupos mostrado en la figura 4.1, para 5000 y 10000 estructuras se puede observar que el análisis de grupos produce grupos con tamaño diferente. Esto evidencia la posibilidad de existencia grupos adicionales, ya que se comprueba una dependencia del tamaño de la muestra utilizada para el agrupamiento. En particular existe la posibilidad de grupos adicionales en el grupo 11, pero también en los grupos $7, 8, 9$. Las estructuras representativas de los grupos obtenidos en el análisis del presente trabajo, se encuentran dentro del rango de las estructuras abarcadas por los grupos obtenidos en el trabajo anterior, a excepción del grupo $8B$ y $10B$; por otra parte, el valor de diferencia del RMSD entre las estructuras representativas obtenidas, es significativo únicamente para las estructuras representativas correspondientes a los grupos $(2, 2B)$ y $(8, 8B)$. Por tanto no sería necesario para posteriores trabajos, simular las estructuras representativas correspondientes al análisis del presente trabajo, sino resultaría conveniente, realizar simulaciones con las estructuras representativas de los posibles nuevos grupos encontrados con diferentes criterios de agrupación.

El análisis en el plano de Ramachandran para cada grupo, muestra que los grupos correspondientes a las conformaciones $1, 2, 3, 4, 5, 6, 10$ tienen una distancia euclídea en el plano de

Ramachandran de alrededor de 12 grados, a partir de la cual comienzan a aparecer grupos adicionales. Por otra parte los grupos $7, 8, 9, 11$ tienen una distancia euclídea de alrededor de 15 grados, lo que coincide con los resultados del análisis del agrupamiento en base a los grupos encontrados en [26], y reafirma la posibilidad de existencia de grupos adicionales.

Se conoce que en la simulación de dinámica molecular de equilibración, de la estructura representativa de cada grupo, se fijo el último átomo de la molécula con el fin de prevenir que abandone la burbuja de agua en la que se realizó la simulación. Esto implica que las coordenas (x, y, z) del último átomo son constantes, por tanto al aplicar el PCA se esperan al menos 3 coordenadas principales próximas a 0. La aplicación del PCA en este caso produjo 2 coordenadas próximas a cero, y una tercera coordenada, mucho menor que el resto, pero a su vez mucho mayor que cero. Una alternativa que se probó, fue con las estructuras alineadas de simulaciones MD, tomando como referencia la primera estructura. A priori se podría esperar obtener tras la aplicación del PCA, 6 coordenadas próximas a cero adicionales, 3 debidas a la simetría de traslación, 3 a la simetría de rotación, además de las 3 debidas a que el último átomo se encuentra fijo. Sin embargo, al aplicar el PCA se obtuvieron 6 coordenadas cuyo valor es mucho menor que los valores anteriores, pero mucho mas grandes que cero, y una coordenada adicional cuyo valor difiere significativamente en comparación a los anteriores, pero no coincide en orden de magnitud con los 6 planteados. Esto se debe a que el proceso de alineación mueve el átomo que estaba fijo, de forma que se pierden las restricciones correspondientes. Cabe notar que la entropía de las estructuras alineadas es mayor que las no alineadas, lo cual se puede esperar debido a que no hay coordenadas independientes (próximas a cero) en el primer caso.

Las frecuencias obtenidas a través de la aproximación quasi-armónica se encuentra en el rango entre 10^6 y $10^{15}[Hz]$, con la mayoría de frecuencias en el rango entre 10^{11} y $10^{14}[Hz]$, lo cual permite establecer una relación directa entre las frecuencias encontradas y el comportamiento molecular, ya que coinciden con el rango de las frecuencias de vibración molecular. Por otra parte, se comprobó que las frecuencias de menor valor, corresponden a modos colectivos de vibración, entre los cuales se analizó la oscilación del centro de masa, y se comprobó que esta oscilación tiene un máximo de frecuencia en torno a $10^8[Hz]$. De esta forma se puede verificar que la aproximación quasi-armónica, produce resultados consistentes con los resultados experimentales.

