Análisis de trajectorías largas de plegamiento de Proteínas a través de rasgos conformacionales

por

Mauricio Martínez Jiménez

TESIS PRESENTADA EN CUMPLIMIENTO PARCIAL DE
LOS REQUERIMIENTOS PARA EL GRADO DE

MAGISTER EN INGENIERÍA CON ÉNFASIS EN INGENIERÍA DE SISTEMAS

en

Escuela de Ingeniería de Sistemas y Computación

(Ingeniería de Sistemas)

UNIVERSIDAD DEL VALLE

(Valle del Cauca, Colombia)

20 de noviembre de 2018

Agradecimientos

Agradezco al mi director Pedro Moreno y a mi codirector Luis Garreta por su apoyo para la realización de este trabajo.

Resumen

Esta tesis toma como referencia los conceptos creados en un trabajo anterior sobre análisis de trayectorias de plegamiento de proteínas que concluye sobre la presencia de estados intermedios durante su plegamiento. Con estos conceptos analizamos aquí trajectorias grandes de plegamiento de proteínas simuladas con Dinámica Molecular que generalmente contienen una inmensa cantidad de conformaciones que dificultan de gran manera su análsis y la manera de obtener información útil de las mismas. Para solucionar esto, creamos aqúi un algoritmo que permite reducir sustancialmente el número de conformaciones de una trayectoría y así obtener una muestra representativa de la misma enfocándose principalmente en preservar la dinámica de la trayectoria original en cuanto a los eventos principales y la relación de tiempo entre las conformaciones. Utilizando este algoritmo realizamos la reducción de algunas de las trayectorías simuladas por la supercomputador Anton, especializada para simulaciones de Dinámica Molecular, y construimos un flujo de trabajo para analizar estas trajectorias de acuerdo a los conceptos definidos en el trabajo de referencia. Los resultados nos mostraron que los estados intermedios encontrados en el trabajo de referencia no están presentas en el plegamiento de estas proteínas de acuerdo a las simulaciones de Anton, lo cual es coherente con la literatura y con el tipo de proteínas analizadas en el trabajo de referencia y con el tipo de proteínas analizadas en esta tesis.

Índice general

Agradecimientos II

Resumen III

Índice general IV

Índice de cuadros VII

Índice de figuras VIII

1. Introducción 1

2. Marco teórico 5
2.1. Estructura de una proteína 5
2.2. Proceso de plegamiento de proteínas 6
2.3. El problema de plegamiento de proteínas 7
2.3.1. Predicción de estructura de proteínas 8
2.3.2. Modelos de plegamiento de proteínas 8
2.4. Rutas de plegamiento de proteínas 9
2.5. Intermedios de plegamiento 10
2.6. Plegamiento de proteínas computacional 11
2.6.1. Simulaciones de Dinámica Molecular (DM) 12
2.7. Métodos de Comparación de Estructuras de Proteínas 13
2.7.1. RMSD 13
2.7.2. GDT 14
2.7.3. TM-score 14

3. Estado del arte 15
3.1. Rasgos conformacionales inherentes 15

4. Datos y métodos 22
4.1. Datos de plegamiento de proteínas 22
4.1.1. Conjunto de Datos de Anton 22

4.2. Evaluación de propiedades 24
4.3. Preparación de datos 24
4.4. Análisis Estadístico 25
4.4.1. Escala Multidimensional 25
4.4.2. Análisis de Componentes Principales 25
4.4.3. Selección y rotación de componentes 26
4.4.4. Clustering Jerárquico 27
4.4.5. Validación de Cluster 27
4.4.5.1. Índice Silhouette 28
4.4.5.2. Validación por remuestreo 28

5. Algoritmo Rápido de Reducción de Trajectorias 30
5.1. Antecedentes 30
5.1.1. Simulaciones de Plegamiento 31
5.1.2. Algoritmos Rápidos de Agrupamiento de Secuencias . 32
5.2. Algoritmo de Reducción de Trayectorias de Plegamiento . . . 33
5.3. Datos y Métodos 33
5.3.1. Comparación de Estructuras de Proteínas 33
5.3.2. Selección de Estructuras Representativas 34
5.3.3. Trayectorias de Plegamiento de Proteínas 34
5.4. Detalles de Implementación 34
5.4.1. Descripción de Programas 34
5.4.2. Ejecución 35
5.4.3. Requisitos 36
5.5. Resultados y Discusión 36

6. Aplicación Virtual 39

7. Software Toolkit para análisis de trayectorias de plegamiento de proteínas 41
7.1. Propiedades de proteínas 41
7.1.1. Contactos Nativos 43
7.1.2. Orden de Contacto 43
7.1.3. Radio de giro 44
7.1.4. Enlaces de Hidrógeno 44
7.1.5. Área de superficie Accesible 45
7.1.6. Vacíos 45
7.1.7. Error cuadrático Medio 46
7.1.8. Error cuadrático Medio Local (Local RMSD) 47
7.1.9. La energía potencial 47

7.1.10. El momento dipolar 48
7.1.11. Residuos en estructuras secundarias correctas o en cualquiera 48
7.1.12. Puntaje estructural 49
7.1.13. Clusters rígidos, regiones estresadas y grados de libertad 49
7.2. Toolkit para cálculo de métricas 50
7.2.1. Funciones 50
7.2.2. Implementación 50
7.3. Instalación y despliegue 51
7.3.1. Resultados 53

8. Resultados 54

9. Conclusiones 56

Bibliografía 58

Índice de cuadros

7.1. Resumen de las propiedades seleccionadas para describir el plegamiento. 42
7.2. Principales funciones del toolkit de análisis de proteínas . . . 51

Índice de figuras

2.1. Un polipéptido como cadena de residuos de aminoácidos. . . 6
2.2. Los cuatro niveles de estructura de las proteínas. 7
2.3. La vista del embudo de plegamiento. 9

3.1. Distribución final de las propiedades de los componentes. . . 16
3.2. Interpretación de los componentes. 17
3.3. Representación de los rasgos IC en términos de sus propiedades. 18
3.4. Niveles de plegamiento de proteínas identificados. 20
3.5. Distribución del puntaje ICF para los cuatro niveles de plegamiento. 21

4.1. Proteínas seleccionadas del conjunto de datos de Anton. . . . 23

5.1. Reducción de las trayectorias cortas de plegamiento para las proteínas 1FCA1 y 2YCC. 37
5.2. Reducción de una trayectoria larga de plegamiento. 38

8.1. Niveles de plegamiento de la trayectoria de la proteína de la vilina. 55
8.2. Niveles de plegamiento para una ruta de 4 estados 55

Capítulo 1

Introducción

En general la bioinformática se ha caracterizado por trabajar con conjuntos de datos voluminosos e incrementales y métodos complejos de análisis de datos[51]. Los avances tecnológicos de la actualidad han abierto muchas posibilidades, especialmente en el campo de la bioinformática, aumentando aún más las oportunidades de análisis, pero también los retos que esto conlleva. Diversos sistemas biológicos pueden ser analizados con gran eficiencia y en general a bajo costo. La consecuencia natural de esto es una gran cantidad de datos por procesar, de los cuales se espera extraer información que permita conocer más sobre algún proceso biológico en particular[71, 52]. Esto conduce además a otros retos: el procesamiento de grandes cantidades de datos requiere de técnicas que hagan viable su ejecución teniendo en cuenta las restricciones técnicas que se tienen al ser la capacidad computacional un recurso limitado.

Dentro de los diversos temas que se trabajan en bioinformática, el problema del plegamiento de proteínas es uno de los que mayor atención ha captado y uno de los que lleva más tiempo sin resolver[18, 29]. Básicamente, se trata de endender como una proteína, partiendo de una secuencia lineal de aminoácidos, logra plegarse rápidamente para tomar una forma tridimensional específica que determina su función[28, 27]. Y más que eso: ¿se puede este proceso biológico simular computacionalmente? Desde hace años se han desarrollado técnicas que permiten, con distinto grado de precisión, predecir dicha estructura tridimensional a partir de la información sobre los aminoácidos que conforman la proteína. Es un problema aún abierto porque se trata de procesos que demandan gran cantidad de recursos computacionales, y su resolución tendría un impacto en el entendimiento de ciertas enfermedades y en el desarrollo de determinados medicamentos.

Sobre este tema, el trabajo de rasgos conformacionales desarrollado por Luis Garreta en el grupo de Bioinformática y Biocomputación de la Universidad del Valle ofrece una visión importante del proceso de plegamiento al identificar estados de plegamiento en diversas simulaciones de plegamien-

to de proteínas. Ofrece además un puntaje de plegamiento que puede ser aplicado a cualquier conformación individual de los miles que pueden hacer parte de una simulación (trayectoria) y calcular que tan plegada o desplegada se encuentra, usando para esto la información obtenida y analizada estadísticamente acerca de propiedades físicas y estructurales de dichas conformaciones. Como fuente de datos se ulilizaron trayectorias de plegamiento cortas (200∼500 conformaciones) de protéinas relativamente grandes (alrededor de 200 amonoácidos) que fueron simuladas con el método conocido como Probabilistic Roadmap Method[96]. Estas proteínas y sus trayectorías se diferencian completamente de las que usamos en esta tesis ya que primero, son simuladas con Dinámica Molecular[63]; segundo, corresponden a proteínas pequeñas (30∼100) aminoácidos; y tercero, son trayectorias muy largas (32000 a más de 1 millón de conformaciones).

En el trabajo de rasgos conformacionales se encontró que las proteínas analizadas en la trayectoria pasan por 4 estados en todo el proceso, cada uno con caracteísticas específicas que permite realizar la asociación de cada conformación en la trayectoría a uno de esos cuatro estados. La pregunta que surge al observar los resultados es si este número de estados es algo que se puede considerar fijo para todas las proteínas, o si por el contrario, es algo específico al tipo de proteínas que se analizaron en el momento. Que distintas proteínas se comporten distinto en cuanto a los estados observables en su plegamiento es importante porque puede ser un aspecto relevante a tener en cuenta en el diseño de nuevos algoritmos de predicción de estructura terciara, de manera que puedan comportarse de manera distinta de acuerdo al tipo de proteína que se esté evaluando. Sin embargo, el calculo de las propiedades que se requieren para obtener los puntajes de plegamiento pueden ser extensos si se trabaja con simulaciones de varios microsegundos, las cuales pueden constar de millones de archivos de secuencias de proteínas. Actualmente dichas trayectorías son más fáciles de obtener que antes, especialmente con la aparición de supercomputadores especialmente diseñados para ejecutar simulaciones de plegamiento de proteínas. De esta manera, se hace nesesario indagar acerca de nuevas técnicas que nos permitan la manipulación de grandes cantidades de datos de simulaciones, con la limitación de la capacidad con la que cuentan los sistemas de cómputo que tenemos al alcance.

Para reducir estas trayectorias los métodos actuales buscan conjuntos de conformaciones representativas, que generalmente utilizan métodos de agrupamiento donde se construye una matriz con las distancias entre cada

una de las conformaciones, usualmente se usa la distancia conocida como RMSD o *Root Mean Square-Deviation*. Estos agrupamientos se vuelven muy costosos en tiempo y recursos computacionales cuando se trata de muchas conformaciones y por esta razón los algoritmos buscan simplificar estos costos, como por ejemplo, reducir el número de átomos que comparar en las conformaciones (solo carbonos alfa).

Otra forma de reducir estas trayectorias es crear agrupamientos rápidos que no tengan que comparar todas las conformaciones, parecido a lo que realiza el algoritmo de Hobohm&Sander [43] para comparar secuencias de ADN. En este trabajo presentamos un algoritmo rápido para reducción de trayectorias de plegamiento de proteínas que toma como base la idea del algoritmo de Hobohm&Sander y que se basa en tres estrategias: primero una partición de la trayectoria en múltiples secciones; segundo, una reducción local muy rápida sobre cada una de ellas que aprovecha el tiempo de ocurrencia de las conformaciones; y tercero, una reducción global que busca encontrar las conformaciones más representativas de cada partición. Estas tres estrategias permiten que este algoritmo sea fácilmente paralelizable, obtenga unos resultados previos de forma rápida, y de esos resultados seleccione los más importantes.

Nuestro enfoque trabaja sobre las trayectorias de plegamiento ya realizadas y disponibles como un conjunto de snapshots a determinado *time step* y sobre estás realiza la selección de las más representativas por segmentos de tiempo que se ingresan como parámetros. El resultado es otra trayectoría con estructuras que representan de forma resumida la trayectoria original y que contiene los eventos o estructuras principales. El algoritmo se puede ajustar con diferentes parámetros para definir el tamaño de las particiones en número de conformaciones y número de representativas por cada partición, por ejemplo se puede definir que cada 1000 conformaciones se seleccionen 500 representativas, lo que reduce a la mitad el número de conformaciones, o cada 10000 conformaciones se seleccionen 100, lo que la reduce a la trayectoria en 99 %.

En este trabajo presentamos el análisis de las trayectorias de simulación de plegamiento de proteínas de Anton, un supercomputador especialmente diseñado para simulaciones de dinámica molecular, con los conceptos definidos en el trabajo de rasgos conformacionales. Para ello creamos un algoritmo para reducir trayectorias, inspirado en uno existente usado para reducción en secuencias de ADN. Creamos un pipeline con los las herramientas

del framework del trabajo citado y aplicamos el pipeline a las trayectorias reducidas. Con esto pudimos obtener los estados por los que pasan dichas proteínas, y los resultados fueron diferentes a los del trabajo previo al no encontrar cuatro estados sino dos, lo que coincide con la bibliografía para este tipo de proteínas que son más pequeñas. El pipeline se despliega en una aplicación virtual lo que facilita el uso de la aplicación al independizarlo de un sistema operativo específico.

El siguiente trabajo se divide en varios capítulos. Primero se presentan los principales conceptos relacionados con el plegamiento de proteínas y algunas de las aproximaciones computacionales más utilizadas en este campo. Luego se muestra el estado del arte enfocandose principalmente en los principales aspectos del trabajo de sobre rasgos conformacionales. Después se describen las trayectorias usadas como datos y los métodos para analizarlos. Seguidamente se detalla el tookit utilizado para el cálculo de las propiedades de las proteínas. Luego se muestran los resultados obtenidos y por último se presentan las conclusiones a las que se llegaron.

Capítulo 2

Marco teórico

En este capítulo presentamos algunos conceptos básicos de la estructura de la proteína y el plegamiento de proteínas. Describimos las dos visiones del problema del plegamiento de proteínas, una relacionada con la predicción de la estructura, y la otra relacionada con la explicación del proceso de plegamiento de proteínas inherente. A continuación, presentamos el concepto de rutas e intermedios. Y finalmente, describimos brevemente la técnica de dinámica molecular para estudiar y analizar el movimiento de las moléculas.

2.1. Estructura de una proteína

Las proteínas son cadenas lineales de residuos de aminoácidos (residuos en adelante), que adoptan estructuras tridimensionales (3D) únicas y que están involucradas en casi todas las funciones de todos los organismos vivos. Están asociados a funciones que van desde la replicación del ADN hasta el movimiento muscular, el transporte de moléculas dentro y fuera de la célula, y la catálisis de reacciones importantes, entre muchas otras.

La mayoría de los residuos tienen una estructura común de un *backbone* central con un átomo principal de carbono alfa ($C\alpha$) con cuatro partes adjuntas: un grupo amino;un grupo carboxilo; un átomo de hidrógeno; y el grupo R o cadena lateral que distingue un aminoácido de otro (ver Figure 2.1). Cada par de residuos se conecta entre sí formando un enlace peptídico que establece la cadena principal o *backbone* del polipéptido.Algunos de estos enlaces, los únicos enlaces donde el $C\alpha$ participa, permiten rotaciones que le dan flexibilidad a la proteína y por lo tanto la adopción de diferentes formas de proteína.

Pero esta visión de una proteína como una simple secuencia lineal de residuos o *estructura primaria* es sólo una de las cuatro abstracciones utilizadas para describir la estructura de la proteína (ver Figura 2.2). Las otras tres: secundaria, terciaria y cuaternaria, involucran las regularidades y principios subyacentes que finalmente conducen a su forma final y propiedades funcionales características. La *estructura secundaria* se asocia con conforma-

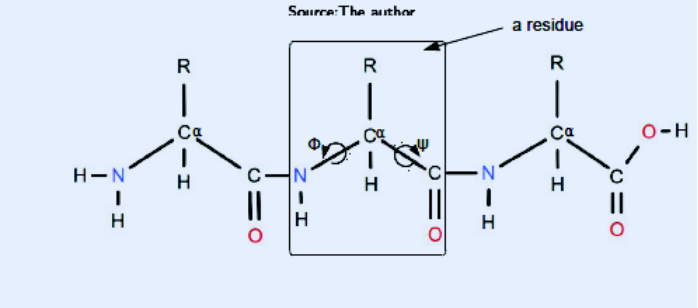


Figura 2.1: **Un polipéptido como cadena de residuos de aminoácidos.** Tres residuos unidos entre sí. La libertad de rotación se puede describir por ángulos de torsión conocidos como ángulos diedros. Las dos rotaciones principales asociadas a cada residuo vienen dadas por el ángulo ϕ (phi, que representa la rotación a lo largo del enlace $N - C\alpha$) y el ángulo ψ (psi, que representa la rotación a lo largo del enlace $C\alpha - C$).

ciones locales que forman estructuras regulares o elementos estructurales secundarios (EESs) como la hélice α, el hilo β y los giros. La *estructura terciaria* describe cómo estos EESs están organizados espacialmente dando la forma 3D final de la proteína llamada el *estado plegado* o el *estado nativo*. Y la *estructura cuaternaria* describe la disposición de varias cadenas de proteínas en un complejo de proteínas.

2.2. Proceso de plegamiento de proteínas

Una proteína se pliega desde una cadena no estructurada (estado desplegado), con alta energía libre, hasta un estado estable final (estado nativo), con baja o mínima energía libre. Durante este proceso, denominado *proceso de plegamiento*, una proteína experimenta interacciones sucesivas entre los átomos de sus residuos (y también los átomos del disolvente circundante) hasta alcanzar un estado estable que determina fuertemente su función biológica. Estas interacciones pueden ser de diferentes maneras: interacciones cortas entre residuos cercanos en secuencias, e interacciones grandes que involucran partes de la proteína que están muy distantes en secuencia.

Aunque los residuos de proteínas se mantienen unidos por enlaces peptídicos, las interacciones más importantes que estabilizan las proteínas provienen de fuerzas no covalentes débiles como las interacciones electrostáticas, los enlaces de hidrógeno, los puentes salinos y las fuerzas de van-der-Waals.[106].Las interacciones electrostáticas o carga-carga ocurren en-

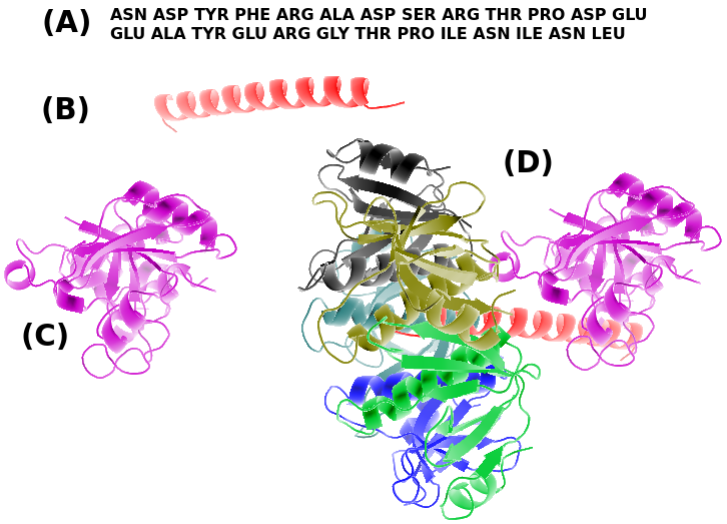


Figura 2.2: **Los cuatro niveles de estructura de las proteínas.** A) Estructura primaria (residuos de la cadena de polipéptidos. (B) Estructura secundaria (una hélice alfa). (C) Estructura terciaria (Una cadena completa de proteínas) (D) Estructura cuaternaria (Múltiples cadenas)

tre átomos debido a la atracción y repulsión de cargas opuestas y las mismas cargas parciales, respectivamente. Los enlaces de hidrógeno ocurren entre átomos unidos a átomos de hidrógeno positivamente polarizados (donantes) y negativamente polarizados (aceptadores). Los puentes salinos se producen entre los átomos donantes y los aceptadores totalmente cargados. Las interacciones van-der-Waals ocurren entre átomos adyacentes que no están cargados y que no están enlazados, y que surgen de las fluctuaciones en la distribución de los electrones.

2.3. El problema de plegamiento de proteínas

El problema del plegado de proteínas se centra en la forma en que una proteína se pliega desde su secuencia de aminoácidos a una estructura 3D estable. Es uno de los rompecabezas más importantes, complejos y fascinantes de la historia de la ciencia. Muchas áreas de la ciencia se han unido para buscar una solución, pero hasta hoy sólo se obtienen aproximaciones a la estructura tridimensional de las proteínas. El proceso de plegado es relevante para dos problemas diferentes pero relacionados[76, 16]: la predicción de

la estructura 3D a partir de su secuencia de aminoácidos y la comprensión del proceso subyacente de plegado. En el primero, el interés principal radica únicamente en la predicción de la estructura de una proteína diana desconocida. En este último, el enfoque se centra en la comprensión de los mecanismos, fuerzas e interacciones que impulsan a las proteínas a alcanzar un estado estable final.

2.3.1. Predicción de estructura de proteínas

La predicción de la estructura de la proteína se ha llevado a cabo principalmente mediante técnicas de comparación y simulación. La comparación es útil cuando existe conocimiento de otras proteínas, con propiedades similares a las de la proteína diana. Si no existe una proteína similar, la principal técnica teórica es simular el proceso de plegado. Estas simulaciones se realizan principalmente mediante dinámica molecular, que es un método computacional que permite predecir las propiedades estáticas y dinámicas de la proteína a partir de las interacciones subyacentes entre sus moléculas. Además de esta técnica, existen otros métodos para simular y estudiar el plegamiento de proteínas, como el Probabilistic Roadmap Method que construye la ruta de plegamiento de proteínas o secuencia de eventos que la proteína probablemente sigue para alcanzar su estado nativo.

2.3.2. Modelos de plegamiento de proteínas

Para el segundo problema, han surgido varios modelos para explicar y comprender el mecanismo de plegamiento de proteínas[76, 23], tres representantes son el modelo framework, en el cual las estructuras secundarias se forman primero y luego se difunden rápidamente y colisionan para propagarse a la estructura terciaria; el modelo de nucleación, proponiendo que un núcleo inicial de estructura secundaria local se forme primero y luego se propague y crezca para formar las estructuras terciarias; y finalmente el modelo de colapso hidrofóbico, en el cual el colapso hidrofóbico impulsa el plegamiento formando un núcleo compacto o glóbulo fundido que se reordena para formar la estructura terciaria.