La estimación de la entropía a través de la aproximación quasi-armónica, provee valores de entropía del orden de $S \approx 17[kcal/molK]$, y valores de diferencia de entropía con un máximo de $\Delta S = 0{,}8[kcal/molK]$, lo cual produce una contribución energética de la entropía máxima de $-T\Delta S = -240[kcal/mol]$, este valor difiere en 2 órdenes de magnitud del valor reportado en trabajos anteriores [26, 19]; sin embargo la distribución de los valores obtenidos,

siguen una tendencia similar, con un cambio abrupto de $-T\Delta S$ en la conformación 12 cuya estructura representativa es la 9240.

Para la descripción de la barrera energética se probaron 6 posibles coordenadas de reacción, tanto para la primera coordenada principal como para el $RMSD_0$ se tiene un comportamiento similar, por lo cual se puede ver una posible correlación entre las dos coordenadas de reacción. En el análisis en el espacio de Ramachandran, la helicidad oscila para las primeras 11 conformaciones intermedias, y su valor incrementa únicamente para las dos últimas conformaciones. Por otra parte el análisis del $RMSD$ en el espacio de Ramachandran calculado tanto con respecto a la conformación no-fusogénica como a la conformación fusogéncia, muestra que las conformaciones $2\ hasta\ 7$ tienen distancias muy próximas, en tanto que a partir de la conformación 8 la distancia se incrementa. Es notable también que en tanto que el $RMSD_{a0}$ para las conformaciones 12 y 13 tiene valores muy próximos, para el $RMSD_{af}$ se puede observar la transición de forma mas marcada, por lo cual esta coordenada describe mejor la transición en sus últimas etapas. Por esto se puede ver que de las 6 coordenadas probadas, la coordenada que mejor describe el sistema es $RMSD_f$, ya que permite apreciar la barrera energética durante todas las etapas de la transición, y que hay una posible correlación entre la primera coordenada principal y el $RMSD_0$.

Para trabajos futuros, sería relevante utilizar las estructuras representativas encontradas mediante análisis de grupos, considerando como parámetro el RMSD para todas las estructuras disponibles de la simulación TMD, así como considerando también los resultados de usar como parámetro la distancia euclídea en el plano de Ramachandran, ya que los resultados de utilizar los dos criterios se pueden considerar como complementarios, a partir de los resultados obtenidos en el presente trabajo. Esto permitiría además encontrar valores adicionales de entropía, utilizando los mismos datos de simulaciones de dinámica molecular para cada conformación, ya que se podría aplicar la aproximación quasi-armónica sobre particiones de la trayectoria en subconjuntos de estructuras, que correspondan a los grupos adicionales. Otro aspecto a tomar en cuenta es que, en la aproximación quasi-armónica, la componentes de frecuencias bajas contribuyen mayoritariamente a la entropía, por ello resultaría relevante ampliar el análisis espectral presentado en la sección 4.2.1, de modo que incluya no solo la oscilación del centro de masa, sino también otros aspectos, por ejemplo la rotación de la molécula.

Bibliografía

[1] P. J. RUSSELL, *Genetics: A Molecular Approach*. Pearson, 2011.

[2] C.-I. BRÄNDÉN y J. TOOZE, *Introduction to Protein Structure*. Taylor & Francis, 1999.

[3] D. L. NELSON, A. L. LEHNINGER y M. M. COX, *Lehninger principles of biochemistry*. Macmillan, 2008.

[4] J. KURIYAN, B. KONFORTI y D. WEMMER, *The molecules of life: Physical and chemical principles*. Garland Science, 2012.

[5] G. t. RAMACHANDRAN y V. SASISEKHARAN, «Conformation of polypeptides and proteins», *Advances in protein chemistry*, vol. 23, págs. 283-437, 1968.