En los últimos años, ha surgido una "nueva visión" del plegado de proteínas conocida como la teoría del *embudo de plegamiento*.Visualiza el proceso de plegado como una especie de embudo de libre energía, en el que varias conformaciones desplegadas de una proteína en el borde, caracterizadas por una alta energía libre y un alto grado de entropía conformacional, pueden

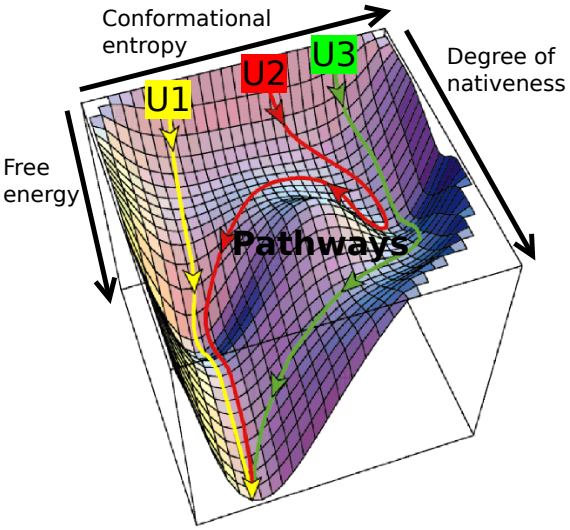


Figura 2.3: **La vista del embudo de plegamiento.** A medida que la proteína se pliega, su energía libre disminuye y su proporción de contactos nativos crece (Grado de cercanía al estado nativo). El ancho del embudo (entropía conformacional) representa la libertad conformacional de la cadena. Varias conformaciones desplegadas U1, U2, y U3 pueden seguir muchos caminos pasando a través de estados intermedios y de transición que pueden estar presentes a lo largo del embudo de libre energía. Estos estados pueden ralentizar o acelerar la ruta o senda de plegado, redirigiendo la proteína a una ruta más empinada (izquierda amarilla 'más rápida') o a rutas redondas (rojas y verdes más lentas)

plegarse siguiendo muchos caminos, hasta alcanzar su estado plegado (nativo) en un solo mínimo global en el fondo. (ver Figura 2.3). Esta teoría da otra visión del plegamiento de proteínas sin invalidar los modelos anteriores. A medida que una proteína se pliega, las trayectorias del movimiento se encuentran en embudos más estrechos con presencia reducida de estados conformacionales. Una proteína puede seguir muchos caminos pasando a través de intermedios y estados de transición presentes a lo largo del embudo.

2.4. Rutas de plegamiento de proteínas

Muchas proteínas se pliegan a sus conformaciones nativas en menos de unos segundos porque el plegado de proteínas no es una búsqueda aleatoria del nativo entre todas las conformaciones posibles. Esta afirmación fue

probada por Levinthal a finales de los años 60.[59] con la argumentación de que las proteínas siguen rutas específicas de plegado de proteínas que son secuencias de eventos (intermedios de plegado) que pliegan la proteína desde el estado desplegado hasta el estado nativo.

La visión de un único sendero de plegamiento formulada por Levinthal fue debatida cuando nuevas teorías, como la del embudo de plegamiento, que propone la existencia de múltiples rutas. En la teoría del embudo de plegamiento las proteínas se pliegan, moviéndose desde la parte superior a la inferior del embudo, por un sesgo general de la superficie de la energía que reduce el problema de la búsqueda a escalas de tiempo biológicamente relevantes sin necesidad de seguir un camino específico pero con la posibilidad de llegar al nativo por muchos de ellos.[56].

Sin embargo, el concepto clásico de una ruta de plegado como una secuencia de eventos seguida de proteínas, sigue siendo válido en esta "nueva visión" del plegado. Los múltiples y microscópicos caminos que una proteína puede seguir, sólo están cubriendo una ruta más general y macroscópica.[77, 101]. Los estados por los que pasa una proteína se interpretan en términos de conjuntos de conformaciones. Y la secuencia de eventos seguida por las proteínas corresponde a encontrar uno de los muchos miembros del conjunto, resultando al final en muchas rutas posibles.

2.5. Intermedios de plegamiento

Aunque la importancia de los productos intermedios en el plegado y la función de las proteínas ha sido reconocida desde hace mucho tiempo, siguen abiertas preguntas fundamentales sobre el papel que desempeñan en el plegado. [14, 101]. ¿Pueden considerarse como meros hitos cinéticos que indican el camino a seguir por la proteína? ¿Son responsables de los estados fallidos y, por lo tanto, son los progenitores de las enfermedades humanas? O por el contrario, ¿su presencia sirve para redirigir la proteína a la vía "correcta" y así poder corregir cualquier error acumulado?

En la ruta para alcanzar su estado nativo, una proteína pasa a través de estados intermedios formando estructuras intermedias o *intermedios de plegamiento*, como argumentó Levinthal con las vías de plegamiento de proteínas. Pero su presencia fue objeto de mucha controversia[14] dado que no podían detectarse en muchas proteínas, como por ejemplo el gran grupo de proteínas con un simple plegado de dos estados o una transición directa del estado desplegado al plegado sin población de intermedios. Ahora que los métodos experimentales de plegado han mejorado, su presencia es más

evidente y el debate ha cambiado a las cuestiones de qué papel juegan en el plegado y cómo se caracterizan [19, 101, 14].

Por las razones anteriores, actualmente no existe una definición establecida de lo que es un intermedio plegable, sin embargo muchos autores coinciden en características comunes sobre ellos:

Los intermedios son estructuras metaestables parcialmente plegadas, presentan una estructura secundaria similar a la nativa con interacciones terciarias débiles o ausentes, y una entropía sustancial debido a que el pobre empaquetado de residuos como cadenas laterales no está fijado y los residuos del núcleo hidrofóbico permanecen parcialmente solvatados. [101, 8, 14, 55, 17].

En el mismo sentido, no existe un tipo único de intermedios, ya que pueden presentar características diferentes según el nivel de plegamiento que exhiben. Como las estructuras nativas secundarias y supersecundarias pueden variar de parciales a completas, y el empaquetado de residuos puede ser débil o fuerte, muchos autores han identificado dos tipos de intermedios[9, 101, 87, 102, 34, 86]: *Intermedios tempranos*, más cercanos al estado desplegado, con falta de elementos de estructuras secundarias e interacciones terciarias muy débiles o completamente ausentes, y que evidencian un núcleo hidrofóbico muy poco compacto. Y, *intermedios tardíos*, más cercanos al estado nativo, y caracterizados como una estructura metaestable, semicompacta, parcialmente plegada con estructura secundaria similar a la nativa pero sin interacciones terciarias completas, y evidenciando un núcleo considerable de grupos no polares debido al pobre empaque de residuos.

2.6. Plegamiento de proteínas computacional

Las predicciones de la estructura de una proteína se han llevado a cabo principalmente usando técnicas de comparación y simulación. La comparación es útil cuando existe información de otras proteínas con características similares a la objetivo. En caso contrario la técnica a usar es la simulación del proceso. Estas simulaciones se realizan principalmente usando dinámica molecular, la cual es una técnica computacional que permite predecir propiedades estáticas y dinámicas de la proteína a partir de las interacciones entre sus moléculas.

Si bien es cierto que las técnicas para estudiar el plegamiento de proteínas ha mejorado en los últimos tiempos, aún se encuentran limitados en

cuanto a los detalles del mismo. Por ejemplo, los estudios experimentales de plegamiento de proteínas se ven obstaculizados por el hecho de que sólo se pueden obtener datos estructurales de baja resolución con una resolución temporal suficiente. Por esta razón, varias estrategias computacionaes han emergido como enfoques complementarios, siendo en algunos casos el único mecanismo para obtener una buena aproximación de los detalles de la dinámica de una proteína.

2.6.1. Simulaciones de Dinámica Molecular (DM)

En la actualidad la técnica de Dinámica Molecular (DM) es una de las principales herramientas para el estudio de moléculas y produce movimientos físicos de átomos altamente realísticos, así como trajectorias de plegamiento de proteínas [78]. Uno de us principales resultados es una serie de snapshots de las posiciones y velocidades de los átomos como una función del tiempo, representando una trayectoria del sistema, lo que corresponde a una ruta de transición de alta resolución [78, 49, 66]. DM lleva la traza de las posiciones y las velocidades de los átomos de la proteína y como interactúan entre ellos y responden a fuerzas externas. DM calcula el movimiento de todas las partículas de las proteínas intergrando numéricamente las ecuaciones de movimiento de Newton de manera simultánea y repetidamente en pequeños pasos de tiempo durante un periodo determinado. La ecuación en su forma más simplista establece que:

$$\text{Fuerza} = \text{Masa} \times \text{Aceleración}$$

donde la fuerza en un átomo dado depende de las interacciones con todos los demás. Modelar las fuerzas en este proceso requiere la definición de una función de potencial muy detallada y precisa en la forma de ecuaciones y parámetros que describan las interacciones entre las partículas de un sistema y retornen su energía potencial como función de sus conformaciones. Si bien DM produce simulaciones físicamente realísticas, es computacionalmente costoso y solamente se pueden simular proteínas pequeñas en tiempos factibiles, convirtiéndose de esta manera en un desafío computacional la simulación de proteínas más grandes en escalas de milisegundos [37][78]).

2.7. Métodos de Comparación de Estructuras de Proteínas

Cuantificar la diferencia entre dos estrucutras de una misma proteína puede parecer simple pero no es un problema trivial, y las distintas aproximaciones para abordarlo están en contínua evolución. En 1994 se originó CASP (Critical Assessment of techniques for protein Structure Prediction) como un experimiento de comunidad científica, el cual brinda la posibilidad de evaluar métodos en predicciones «ciegas» de proteínas recientemente resueltas pero sin publicar, convirtiendose así en una medida del estado del arte en modelamiento de estructuras de proteínas a partir de su secuencia de aminoácidos[73]. Con esto se puede decir que los métodos de comparación de estructuras han estado en desarrollo activo y se han usado en el campo de la modelación computacional específicamente para medir la calidad de los modelos predichos. Sin embargo, aunque estos modelos fueron originalmente desarollados para evaluar estos modelos simulados, ahora su aplicación se extiende además a la identificación, evaluación, entendimiento y predicción de cambios conformacionales en proteínas, siendo esto un punto clave en su funcionamiento biológico[54].

Dentro de los métodos más conocidos se puede citar el RMSD global, el GDS y el TM-score. A continuación se presenta una descripción general de cada uno de ellos, así como los casos en los que unos resultan más adecuados que los demás.

2.7.1. RMSD

La medida RMSD (Root Mean Square Deviation) es la medida cuantitativa de similitud entre dos coordenadas atomicas superpuestas más usada comunmente. Esta medida puede ser calculada para cualquier tipo y subset de atomos. Un uso común es calcualr el RMSD sobre los carbonos alfa ($C\alpha$), lo que reduce la cantidad de cálculos a realizar. Una de las desventajas de RMSD radica en que es muy sensible a los errores. Incluso si dos estructuras son casi identicas excepto un tramo, la superimposición de las estrucutras no puede ser realizada de la manera más efectiva por el algoritmo que optimiza el RMSD global.

Otro aspecto importante observado en esta métrica es un efecto observado en sistemas macromoleculares muy grandes, donde los efectos de la dimensionalidad producen lo que se conoce como «la maldición de la dimensionalidad», haciendo referencia a una capacidad decreciente para dis-

criminar la diferencia entre pares de conformaciones a medida que aumenta el tamaño del sistema e impactando de esta manera el análisis basado en RMSD[89].

2.7.2. GDT

También conocido como GDT TS (por total score). Sirve para medir la similitud entre dos estructuras de proteínas con una secuencias de aminoácidos idénticas pero una estructura terciaria distinta. Es una métrica pensada para ser una medida más precisa que RMSD, la cual, como se mencionó, es muy sensible a las regiones atípicas de modelados pobres aun cuando el resto de la estructura sea razonablemente similar. En general, cuanto más alto es el valor de GDT_TS, mejor es un modelo determinado comparado con una estructura de referencia. El valor GDT se calcula como el conjunto más largo de átomos carbonos alfa en la estructura modelo que se encuentran a una determinada distancia de la proteína referencia. Dicha distancia está determinada por un umbral específico. En general el GDT score se calcula usando varios umbrales, y típicamente un aumento en el umbral implica un aumento en el score. Si esto no es así, es posible que sea un indicador de que la divergencia entre la estructura referencia y el modelo es muy alta. Uno de los problemas de la técnica es que la elección del valor de los umbrales es subjetiva y necesitan ser ajustados manualmente para diferentes categorías de objetivos de modelado[108].

2.7.3. TM-score

TM-score (Template modeling score) es otra medida de similitud entre dos proteínas que tienen distintas estructuras terciarias. Fue concebida como una medida más precisa que RMSD y GDT. TM-score representa la diferencia entre dos estructuras con un puntaje que va de 0 a 1, donde 1 indica un emparejamiento perfecto de ambas proteínas. Un puntaje de menos de 0.20 indica una comparación entre dos proteínas aleatorias no relacionadas entre sí. Esta métrica está diseñada para ser independiente de la longitud de las proteínas que se están comparando[111].

Capítulo 3

Estado del arte

En el trabajo realizado por Luis Garreta se aplicaron diversas técnicas estadísticas para obtener un puntaje de grado de plegamiento en diversas trayectorias de plegamiento de proteínas. Con este puntaje se pudo identificar varios estados del proceso de plegado los cuales no eran directamente observables a través de el cálculo de propiedades estructurales y energéticas en las conformaciones. A continuación se presentan los principales conceptos de dicha investigación, la cual es usada en este trabajo, donde se extiende a otras trayectorías, y donde se usa en conjunto con técnicas de reducción de datos para ampliar su capacidad de aplicación.

3.1. Rasgos conformacionales inherentes

Definiciones

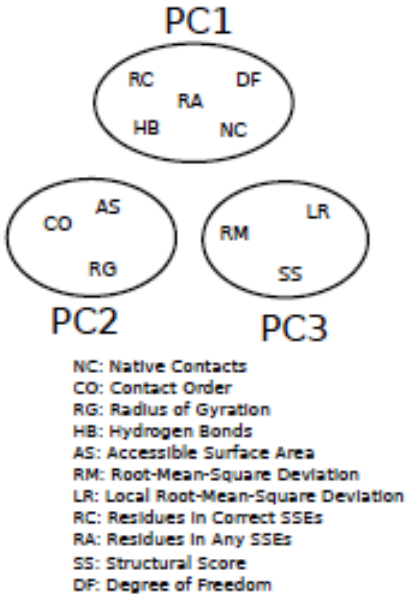
Rasgos de plegamiento de proteínas

El Análisis del Componente Principal realizado por el trabajo de Luis Garreta mostró tres componentes principales PC1, PC2 y PC3. El componente PC1 se cargó principalmente con cinco propiedades; mientras que PC2, y PC3, se cargaron principalmente cada uno con tres propiedades, como se muestra en la matriz de cargas de la Figura 3.1. El componente PC1 se cargó con contactos nativos, enlaces de hidrógeno, residuos en estructuras secundarias correctas, residuos en cualquier estructura secundaria y grados de libertad (NC, HB, RC, RA y DF, respectivamente). El componente PC2 incluye orden de contacto, radio de giro y área de superficie accesible (CO, RG y AS, respectivamente). El componente PC3 se cargó con la desviación cuadrática media de la raíz (RMSD), la RMSD local y la puntuación estructural (RM, LR, SS, respectivamente).

Inspeccionando las propiedades principales que se cargaron en cada uno de los tres componentes conservados (Figure 3.1.b) se puede observar que hubo una asociación con tres rasgos de plegamiento ocultos o no medibles directamente: *estabilidad, compactibilidad, and semejanza al estado nativo*

Properties	Components		
ID	PC1	PC2	PC3
NC	.72		
CO		-.93	
RG		.64	
HB	.72		
AS		.82	
RM			-.61
LR			-.84
RC	.73		
RA	.89		
SS			.81
DF	-.73		

(a) Matrix of Loadings Final PCA.



(b) Component as Groups of Properties.

Figure 3.1: **Distribución final de las propiedades de los componentes.** (a) Matriz de cargas del PCA final con las principales propiedades (alto peso) de cada componente. b) Una vista de los componentes como grupos de propiedades interrelacionadas que muestran alguna estructura oculta.

(ver Figura 3.2). El ***componente de estabilidad*** fue influenciado principalmente por la propiedad relacionada con los enlaces de hidrógeno (HB) cuyas interacciones estabilizan las estructuras proteicas. Sus otras propiedades muestran las consecuencias de estas interacciones, como el número de contactos nativos, los residuos en la estructura correcta o cualquier estructura secundaria, y los grados de libertad que se reducen a medida que los enlaces de hidrógeno imponen restricciones a las rotaciones de los enlaces [42]. El ***componente de compactibilidad*** fue influenciado por el radio de giro y la superficie accesible, dos propiedades relacionadas con la compactibilidad de las proteínas [65]. El ***componente de semejanza al estado nativo*** estuvo influenciado por propiedades relacionadas con la similitud estructural entre una estructura objetivo y la nativa, descrita por RMSD, RMSD local y la puntuación estructural.

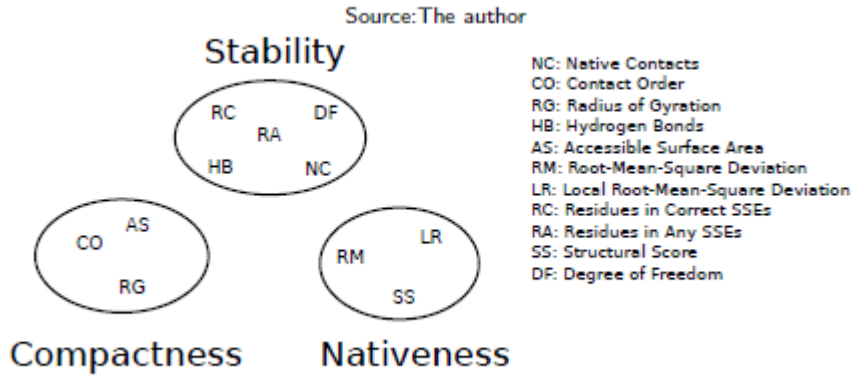


Figure 3.2: **Interpretación de los componentes.** De acuerdo con las propiedades que se cargaron en los tres componentes se asociaron con tres características intrínsecas del plegamiento de proteínas: estabilidad, compactibilidad y semejanza al estado nativo.

Rasgos conformacionales inherentes

Con la interpretación anterior se propuso los componentes encontrados como rasgos conformacionales inherentes del plegamiento de proteínas, los cuales se denominaron *rasgos IC*. Los valores de cada característica en términos de sus propiedades físicas pueden calcularse siguiendo el formulario general utilizado en PCA para calcular las puntuaciones de los componentes[41]: Cada característica se toma como una suma ponderada de las propiedades de la proteína, con cargas (pesos) proporcionales a la fuerza de la relación entre la característica y las propiedades de la proteína (ver Figura 3.3).

El vector de tres rasgos $[F_1, F_2, F_3]$ para una conformación de proteína correspondiente con los tres componentes principales PC1, PC2 y PC3 es calculado al multiplicar el vector de sus 11 propiedades $[p_1, ..., p_{11}]$ por la matriz de cargas calculada para los tres componentes $[[L_{1,1}, ...L_{1,11}], [L_{2,1}, ...L_{2,11}], [L_{3,1}, ...L_{3,11}]]$ como:

$$\begin{aligned} F_1 &= p_1 * L_{1,1} + ... + p_k * L_{1,k} + ... + p_{11} * L_{1,11} \\ F_2 &= -(p_1 * L_{2,1} + ... + p_k * L_{2,k} + ... + p_{11} * L_{2,11}) \\ F_3 &= p_1 * L_{3,1} + ... + p_k * L_{3,k} + ... + p_{11} * L_{3,11} \end{aligned} \tag{3.1}$$

donde $F_1, F_2,$ y $F3$ corresponden a los rasgos de IC de estabilidad, com-

Figure 3.3: **Representación de los rasgos IC en términos de sus propiedades.** Los tres componentes definen un espacio en 3D que se interpreta como los tres rasgos de IC: estabilidad (PC1, eje de rojo), compactibilidad (PC2, eje verde) y semejanza con el nativo (PC3, eje azul). Pueden ser representados por un valor numérico en términos de la distancia de sus propiedades físicas desde el origen (líneas de puntos) a lo largo del eje de la característica IC correspondiente.

pactibilidad y semejanza al estado nativo, respectivamente. El signo negativo en la definición del segundo rasgo (F_2), correspondiente a la compactibilidad, fue introducido para unificar la tendencia del rasgo IC a incrementar en una ruta que se aproxima al estado nativo. Su componente principal PC2 decrece a medida que el plegamiento progresa, contrario al comportamiento de los otros dos componentes.

Ahora podemos describir cualquier conformación de la proteína en términos de los tres rasgos IC: estabilidad, compactibilidad y semejanza con el estado nativo, en lugar de utilizar el gran conjunto de propiedades físicas. Además, estas características del IC se convierten en propiedades observables que pueden ser medidas. Al vector de rasgos tridimensionales compuesto de estabilidad, compacidad y semejanza con el nativo nos referimos como *estado de plegado* de una conformación. El estado de plegado ofrece una representación concisa y compacta del grado en que se pliega una determinada conformación, resumiendo sus propiedades individuales y sin depender de su tamaño o topología de la conformación.

El puntaje de grado de plegamiento

Los hallazgos anteriores sugieren que los rasgos miden en cierta medida la cantidad o grado de plegamiento para las conformaciones de las rutas. Con esto se creó un puntaje que usa los tres rasgos para obtener una aproximación del grado de plegamiento (ver abajo). El puntaje definido fue nombrado *inherent conformational feature score,* o ***ICF score*** y ofrece la aproximación al grado de plegamiento de una conformación de proteína al calcular la información numérica de los tres rasfos. El puntaje toma en cuenta la importancia de los rasgos de acuerdo a la varianza de cada componente y es calculado como la suma ponderada de los valores individuales de la rasgo IC, como se muestra en la fórmula siguiente:

$$ICF(c) = \frac{v_1 * F_1(c) + v_2 * F_2(c) + v_3 * F_3(c)}{V} \tag{3.2}$$

Medida de similitud conformacional

La transformación anterior muestra una conformación proteica en términos de un conjunto reducido de rasgos que incorpora un conjunto mayor de diferentes propiedades físicas estructurales y energéticas. Ahora, usamos esta transformación para definir una medida de similitud para comparar dos conformaciones de proteínas con respecto a sus rasgos de IC. La medida nos permite saber cuán lejos o cerca están un par de conformaciones con respecto a su proceso de plegado y permite caracterizar las conformaciones de una vía de plegado de proteínas en términos de grupos de conformaciones con características conformacionales comunes de acuerdo a su grado de plegado.

Con esto se define la medida de similitud de dos proteínas como la distancia Euclidiana en tre los vectores de sus rasgos IC: estabilidad (ST), compactibilidad (CP) y semejanza con la estructura nativa (NT). Formalmente, dadas dos conformaciones de proteínas a y b y sus tres puntajes de los rasgos (ST_a, CP_a, NT_a), y (ST_b, CP_b, NT_b), respectivamennte, se define la distancia Euclidiana (δ) entre ellos como:

$$\delta(a,b) = \sqrt{(ST_a - ST_b)^2 + (CP_a - CP_b)^2 + (NT_a - NT_b)^2} \tag{3.3}$$

Niveles de plegamiento

El plegamiento de la proteína puede ser tomado como una acumulación de cambios del estado desplegado al estado nativo. Consecuentemente, una

proteína asume estados conformacionales dinámicamente diferentes que comparten características conformacionales similares. Utilizando esta visión del plegamiento de proteínas y las definiciones anteriores, en los análisis realizados por Luis Garreta[XXXX] se encontró que el plegamiento sigue cuatro niveles principales: desplegado, intermedio temprano, intermedio tardío y plegado, como se muestra en la figura 3.4.