[6] L. PAULING, R. B. COREY y H. R. BRANSON, «The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain», *Proceedings of the National Academy of Sciences of the United States of America*, vol. 37, num. 4, pág. 205, 1951.

[7] K. A. DILL y H. S. CHAN, «From Levinthal to pathways to funnels», *Nature structural biology*, vol. 4, num. 1, págs. 10-19, 1997.

[8] K. A. DILL y J. L. MACCALLUM, «The protein-folding problem, 50 years on», *Science*, vol. 338, num. 6110, págs. 1042-1046, 2012.

[9] M. LEVITT, M. GERSTEIN, E. HUANG, S. SUBBIAH y J. TSAI, «Protein folding: the endgame», *Annual review of biochemistry*, vol. 66, num. 1, págs. 549-579, 1997.

[10] C. LEVINTHAL, «Are there pathways for protein folding», *J. Chim. phys*, vol. 65, num. 1, págs. 44-45, 1968.

[11] C. ANFINSEN, «Principles that govern the protein folding chains», *Science*, vol. 181, págs. 233-230, 1973.

[12] P. A. BULLOUGH, F. M. HUGHSON, J. J. SKEHEL y D. C. WILEY, «Structure of influenza haemagglutinin at the pH of membrane fusion», *Nature*, vol. 371, num. 6492, págs. 37-43, 1994.

[13] C. M. CARR y P. S. KIM, «A spring-loaded mechanism for the conformational change of influenza hemagglutinin», *Cell*, vol. 73, num. 4, págs. 823-832, 1993.

[14] N. K. SAUTER, J. E. HANSON, G. D. GLICK, J. H. BROWN, R. L. CROWTHER, S. J. PARK, J. J. SKEHEL y D. C. WILEY, «Binding of influenza virus hemagglutinin to analogs of its cell-surface receptor, sialic acid: analysis by proton nuclear magnetic resonance spectroscopy and X-ray crystallography», *Biochemistry*, vol. 31, num. 40, págs. 9609-9621, 1992.

[15] H. M. BERMAN, J. WESTBROOK, Z. FENG, G. GILLILAND, T. BHAT, H. WEISSIG, I. N. SHINDYALOV y P. E. BOURNE, «The protein data bank», *Nucleic acids research*, vol. 28, num. 1, págs. 235-242, 2000.

[16] Y. ZHOU, C. WU, L. ZHAO y N. HUANG, «Exploring the early stages of the pH-induced conformational change of influenza hemagglutinin», *Proteins: Structure, Function, and Bioinformatics*, vol. 82, num. 10, págs. 2412-2428, 2014.

[17] C. D. HIGGINS, V. N. MALASHKEVICH, S. C. ALMO y J. R. LAI, «Influence of a heptad repeat stutter on the pH-dependent conformational behavior of the central coiled-coil from influenza hemagglutinin HA2», *Proteins: Structure, Function, and Bioinformatics*, vol. 82, num. 9, págs. 2220-2228, 2014.

[18] H. S. CHOI, J. HUH y W. H. JO, «Electrostatic energy calculation on the pH-induced conformational change of influenza virus hemagglutinin», *Biophysical journal*, vol. 91, num. 1, págs. 55-60, 2006.

[19] H. MORALES y M. BAYAS, «Efecto de mutaciones puntuales en las transiciones conformacionales de un nanomotor biologico», *Revista Politecnica*, vol. 32, num. 3, págs. 63-70, 2013.

[20] S. BANTA, Z. MEGEED, M. CASALI, K. REGE y M. L. YARMUSH, «Engineering protein and peptide building blocks for nanotechnology», *Journal of nanoscience and nanotechnology*, vol. 7, num. 2, págs. 387-401, 2007.

[21] S. BANTA, I. R. WHEELDON y M. BLENNER, «Protein engineering in the development of functional hydrogels», *Annual review of biomedical engineering*, vol. 12, págs. 167-186, 2010.

[22] A. DUBEY, C. MAVROIDIS, A. THORNTON, K. NIKITCZUK y M. YARMUSH, «Viral protein linear (VPL) nano-actuators», en *Nanotechnology, 2003. IEEE-NANO 2003. 2003 Third IEEE Conference on*, IEEE, vol. 1, 2003, págs. 140-143.

[23] M. HAMDI, A. FERREIRA, G. SHARMA y C. MAVROIDIS, «Prototyping bio-nanorobots using molecular dynamics simulation and virtual reality», *Microelectronics Journal*, vol. 39, num. 2, págs. 190-201, 2008.

[24] W. HUMPHREY, A. DALKE y K. SCHULTEN, «VMD: visual molecular dynamics», *Journal of molecular graphics*, vol. 14, num. 1, págs. 33-38, 1996.

[25] M. MADHUSOODANAN y T. LAZARIDIS, «Investigation of pathways for the low-pH conformational transition in influenza hemagglutinin», *Biophysical journal*, vol. 84, num. 3, págs. 1926-1939, 2003.

[26] J. R. CALDERÓN FIGUEROA, «Descripción estadística del proceso de plegamiento del lazo 36 de la cadena AH2 de la Hemaglutinina», *Tesis EPN*, 2015.

[27] M. B. EISEN, P. T. SPELLMAN, P. O. BROWN y D. BOTSTEIN, «Cluster analysis and display of genome-wide expression patterns», *Proceedings of the National Academy of Sciences*, vol. 95, num. 25, págs. 14 863-14 868, 1998.

[28] R. MOJENA, «Hierarchical grouping methods and stopping rules: An evaluation», *The Computer Journal*, vol. 20, num. 4, págs. 359-363, 1977.

[29] M. KARPLUS y J. N. KUSHICK, «Method for estimating the configurational entropy of macromolecules», *Macromolecules*, vol. 14, num. 2, págs. 325-332, 1981.

[30] I. T. JOLLIFFE, *Principal Component Analysis*. Springer, 1986.

[31] J. N. R. JEFFERS, «Two Case Studies in the Application of Principal Component Analysis», *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 16, num. 3, págs. 225-236, 1967.

[32] R. E. MADSEN, L. K. HANSEN y O. WINTHER, «Singular value decomposition and principal component analysis», 2004.

[33] H. ABDI y L. J. WILLIAMS, «Principal component analysis», *Wiley interdisciplinary reviews: computational statistics*, vol. 2, num. 4, págs. 433-459, 2010.

[34] D. POOLE, *Álgebra lineal. Una introducción moderna*. Cengage Learning Editores, 2011.

[35] J. SCHLITTER, «Estimation of absolute and relative entropies of macromolecules using the covariance matrix», *Chemical Physics Letters*, vol. 215, num. 6, págs. 617-621, 1993.

[36] J. NUMATA, M. WAN y E. KNAPP, «Conformational entropy of biomolecules: beyond the quasi-harmonic approximation.», en *Genome informatics. International Conference on Genome Informatics*, vol. 18, 2007, pág. 192.

[37] H. SCHÄFER, A. E. MARK y W. F. van GUNSTEREN, «Absolute entropies from molecular dynamics simulation trajectories», *The Journal of Chemical Physics*, vol. 113, num. 18, págs. 7809-7817, 2000.

[38] I. ANDRICIOAEI y M. KARPLUS, «On the calculation of entropy from covariance matrices of the atomic fluctuations», *The Journal of Chemical Physics*, vol. 115, num. 14, págs. 6289-6292, 2001.

[39] H. MEIROVITCH, S. CHELUVARAJA y R. P. WHITE, «Methods for calculating the entropy and free energy and their application to problems involving protein flexibility and ligand binding», *Current Protein and Peptide Science*, vol. 10, num. 3, págs. 229-243, 2009.