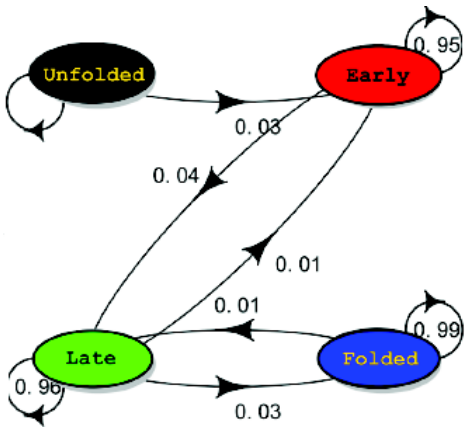


Figure 3.4: **Niveles de plegamiento de proteínas identificados.** Los grupos de conformaciones pueden verse como estados metaestables de niveles de plegado o plegado. Las proteínas comienzan en un nivel desplegado, pasan a través de intermediarios tempranos y tardíos, y terminan en un nivel plegado. Los números representan las probabilidades de transición entre estados de acuerdo con los análisis antes mencionados.

Usando un enfoque estadístico sobre el conjunto de las vías de plegamiento de proteínas simuladas de Amato[[96]], Luis Garreta resumió la distribución de la puntuación del grado de plegado (ICF) en los rangos que se muestran en la figura 3.5.

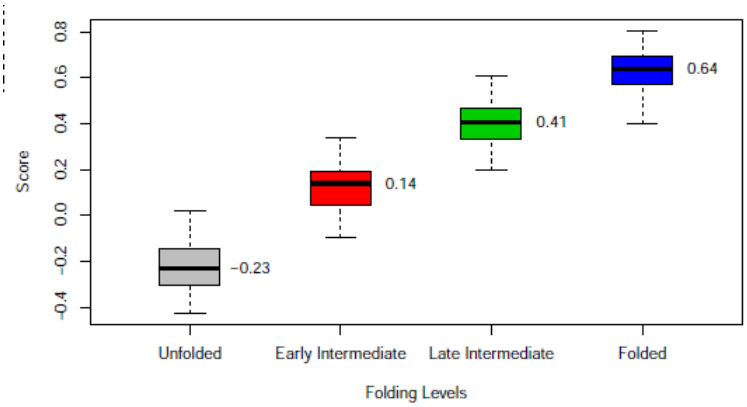


Figure 3.5: **Distribución del puntaje ICF para los cuatro niveles de plegamiento**. Los valores de las puntuaciones aumentan a medida que los niveles de plegado pasan de desplegado a plegado. Cada nivel de plegamiento está bien diferenciado de los demás, como muestra la media para cada nivel.

Clasificador de niveles de plegamiento

Utilizando la medida de similitud introducida anteriormente (3.3) se construyó un clasificador que asigna el nivel de plegado a una determinada conformación de la proteína: para cada conformación de la proteína c se puede evaluar su distancia a la conformación central g_k como $\delta(c, g_k)$. El nivel de plegamiento $lev(c)$ se asigna como el valor de k que minimiza la distancia entre c y g_k.

$$lev(c) =_{g_k} \delta(c, g_k) | k = 1, ..., 4 \tag{3.4}$$

Capítulo 4

Datos y métodos

Este capítulo contiene información sobre los datos de plegado de proteínas en los que se basa este trabajo (proteínas, rutas de plegado y trayectorias de plegado); y los métodos utilizados para los diferentes análisis estadísticos realizados sobre estos datos. Los detalles específicos del tratamiento de datos y la aplicación de los métodos se describen en los capítulos correspondientes en los que se utilizan.

4.1. Datos de plegamiento de proteínas

Se han aplicado varios enfoques computacionales para simular el plegado de proteínas y el movimiento molecular al plegado de proteínas[78, 49], entre los que se destacan aquellos que ofrecen una representación a nivel atómico como Dinámica Molecular (DM). En nuestra investigación usamos un conjunto de datos de plegamiento del supercomputador Anton, como se detallará más adelante. Cabe señalar que la recopilación de datos de este tipo de simulaciones es difícil, ya que las trayectorias y caminos resultantes no se publican comúnmente, y la gran cantidad de datos generados en estas simulaciones es difícil de almacenar y manejar.

4.1.1. Conjunto de Datos de Anton

Anton es un supercomputador especializado que acelera enormemente la ejecución de simulaciones de dinámica molecular. Además de las adaptaciones en hardware, el grupo que desarrolló dicho computador también modificó el campo de fuerza CHARMM para usarlo en dichas simulaciones[64].

De simulaciones realizadas en el super computador Anton obtuvimos una serie de trayectorias de simulaciones de proteínas. Estas trayectorias correponden a una secuencia de estados intermedios que la proteína sigue para pasar del estado no plegado al estado plegado o nativo. Una proteína puede tener más de un camino o ruta de plegamiento, siendo cada ruta una serie de estructuras 3D o conformaciones (snapshots) de la proteína.

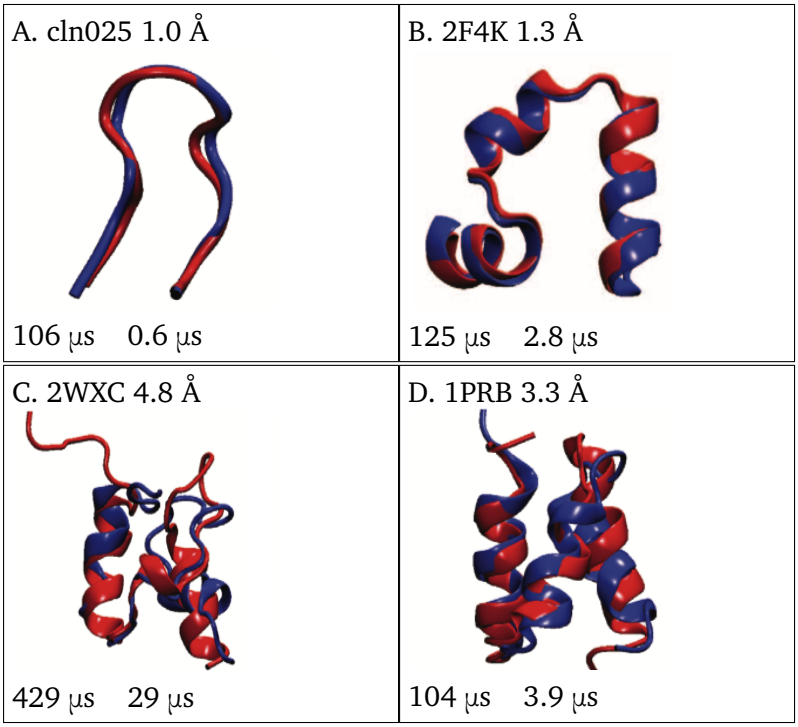


Figura 4.1: **Proteínas seleccionadas del conjunto de datos de Anton.** Estructuras representativas del estado plegado en algunas simulaciones de plegamiento de Anton. Para cada proteína se muestra la estructura obtenida en la simulación (azul) superimpuesta en la determinada experimentalmente (rojo). Se muestra también la entrada PDB de la estructura experimental, el RMSD entre las dos estructuras, el tiempo total de la simulación y el tiempo de plegamiento (tiempo promedio que permaneció en estado no plegado).

4.2. Evaluación de propiedades

Para la implementación de los métodos de cálculo de las propiedades, desarrollamos un kit de herramientas propio para el análisis de proteínas. El kit de herramientas ofrece funciones Python para cada propiedad que se pueden utilizar como programas independientes o se pueden llamar como bibliotecas desde programas externos. Muchas de estas funciones se implementan directamente en un lenguaje de computación y otras envuelven la funcionalidad de programas, bibliotecas y frameworks bien conocidos que ofrecen la funcionalidad indirectamente a través de procesos más complejos de los que nuestra función extrae la información necesaria para devolver el resultado apropiado.

Dado que la evaluación de las propiedades se realiza para muchas conformaciones de proteínas, se trata de una tarea que lleva mucho tiempo cuando se ejecuta en sistemas de uniprocesador como los que están disponibles en la mayoría de nuestros laboratorios de computación. Luego desarrollamos un framework distribuido propio que ejecuta distributivamente las evaluaciones reduciendo considerablemente los tiempos totales de ejecución. El framework se basa en el enfoque de editor/suscriptor de la informática distribuida y utiliza como middleware los servicios de almacenamiento en nube de dropbox [31] para comunicar cualquier PC de escritorio o servidor dentro o fuera de nuestra institución.

4.3. Preparación de datos

Todas las conformaciones de proteínas se minimizaron energéticamente para eliminar las restricciones estéricas mediante un algoritmo de minimización de gradiente conjugado del paquete de Gromacs. [103]). Las evaluaciones realizadas tanto en el análisis preliminar como en la definición de los componentes se escalaron para utilizar el mismo rango de valores para todas las propiedades mediante una normalización *min-max*. [48], unificando de 0 a 1 todas las propiedades. A continuación, fueron promediados por un filtro móvil medio modificado a valores de ruido suave que pueden resultar de la evaluación de conformaciones con eventos de plegamiento brusco que producen propiedades proteicas extremas.

En la definición de los componentes, el conjunto completo de las rutas se dividió en dos conjuntos de datos: uno para la formación y otro para la realización de pruebas. El conjunto de entrenamiento fue conformado por 5399 conformaciones de 29 rutas (80 %); se utilizó para filtrar y definir propie-

dades finales, componentes, grupos, distribuciones, y los parámetros para la definición de una medida de similitud. El conjunto de pruebas fue conformado por 2610 conformaciones de 7 rutas (20 %) y se utilizó para asignar el nivel de plegado utilizando la medida de similitud con sus conformaciones y analizar los resultados.

4.4. Análisis Estadístico

4.4.1. Escala Multidimensional

Se realizaron dos análisis clásicos (métricos) de Escala Multidimensional (MDS) (descritos a continuación) con datos de las rutas de Anton evaluadas. El primer MDS se ejecutó para seleccionar las propiedades más relevantes. Y el segundo MDS se llevó a cabo para encontrar una estructura preliminar de propiedades interrelacionadas. Ambos MDSs utilizaron como entrada una matriz de datos de las rutas evaluadas de Anton, la primera con las 16 propiedades y la segunda con un subconjunto de 11 de ellas. Además, utilizaban las diferencias entre propiedades calculadas con una distancia euclídea para aproximar un conjunto de puntos que representaban las propiedades de las proteínas a distancias entre puntos. Los puntos resultantes se graficaron en un gráfico 2D (XY) a partir del cual se interpreta la solución. Usamos para ambos MDSs la función *cmdscale* del paquete de estadísticas R.

La Escala Multidimensional (MDS) es una técnica para el análisis exploratorio de datos que representa las relaciones entre objetos (datos de similitud o disimilitud) como distancias entre pares de puntos en un espacio de dimensiones más bajas, haciendo que los datos sean accesibles para la inspección visual y la exploración[13]. Dos técnicas MDS se diferencian por la forma en que se generan los datos de proximidad (distancias entre puntos): la métrica clásica y la no métrica. Si muestra propiedades métricas, como distancias (por ejemplo, distancia euclídea), pertenece a la primera; pero si sólo asume que el orden de las proximidades es significativo (por ejemplo, orden de rango), entonces pertenece a la segunda[107].