[40] R. BARON, P. H. HÜNENBERGER y J. A. MCCAMMON, «Absolute single-molecule entropies from quasi-harmonic analysis of microsecond molecular dynamics: correction terms and convergence properties», *Journal of chemical theory and computation*, vol. 5, num. 12, págs. 3150-3160, 2009.

[41] K. HUANG, *Lectures on statistical physics and protein folding*. World Scientific, 2005.

[42] R. SUZUKI y H. SHIMODAIRA, «Pvclust: an R package for assessing the uncertainty in hierarchical clustering», *Bioinformatics*, vol. 22, num. 12, págs. 1540-1542, 2006.

[43] J. BIEN y R. TIBSHIRANI, «Hierarchical clustering with prototypes via minimax linkage», *Journal of the American Statistical Association*, vol. 106, num. 495, págs. 1075-1084, 2011.

[44] J. C. PHILLIPS, R. BRAUN, W. WANG, J. GUMBART, E. TAJKHORSHID, E. VILLA, C. CHIPOT, R. D. SKEEL, L. KALE y K. SCHULTEN, «Scalable molecular dynamics with NAMD», *Journal of computational chemistry*, vol. 26, num. 16, págs. 1781-1802, 2005.

[45] R CORE TEAM, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2015. dirección: http://www.R-project.org.

[46] G. VAN ROSSUM y F. L. DRAKE JR, *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam, 1995.

[47] J. D. HUNTER, «Matplotlib: A 2D graphics environment», *Computing In Science & Engineering*, vol. 9, num. 3, págs. 90-95, 2007. DOI: 10.1109/MCSE.2007.55.

[48] S. v. d. WALT, S. C. COLBERT y G. VAROQUAUX, «The NumPy array: a structure for efficient numerical computation», *Computing in Science & Engineering*, vol. 13, num. 2, págs. 22-30, 2011.

[49] T. WILLIAMS, C. KELLEY y MANY OTHERS, *Gnuplot 4.6: an interactive plotting program*, http://gnuplot.sourceforge.net/, abr. de 2013.

[50] R. T. MCGIBBON, K. A. BEAUCHAMP, M. P. HARRIGAN, C. KLEIN, J. M. SWAILS, C. X. HERNÁNDEZ, C. R. SCHWANTES, L.-P. WANG, T. J. LANE y V. S. PANDE, «MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories», *Biophysical Journal*, vol. 109, num. 8, págs. 1528-1532, 2015. DOI: 10.1016/j.bpj.2015.08.015.

[51] A. BARTH, «Infrared spectroscopy of proteins», *Biochimica et Biophysica Acta (BBA)-Bioenergetics*, vol. 1767, num. 9, págs. 1073-1101, 2007.

[52] K. DILL y S. BROMBERG, *Molecular driving forces: statistical thermodynamics in biology, chemistry, physics, and nanoscience*. Garland Science, 2010.

[53] J. THIJSSEN, *Computational physics*. Cambridge university press, 2007.

Anexos

A. Máximo de la Entropía para una Desviación Estandar Nula y Varianza Conocida

Para la estimación de la entropía se puede considerar un sistema cuántico, con estados discretos $|n\rangle$, con autovalores de energía ϵ_n asociados a una función de partición para un ensamble canónico de la forma [35]:

$$Z = \sum_{0}^{n} exp(-\beta\epsilon_n) \tag{1}$$

donde $\beta = 1/k_B T$.

Considerando la definición de la entropía:

$$S = -k_B \sum_{0}^{n} p_n ln(p_n) \tag{2}$$

donde la probabilidad p_n para cada estado viene dada por la expresión:

$$p_n = \frac{exp(-\beta\epsilon_n)}{Z} \tag{3}$$

podemos considerar que el sistema en su conjunto posee una masa m y una única coordenada x con desviación media nula $\langle x\rangle = 0$ y varianza $\langle x^2\rangle$ cuyo valor se asume conocido, y además se puede expresar en función de los valores de cada elemento individual del sistema:

$$\langle x^2\rangle = \sum_{0}^{n} p_n x_n^2 \tag{4}$$

Considerando como restricciones: la ecuación anterior, y la condición de normalización para las probabilidades $\sum_0^n p_n = 1$, se puede plantear el problema optimización, el cual siempre

producirá un máximo ya que la segunda derivada de la entropía es siempre negativa [37]:

$$\max_{p_n} \quad F : S = -k_B \sum_0^n p_n ln(p_n) \tag{5a}$$

$$\text{sujeto a} \quad g_1 : \langle x^2 \rangle = \sum_0^n p_n x_n^2 \tag{5b}$$

$$g2 : \sum_0^n p_n = 1 \tag{5c}$$

Utilizando los parámetros de Lagrange λ y μ, se obtiene la siguiente condición para cada n:

$$ln(p_n) + 1 + \lambda x_n^2 + \mu = 0 \tag{6}$$

Expresando p_n en términos de la energía, la condición puede ser expresada de la forma:

$$\beta(\epsilon_n) - \lambda x_n^2 = \mu - ln(Z) + 1 \tag{7}$$

B. Script para generar mapas de Ramachandran para cada grupo

Código en Python[46] para generar mapas de Ramachandran para cada grupo.

```
import csv
from math import *
import numpy as np
from pylab import *
import matplotlib.pyplot as plt

def main():
    pathAngulos =
     ↪ '/Tesis/Objetivo1/Ramachandran/angulos.txt'
    pathGrupos =
     ↪ '/Tesis/Objetivo1/TMDAlineado/num_gruposJC.txt'

    g = np.fromfile(pathGrupos,sep=" ")
    G = np.reshape(np.array([g, g]), 10000, order='F')

    ang = np.zeros((360036,3))
    i = 0

    with open(pathAngulos, 'rb') as csvfile:
        reader = csv.reader(csvfile, delimiter=' ',
         ↪ quotechar='|')
```

```
        for fila in reader:
            ang[i] = fila
            i += 1
    print ang
    print(ang.shape)

    grup = np.append(np.repeat(G, 36), np.repeat([11], 36))
    A = np.concatenate((ang, grup[:, None]), axis=1)

    for grupo in range(1,12):
        estructuras = A[A[:,3] == grupo]

        figure()
        scatter(estructuras[:, 1], estructuras[:, 2],
         ↪ marker='.', c=estructuras[:, 0], s=3,
         ↪ cmap='Spectral')
        cbar = colorbar()
        cbar.set_label('Estructuras Res')
        xlabel(r'$\Phi$ [grados]')
        xlim(-180, 180)
        ylabel(r'$\Psi$ [grados]')
        ylim(-180, 180)
        savefig('grupo'+str(grupo)+'.png')
        close()

if __name__ == "__main__": main()
```

C. Script para análisis de grupos utilizando la distancia en el plano de Ramachandran

Script en R [45] para realizar análisis de grupos, utilizando la distancia euclídea en el plano de Ramachandran como parámetro.

```
library(pracma)
library(readr)
library(sp)
library(mgcv)
library(protoclust)

angulos <-
 ↪ read_delim("~/Tesis/Objetivo1/Ramachandran/angulos.txt",
 ↪ " ", escape_double = FALSE, col_names = FALSE, trim_ws =
 ↪ TRUE)

```

```
numFrames <- as.integer(max(angulos[,1]))

matrizAngulos <- matrix(NA, nrow=10000, ncol=72)

for(frame in 1:10000){
  matrizFrame <- angulos[angulos[,1] == frame , ]
  angulosFrame <- t(matrizFrame[,c(2,3)])
  rowAngulos <- t(c(angulosFrame))
  matrizAngulos[frame, ] <- rowAngulos
}

distancia <- dist(matrizAngulos, method = "euclidean")/36
mat <- protoclust(distancia)
cut <- protocut
plot(mat)
```