4.4.2. Análisis de Componentes Principales

Realizamos tres análisis de componentes principales (PCA). El primero con 9 propiedades evaluadas sobre los datos de la trayectoria de la proteína de la villina en la cabeza. La segunda y tercera para definir una medida de similitud con 11 propiedades evaluadas en los datos de las rutas de PMR. En los tres PCA utilizamos un método general para realizar PCA llamado

descomposición espectral o *eigendecomposition*. Este método examina las covarianzas y correlaciones entre variables, e identifica componentes independientes que explican la cantidad máxima de correlación mutua (variación) que se mide por su *eigenvalue* [15] que será mayor para el primer componente y menor para los siguientes. El análisis de componentes principales (PCA) es una técnica estadística exploratoria para la reducción de variables que permite la identificación de grupos de variables interrelacionadas a través de una estructura subyacente que no es directamente observable. [40, 15]. El PCA reduce la dimensionalidad de un conjunto de observaciones sobre n variables en r nuevos ($r < n$) llamados *componentes* principales, que explican la mayor parte posible de la varianza en las *n* variables originales [97]. Los pasos generales para llevar a cabo el PCA son: una extracción inicial de componentes; determinación del número de componentes; rotación a una solución final; interpretación de la solución rotada y cálculo de las puntuaciones de los factores.

Utilizamos tres funciones diferentes para ejecutar el PCA desde el sistema R. En la primera, utilizamos la función *princomp*, ya que es simple y no rota componentes. Para el segundo, utilizamos la función *principal* del paquete *psych* para obtener una solución rotativa y seleccionar el número de componentes; y para el PCA final implementamos una función propia, basada en eigenvalues, que nos da más información que las otras y hace cálculos adicionales.

4.4.3. Selección y rotación de componentes

En todos nuestros análisis PCA seguimos una combinación de tres métodos para decidir cuántos componentes serán retenidos (seleccionados) para explicar en la medida de lo posible la información en los datos originales: criterio de valor propio - un criterio, porcentaje de varianza acumulada y regla de factores no triviales. El primero propone seleccionar el componente cuando su valor propio es superior a 1,0. En el segundo, sólo se conservan los primeros N componentes si son capaces de explicar un porcentaje específico de la varianza de los datos (normalmente se recomienda un 70 % u 80 %). Y la tercera define un criterio de agrupación en el que se eliminan los componentes que contienen una sola variable, dado que no contribuyen al objetivo de encontrar patrones de agrupación.

Por otro lado, después de seleccionar el número de componentes, utilizamos una rotación ortogonal de Varimax para hacer la solución más interpretable.[26, 21]. Una rotación se refiere a realizar aritmética para obtener un nuevo conjunto de cargas factoriales a partir de un conjunto dado, correspondiendo las

cargas factoriales a los pesos (correlaciones) entre las variables originales y los nuevos componentes.

4.4.4. Clustering Jerárquico

Realizamos un clustering jerárquico (HC) sobre las conformaciones de las rutas de plegado de proteínas del servidor Anton para buscar grupos que compartan características de plegado similares de estabilidad, compacidad y semejanza con el nativo. Un método de enlace completo basado en una distancia euclídea cuadrada para fusionar dos *clusters* [11] fue usando como estrategia de *clustering*. Selecciona la distancia más larga posible entre *clusters* (miembros más disímiles) para asegurar que más conformaciones en un *cluster* estén dentro de una distancia máxima entre sí, produciendo *clusters* bien distribuidos.

El criterio del coeficiente de inconsistencia se utilizó para determinar el número de clusters cortando el dendrograma a la altura de 1.0 y resultando en 4 clusters. Este coeficiente se calcula comparando la altura de cada eslabón de un clúster con las alturas medias de los enlaces vecinos que se encuentran debajo de él. El valor del coeficiente se selecciona para la mayor inconsistencia de un enlace, conectando los clusters más disímiles.

La agrupación jerárquica funciona con un enfoque ascendente que comienza con cada observación como una agrupación única distinta y fusiona sucesivamente las dos más cercanas en cada nivel hasta que sólo queda una agrupación.[11]. Un árbol binario, llamado dendrograma, se forma en cada paso a medida que dos clústeres se fusionan en uno nuevo. Las hojas del dendrograma (observaciones individuales) tienen alturas o nivel 0, y los nodos internos que resultan después de fusionar dos clústeres tienen una altura igual a la distancia de su intercluster. Existen cuatro estrategias comunes de agrupamiento (o vínculos) para medir esta distancia: único, toma la distancia mínima; completo, la distancia más grande; promedio, la distancia promedio; y centroide, toma los centroides de los clústeres.

4.4.5. Validación de Cluster

Aunque no existe un proceso o método general para validar las soluciones de clustering, se utilizan comúnmente tres enfoques principales: validaciones externas, internas y relativas (ver abajo). En nuestro trabajo, la solución de clustering fue validada tanto por un enfoque interno como por otro relativo (ver abajo). La validación interna utiliza sólo características inherentes al conjunto de datos para medir la calidad de la solución, como

la cohesión y la separación de los grupos obtenidos. Y la validación relativa compara la estructura de clúster con otras estructuras de clúster que resultan con el mismo algoritmo pero con parámetros diferentes. [94]. Para el primero, utilizamos el índice *Silhouette*; y para el segundo, un método de remuestreo.

No se utilizó la validación externa ya que necesita un conocimiento predefinido, como por ejemplo las etiquetas de clase correctas, y nuestra agrupación siguió un enfoque no supervisado sin ningún conocimiento biológico previo, específicamente nuestro proceso pretende definir el nivel de plegado de las conformaciones de una proteína, no asignar niveles conocidos.

4.4.5.1. Índice Silhouette

El índice Silhouette Si se calculó para cada grupo promediando el Si de cada una de sus conformaciones de proteínas. Al final, el promedio Si se calculó para el clustering completa promediando the groups de Si. El índice Si para la pconformación de proteína i fue calculada como $Si = (bi - ai)/max(ai, bi)$, donde ai es la distancia media entre la i-ésima conformación y las otras conformaciones en el mismo clúster que i, y bies la distancia media mínima entre la i-ésima conformación y las conformaciones en un clúster diferente. [69]. El índice de Silhouette combina ideas de cohesión (cuán estrechamente relacionados están los objetos en un clúster) y de separación (cuán distinto es un clúster de los demás). El valor de este índice se mide comparando la cercanía del objeto con los objetos de su propio clúster en comparación con los objetos de otros clústeres. [88]. El índice oscila entre 1 y -1, con valores cercanos a 1 para una correcta clasificación, cercanos a -1 mala clasificación y cercanos a cero, lo que indica que los objetos pueden ser clasificados igualmente en cualquier otro grupo.

4.4.5.2. Validación por remuestreo

Esta validación de conglomerados fue adaptada del enfoque de remuestreo propuesto por[58] en el que el criterio básico se basa en el hecho de que si un clustering es válido, entonces cada uno de sus subconjuntos también debe ser válido. [5]. En general, para la estimación no supervisada de la validez del clúster, se vuelven a muestrear los datos disponibles y se utiliza una cifra de mérito para medir la estabilidad de la solución frente a la dada por las soluciones de remuestreo.

En nuestra validación de clúster, utilizamos el conjunto completo de conformaciones de rutas como un conjunto de datos completo que se dividió en

varios subconjuntos. El procedimiento de clustering descrito anteriormente se aplicó tanto al conjunto completo como a cada subconjunto de datos, y para comparar las soluciones, creamos su respectiva matriz de conectividad, que consistía en una matriz simétrica conformada por las conformaciones de cada conjunto de datos. Las matrices estaban marcadas con 1 y 0; 1 si la conformación de la columna *i* estaba en el mismo grupo que la de la fila *j*; 0 en caso contrario. Luego, se comparó la matriz del conjunto completo de datos con cada una de las matrices de los subconjuntos, obteniendo una medida de concordancia entre las soluciones de clústeres, con un rango de 0 a 1 (consenso de 0 % a 100 %). Finalmente, los resultados individuales se promediaron para obtener una media global de consenso.

Capítulo 5

Algoritmo Rápido de Reducción de Trajectorias

5.1. Antecedentes

Para la reducción de trayectorias de simulación de plegamiento se han utilizado varios métodos como los basados en agrupamientos, basados en transformaciones lineales como el análisis de componentes principales (PCA) y el escalamiento multidimensional (MDS), y los que cambia la representación de la estructura como los basados en mapas de contactos.

Los basados en agrupamientos de estructuras se implementa en varias herramientas de simulación de plegamiento como el algoritmo de agrupamiento de GROMACS [25] donde toman todas las estructuras, miden la distancia entre ellas, toman como representativa la que más vecinos tenga de acuerdo a un valor de corte *(cutoff),* la eliminan junto a sus vecinos, y repiten el prtooceso para las restantes. Sin embargo, este tipo de algoritmos generalmente dependen de distintos parámetros tales como la especificación inicial del rádio o número de grupos, o la medida de similaridad para comparar las estructuras. Estos parámetros tienden a hacer artificial el agrupamiento donde los cambios del valor de alguno de los parámetros, pueden producir resultados que varian de forma considerable.

Entre los basados en transformaciones tanto lineales como no-lineales están los que usan PCA y MDS. Los que usan PCA [32] transforman la estructura de la proteína desde un espacio N-dimensional–dado por los puntos de datos de las N coordenadas de sus átomos–a un espacio lineal K-dimensional ($K < D$) que corresponde a un nuevo sistema de coordenadas llamado componentes principales. Estos componentes representan los vectores tangentes que describen un hiperplano que pasa a través de los puntos de datos tanto como sea posible cuando se evalúan sus mínimos cuadrados. Los componentes se ordenan de acuerdo a su varianza y los primeros (los de mayor varianza) son los que resumen mejor los cambios conformacionales globales de la proteína. Sin embargo PCA tiene problemas cuando los espacios son

no-lineales, como se piensa que es el espacio conformacional de la proteína y por lo tanto el nuevo espacio K-dimensional puede resultar distorsionado [24].

Para evitar el problema de la linearidad con el PCA, Rajan et al. 2010 [84] adaptan un método de escalamiento multidimensional no métrico (*nMDS)* para obtener una representación reducida 2D de toda la trayectoria. Inicialmente transforman la estructura 3D de la proteina a sus respectivos ángulos dihédricos para luego aplicarles el método de escalamiento y obtener un conjunto de puntos que representa las estructuras de proteínas. Estos puntos se despliegan sobre un espacio métrico (generalmente 2D) que representa la trayectoria de tal manera que la distancia cada par de puntos *x,y* es consistente con las distancias de cada par de estructuras *X,Y* representadas por los respectivos puntos. Aunque esta forma de reducción simplifica a 2D las estructuras N dimensionales (N coordenadas XYZ de sus átomos), la información de la estructura se pierde y la reducción se vuelve específica para ciertos análisis como el de analizar visualmente la ocurrencia de eventos en el tiempo.

Entre los que cambian la representación de la estructura Yang et al. 2007 [109] transforman la estructura a un mapa de contacto (matriz 2D binaria) a través de lo que definen como SOAPs o patrones 2D no locales que se encuentran en los mapas de la estructura. Se encuentran los SOAPs de todas las estructuras, se los agrupa por SOAPs comúnes de acuerdo a una medida de distancia, y se obtienen los más frecuentes. Así, las principales partes de la estructura se representa con los SOAPs más frecuentes y las otras partes se eliminan, lo que lleva a una representación más concisa. Sin embargo, la reducción cambia sustancialmente los elementos de la trayectoria al trasladar las estructuras 3D a representaciones 2D, perdiendo información implicita en la estructura.

5.1.1. Simulaciones de Plegamiento

Las simulaciones del plegamiento de proteínas son complejas y demandan gran cantidad de tiempo y recursos computacionales. Debido a estas limitaciones tecnológicas, hasta hace unos años estas simulaciones se realizaban para proteínas pequeñas y los tiempos simulados eran muy cortos, en el orden de los microsegundos mientras que una proteína se pliega en el orden de los milisegundos [1]. Sin embargo, en los últimos años los avances en el hardware han logrado avances de tal manera que se empiezan a mostrar resultados de simulaciones mucho más largas y de proteínas más grandes. Dos ejemplos de estos avances son los proyectos de folding@home y de la

supercomputadora Anton. El proyecto foldin@home logró realizar hace algunos años una de las primeras simulaciones largas utilizando computación distribuida. Una de sus simulaciones alcanzó el orden de los microsegundos para plegar completamente una proteína pequeña, la Villin Headpiece de 36 residuos [67]. Más recientemente, la supercomupatadora Anton ha usado computación paralela y hardware especializado para simular dinámica molecular [91]. Con está máquina se ha logrado plegar completamente varias proteínas medianas (10-80 residuos), alcanzando tiempos de simulación del orden de los milisegundos. En ambos proyectos los resultados de las trayectorias están disponibles para que la comunidad científica los descargue y los analice para avanzar en el entendimiento del plegamiento de las proteínas.

5.1.2. Algoritmos Rápidos de Agrupamiento de Secuencias

Muchos de los algoritmos para realizar agrupamientos rápidos de secuencias biológicas se basan en las ideas del algoritmo de Hobohm y Sander [43] que creó inicialmente para agrupar de forma rápida secuencias de proteínas. El algoritmo determina las secuencias más representativas a través de dos actividades: un ordenamiento y una selección rápida. En el ordenamiento, las secuencias se organizan por longitud en orden descendiente, luego se toma la primera secuencia (la más larga) como representativa del primer grupo. En la selección rápida, se compara el resto de secuencias con la representativa y se las incorpora al grupo si son cercanas (ejemplo, si son similares a nivel de secuencias), de lo contrario, pasa a ser la representativa de un nuevo grupo y se hace lo mismo con el resto de secuencias hasta terminar. Los aspectos determinantes del éxito del algoritmo son la relación de orden que se establezca al inicio y las propiedades que se tomen para comparar las secuencias. En secuencias de ADN y de proteínas estos aspectos funcionan bien ya que dos secuencias de más o menos de igual longitud tienen mayor probabilidad de ser similares que dos secuencias de longitudes completamente diferentes. Sin embargo en estructuras tridimensionales de proteínas que pertenecen a una misma trayectoria, la longitud y la similaridad de la secuencia va a ser la misma para todas la conformaciones, lo que implica redefinir estos aspectos en términos de las características de las estructuras 3D de proteínas de una misma trayectoria, como vamos a describir más adelante cuando mostremos nuestro algoritmo de reducción de trayectorias de plegamiento.

Dos de las implementaciones más usadas de este algoritmo para agrupamiento rápido de secuencias son los programas CD-HIT [60] y UCLUST

[35]. El programa CD-HIT realiza un ordenamiento por longitud de la secuencia como lo plantea el algoritmo de Hobhon, y para la selección utiliza un filtro de palabras cortas para comparar si dos secuencias son similares–evitando el alineamiento de las mismas–y así asignarlas a un mismo grupo o crear uno nuevo. En el caso de secuencias de proteínas el programa usa por defecto una palabra de 10 aminoácidos o *decapeptido*. En cambio el programa UCLUST utiliza para com parar las secuencias una función creada por los mismos autores que la llaman como USEARCH y que calcula la similitud entre las secuencias a partir de un alineamiento global.

5.2. Algoritmo de Reducción de Trayectorias de Plegamiento

La primera parte del algoritmo realiza un agrupamiento local rápido donde se aprovecha el ordenamiento temporal de las conformaciones implícito en la trayectoria. Pare esto, se toma la idea del algoritmo propuesto por Hobohm et al. [1] para la selección de conjuntos de proteínas. Se particiona la trayectoria en *M bins* o secciones de *N* conformaciones contiguas en el tiempo de simulación. Para cada uno de los *bins* se toma la primera estructura como cabeza del primer grupo y se la compara con la siguiente en orden de tiempo de simulación. Si presentan similaridad se adicionan al grupo; de lo contrario si es disimilar se crea un nuevo grupo y se toma a esta última estructura como cabeza del nuevo grupo. El proceso continua hasta terminar con todas las estructuras del *bin* y esto mismo se realiza para los demás *bins*. En la segunda parte del algoritmo, toma cada conjunto de conformaciones cabeza de grupo seleccionadas en cada *bin* y se crea una matriz de similaridades que se la usa para realizar un agrupamiento para seleccionar las *K* estructuras más representativas de cada conjunto tomando los *k-medoides*. La unión de estas *K* estructuras por *bin* crea un nuevo conjunto mucho más reducido que el creado en el agrupamiento local. El orden temporal no se pierde ya que las *K* estructuras seleccionadas por cada conjunto se las ordena de acuerdo a su tiempo original de simulación.

5.3. Datos y Métodos

5.3.1. Comparación de Estructuras de Proteínas

Tanto para la primera y segunda fase de reducción utilizamos la medida de similitud entre estructuras de proteína llamada TM-score (Template Mo-

deling score) [111]. Esta medida de similitud a diferencia de otras medidas ampliamente usadas en comparación de estructuras como el RMSD (Root Mean Square-Deviation) es más precisa ya que en el TM-score influyen poco sobre el puntaje final las secciones pequeñas de la proteína que alinean incorrectamente, tales como giros simples o términos flexibles, lo que reduce el chance de evaluaciones sesgadas.

5.3.2. Selección de Estructuras Representativas

Las estructuras representativas de cada grupo o *bin* resultante de la selección rápida de la primera fase se obtienen aplicando en cada uno de ellos un algoritmo de particionamiento alrededor de medoides (PAM) [80]. El algoritmo selecciona como representativa la estructura media o central de cada subgrupo resultante para la cual la suma de las distancias entre esta y las demás estructuras del subgrupo es mínima. Así, al final se obtienen por cada grupo o *bin* inicial un conjunto reducido de estructuras que representan los eventos principales de esta sección de la trayectoria de plegamiento.

5.3.3. Trayectorias de Plegamiento de Proteínas

Para mostrar los resultados del algoritmo de reducción propuesto, aplicamos las reducciones a tres trayectorias de plegamiento de proteínas. La dos primeras corresponden a trayectorias cortas (200-300 conformaciones) para las proteínas: ferredoxina desde clostridium acidurici (PDB: 1FCA) y del Cyt férrico de levadura (iso-1-Cytc, PDB: 2YCC) que fueron simuladas por el grupo de Amato mediante el método *Probabilistic Roadmap Method [96]* y que se caracterizan por ser trayectorias cortas que tratan de incluir los eventos principales de la simulación. Por el contrario, la tercera trayectoria corresponde a la simulación de plegamiento mediante la técnica de Dinámica Molecular [63] para la proteína Trp-cage (PDB: 2JOF) y se caracteriza por ser una trayectoria mucho más extensa y detallada (más de 1 millón de conformaciones).

5.4. Detalles de Implementación

5.4.1. Descripción de Programas

El algoritmo está implementado a través de tres scripts:

- pr00_main.py: Script principal en lenguaje Python que toma los parámetros iniciales y llama a los otros scripts enviándoles los parámetros necesarios.
- pr01_createBins.py: Script en lenguaje Python que realiza las particiones
- pr02_localReduction.R : Script en lenguaje R que realiza la reducción local.
- pr03_globalReduction.R: Script en lenguaje R que realiza la reducción global.

5.4.2. Ejecución

La ejecución se realiza llamando al script *reduction.py* así:

```
$ ./reduction.py <InputDir> <Ouputdir> <BinSize> <Threshold>
                        <K> <nCores>
```

Donde:

- **Input Dir:** Nombre del directorio de entrada donde están las conformaciones de la trayectoría de la proteína a reducir.
- **Output Dir:** Nombre del directorio donde quedarán los resultados de la reducción. Si el directorio ya existe lo renombra automáticamente y crea uno nuevo. Dentro del directorio se crean cuatro subdirectorios:
  - bins: donde se crean las particiones con las conformaciones correspondientes a cada bin
  - binsLocal: donde se crean las nuevas particiones con los resultados de la reducción local
  - pdbsLocal donde se copian todas las conformaciones de la nueva trayectoria producto de la reducción local.
  - pdbsGlobal: donde se copian todas las conformaciones de la nueva trayectoria producto de la reducción global.
  - tmp: donde se coloca los archivos temporales resultantes de la creación de las matrices de distancia con TM-score
- **Bin Size:** El tamaño de conformaciones por partición o *bin*. El algoritmo crea el número de *bins* dependiendo del tamaño de la trayectoria.

- **Threshold:** Umbral usado por el TM-score para comparar dos conformaciones y decidir si son similares.
- **K:** Número de conformaciones a seleccionar por el agrupamiento global
- **nCores:** Número de *cores* a utilizar para el procesamiento en paralelo.

5.4.3. Requisitos

Los programas están en python y en R. Del sistema R se necesita instalar las librería para agrupamientos y paralelización: *cluster* y *parallel*, respectivamente.

5.5. Resultados y Discusión

Para mostrar las reducciones que realiza nuestro algoritmo, presentamos aquí los resultados de la reducción realizada a tres trayectorias de proteínas. Las dos primeras son trayectorias cortas de menos de 300 conformaciones, mientras que la tercera es mucho mas larga con más de 1 millón de conformaciones.

En la Figura 5.1 mostramos las reducciones de las dos trayectorias cortas correspondientes a las proteínas 1FCA1 y 2YCC (ver sección 5.3.3). En la parte superior está la trayectoria original completa; en la parte intermedia la trayectoria después de la reducción local; y en la parte inferior la trayectoria final después de la reducción global. Las reducciones logradas son del orden de más del 76 % para la proteína 1FCA1 (de 239 a 57 conformaciones) y más del 90 % para la proteína 2YCC (de 268 a 26 conformaciones). Observamos que los eventos principales en ambas trayectorias se conservan claramente (recuadros rojos en las trayectorias original y final) lo que prueba visualmente que nuestro algoritmo realiza reducciones que reflejan la dinámica de la trayectoria. Además, destaquemos que en la primera reducción, la local (figura intermedia), los eventos principales tienden a desplazarse frente a los originales (recuadros azules), lo cual se logra después corregir en la reducción final. Esto se debe a que la reducción local por ser rápida incluye conformaciones tanto de eventos principales como de eventos secundarios, mientras que la global se enfoca en dejar solo los eventos principales y por lo tanto el desplazamiento se reduce, lo cual va a ser más evidente en el caso de la trayectoria larga descrita a continuación.