D. Script para preparación de trayectorias

Código en Python [46] para convertir las coordenadas desde el formato xyz, a una matriz de trayectorias con la ponderación mediante las masas atómicas.

```
import csv
import numpy as np

def main():
    pathPosiciones = '/Tesis/Scripts/Alineacion/538_a.txt'
    pathMasas = '/Tesis/Objetivo2/MD/masas.txt'
    i = 0
    j = 0
    x = np.zeros((12002,1863))

    m = np.sqrt(np.fromfile(pathMasas,sep=" "))
    M = np.reshape(np.array([m, m, m]), 1863, order='F')

    print M

    with open(pathPosiciones, 'rb') as csvfile:
        reader = csv.reader(csvfile, delimiter=' ',
         ↪ quotechar='|')
        r = np.zeros(1863)

        for fila in reader:
            if(str(fila).__contains__('generated')):
                x[j] = np.multiply(r,M)
                r = np.zeros(1863)
```

```
                j += 1
                i=0
            elif(len(fila)>3):
                for columna in fila:
                    if columna.__contains__('.'):
                        r[i] = np.float(columna)
                        i+=1
    print x
    pathX = '/Tesis/Objetivo2/MD/matriz_538a.txt'
    np.savetxt(pathX, x, fmt='%f', delimiter=' ',
     ↪ newline='\n', comments='# ')

if __name__ == "__main__": main()
```

E. Script para estimación de la entropía

Script en R [45] para estimar la entropía a partir de trayectorias de dinámica molecular, utilizando la aproximación quasi-armónica.

```
library(pracma)
library(readr)
library(sp)
library(mgcv)
library(ggplot2)
library(ggfortify)

h_bar <- 1.054571e-34
k_B <- 1.38064852e-23
Ts <- 300.00
alpha <- h_bar/(k_B*Ts)
limite <- 1.0/alpha

estructuras <-
 ↪ read_delim("~/Tesis/Objetivo2/MD/matriz_538a.txt", " ",
                     escape_double = FALSE, col_names =
                      ↪ FALSE, trim_ws = FALSE)
cp <- prcomp(estructuras)
w <- 5e12*cp$sdev

s <- c()
for (i in 1:length(w)){
  if(w[i]>1){
    s <- c(s, k_B*alpha*w[i]/(exp(alpha*w[i])-1) -
     ↪ k_B*log(1-exp(-alpha*w[i])))
  }
```

```
}

Sq <- sum(s) / 6.9477e-21 # El valor de la entropía se obtiene en kcal/mol K

write.table(format(w, scientific = TRUE), "/Users/dc/Tesis/Objetivo2/ResultadosR1/w_538a.txt",
            row.names = FALSE, col.names = FALSE, quote = FALSE)
```

F. Script para el cálculo de la helicidad

Script en R [45] para calcular la helicidad, a partir de los ángulos de ramachandran de residuos, que se encuentren dentro del área correspondiente a α-hélices.

```
library(pracma)
library(readr)
library(sp)
library(mgcv)

alphaHelice <- matrix(c(-180.0, -34.9, -164.3, -42.9, -133.0, -42.9, -109.4, -32.2, -106.9, -21.4, -44.3, -21.4, -44.3, -71.1, -180.0, -71.1), nrow = 8 , ncol = 2, byrow = TRUE)

angulos <- read_delim("~/Tesis/Ramachandran/angulos.txt", " ", escape_double = FALSE, col_names = FALSE, trim_ws = TRUE)

numFrames <- as.integer(max(angulos[,1]))

helicidad <- c()

for(frame in 1:numFrames){
  matrizFrame = angulos[angulos[,1] == frame , ]
  angulosFrame = data.matrix(matrizFrame[,c(2,3)])
  helicidad <- c(helicidad, sum(in.out(alphaHelice, angulosFrame), na.rm=TRUE))
}

write.table(helicidad, "/Users/dc/Tesis/Ramachandran/helicidad.txt", row.names = FALSE, col.names = FALSE)
```