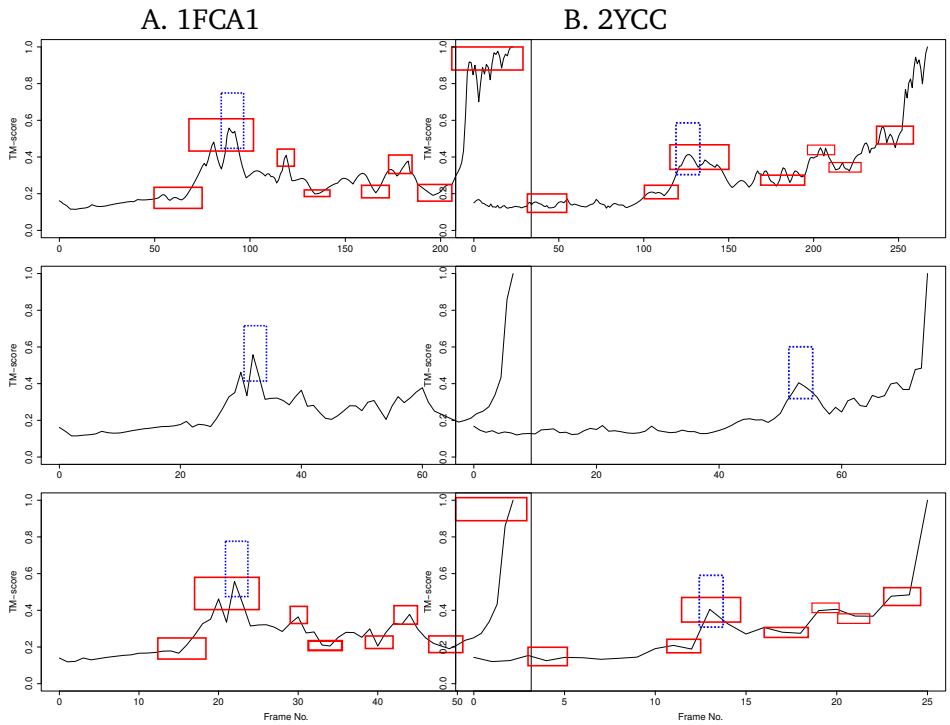


Figura 5.1: **Reducción de las trayectorias cortas de plegamiento para las proteínas 1FCA1 y 2YCC.** En recuadros rojos se resaltan los eventos principales que se conservan tanto en la trayectoria original como en la final. Los recuadros rojos muestran como algunos eventos principales se desplazan en la reducción local, pero logran ajustarse al final en la reducción global. Para la proteína 1FCA1 la reducción se realizó con los parámetros de 40 bins, un umbral de TMscore de 0.5 y un K de 10. Mientra que para la proteína 2YCC se usaron 50 bins, un TMscore de 0.5 y un K de 5

Ahora, en la Figura 5.2 observamos la reducción hecha sobre una trayectoria larga de más de 1 millón de conformaciones para la proteína 2FOF (ver sección 5.3.3). La reducción final fue de más del 97 % (de 1044004 a 20883 conformaciones). Observamos que a pesar de que la simulación presenta bastantes oscilaciones en el plegamiento, en general los eventos principales al final de la reducción global se conservan. Es importante notar aquí que la reducción local no describe claramente los eventos principales, como lo destacamos en las reducciones anteriores, sin embargo la reducción global que toma los datos de la local, logra destacarlos cuando selecciona las conformaciones más representativas de cada partición.

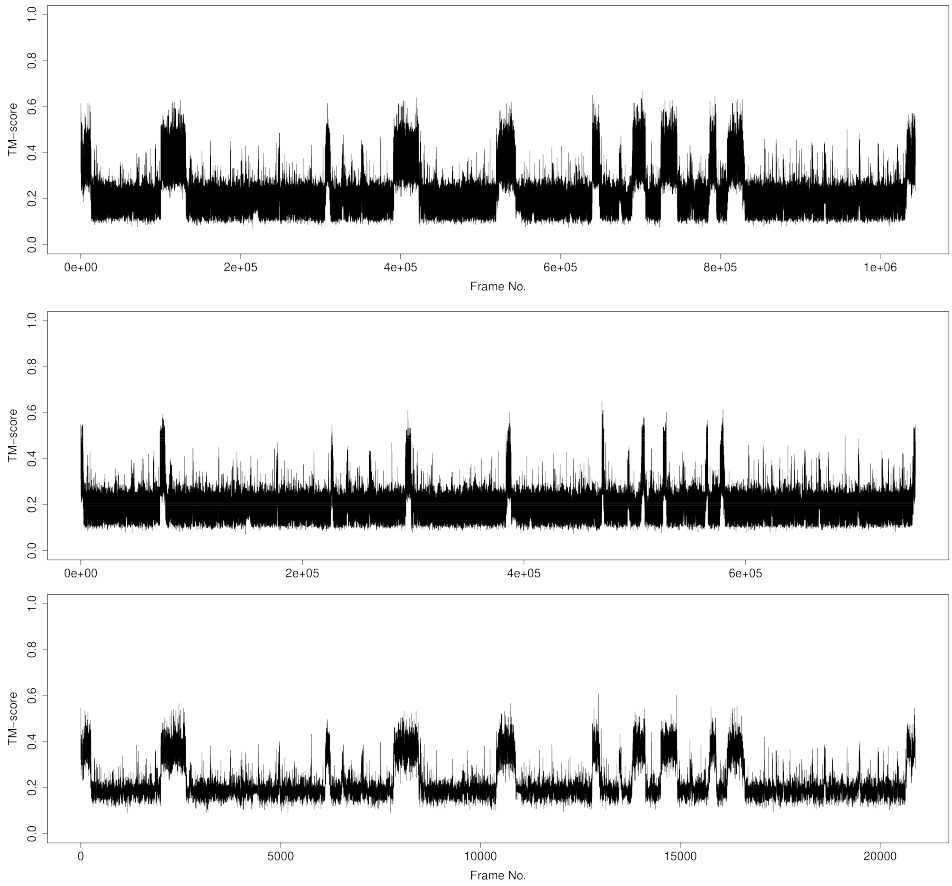


Figura 5.2: Reducción de una trayectoria larga de plegamiento.

Capítulo 6

Aplicación Virtual

En trabajo de rasgos conformacionales propone una serie de propiedades fisicas y estructurales las cuales se evalúan en las proteínas, siendo estos cálculos el insumo para calcular el puntaje de plegamiento de la misma, y con esto, su clasificación en uno de los estados de plegamiento. Dichas herramientas se encuentran escritas en distintos lenguajes de programación, tales como C, C++, python, R, etc. Todas estas funciones de evaluación de propieddes se encuentran envueltas en un programa python, estableciendo así un API que oculta las implementaciones particulares para estas evaluaciones. Esta diversidad de lenguajes hace que sea complejo la distribución de estas herramientas, por ejemplo por el hecho de tener que compilar para la plataforma en la que se desea trabajar. Por esto, y con el objetivo adicional de incorporar en el pipeline la reducción de proteínas, proponemos el uso de una aplicación virtual que empaquete la funcionalidad completa.

Aplicaciones Virtuales

Una aplicación virtual (virtual appliance) es un sistema pre-integrado y auto contenido que combina una o más aplicaciones de software con un sistema operativo reducido o ajustado (en este caso, una distribución liviana de Linux) para que la aplicación se ejecute de forma óptima sobre un hardware estándar (por ejemplo, máquinas con procesadores Intel) o se ejecute sobreuna máquina virtual (por ejemplo, VirtualBox[2], VMware[3], Xen[4], entre otros) [27].

Los beneficios de las aplicaciones virtuales son muchos, entre ellos la facilidadde instalación y uso a través del empaquetamiento de la aplicación con el software necesario para su uso, sin necesidad de recompilar o instalar externamente nuevas librerías, compiladores u otras utilidades.
Otro beneficio es la universalidad que les da a las aplicaciones, pudiéndose ejecutar en cualquier plataforma que maneje máquinas virtuales. Adicionalmente, si junto con la aplicación se empaquetan las herramientas necesarias

para desarrollo (compiladores, librerías y utilidades), entonces se brinda la opción de poder hacer nuevos desarrollos sobre la aplicación. Esto es muy ventajoso desde el punto devista de software libre y abierto, ya que facilita a los desarrolladores todas las herramientas necesarias para seguir aumentado y mejorando la aplicación.

La aplicación

La aplicación virtual está constituida por los siguientes elementos:

- Código fuente del pipeline. Código en C/C++, Perl y shell scripts, junto con otras herramientas propias para su ejecución, tales como scripts en el lenguaje Perl.
- Sistema operativo mínimo que sirve como plataforma para la ejecución y compilación del software. El sistema operativo elegido es la versión mínima de Ubuntu 8.04, conocido como JEOS 8.0425 (Just Enough Operating System).
- Documentación del producto.
- Software externo. Incluye librerías especiales usadas para las distintas implementaciones de los modelos de optimización. Entre estas, están las librerías R para clustering jerárquico bio3d.

Capítulo 7

Software Toolkit para análisis de trayectorias de plegamiento de proteínas

Este capítulo describe un toolkit que fue desarrollado para evaluar 16 propiedades estructurales y energéticas de proteínas seleccionadas de la literatura y otras de desarrollo propio. Las propiedades serán usadas en los próximos capítulos para lograr una representación de una proteína en términos de *características de plegamiento*. Empezamos con la presentación de una descripción de las propiedades mostrando su definición y su importancia asociada con características de pleglamiento; y luego presentamos el toolkit que desarrollamos con los algoritmos y herramientas usadas para calcular sus valores.

7.1. Propiedades de proteínas

Las propiedades descritas en la tabla 7.1 han sido usadas en otros estudios para caracterizar diferentes aspectos del proceso de plegamiento de proteínas, y han sido asociadas con algunas características de plegamiento no observables directamente a partir de la estructura de la proteína.

Nombre	Id	Descripción	Asociación
Contactos Nativos	NC	Contactos formados por una conformación de proteína que también están presentes en la estructura nativa.	Grado de cercanía con la estructura nativa. Cantidad de estructura nativa.
Orden de Contactos	CO	Separación promedio de secuencias entre residuos en contacto en la estrucutra nativa.	Propiedades cinéticas del plegamiento de proteínas.
Radio de Giración	RG	Cuánto se esparce la proteína de su centro.	Nivel de compactación de la estructura.
Enlaces de Hidrógeno	HB	Número de enlaces de hidrógeno formados por la estructura.	Formación y estabilización.
Área de Superficie Accesible.	AS	Superficie accesible para el solvente.	Compactación y estabilidad conformacional.
Vacíos	VD	Espacios sin llenar dentro de una proteína que no son accesibles desde el solvente.	Empaquetado y estabilidad de la proteína.
RMSD	RM	Distancia promedio en posiciones entre los atomos de dos proteínas.	Similitud de las proteínas.
RMSD Local	LR	Promedio del RMSD entre los elementos de la estructura secundaria (desarrollado por Garreta).	Topología de la proteína y similitud estructural.
Energía Potencial	PE	Energía asociada con las fuerzas de atracción y repulsión entre los átomos de la proteína.	Estabilidad termodinámica.
Momento Dipolar	DM	El producto de la magnitud de las cargas y su distancia de separación en una proteína.	Forma de la proteína- Propiedades eléctricas.
Residuos en Estructuras Secundarias Correctas	RC	Residuos formando la misma estructura secundaria que la conformación nativa (desarrollado por Garreta).	Similitud de estructuras entre proteínas.
Residuos en cualquier estructura secundaria	RA	Residuos en la proteína objetivo formando cualquier elemento de estructura secundaria (desarrollado por Garreta).	Similitud de estructuras entre proteínas.
Puntaje Estructural	SS	Residuos que forman “correctamente” motivos locales cortos o bloques de proteínas (desarrollado por Garreta).	Similitud de estructuras entre proteínas.
Clusters Rígidos	CL	Grupos de átomos acoplados uno al otro a través de enlaces rígidos.	Estabilidad mecánica de una proteína.
Grados de libertad	DF	Relacionada con el numero de restricciones necesarias para ha..Related with the number of constraints necessary hacer rígida la estructura de la proteína.	Estabilidad mecánica de una proteína.
Regiones estresadas	SR	Regiones con sobrecarca de restricciones que reducen el grado de libertat de una proteína.	Estabilidad mecánica de una proteína.

Cuadro 7.1: **Resumen de las propiedades seleccionadas para describir el plegamiento.** La tabla muestra el nombre de las propiedad, el id que será usado como nombre corto a lo largo de este documento, una descripción corta y su asociación con el plegamiento de proteínas.

7.1.1. Contactos Nativos

El número de contactos nativos de una conformación proteica puede ser considerado como una medida global de la "cantidad" de estructura que contiene respecto a su estructura nativa. [95]. ...Dos residuos de proteína distintos forman un contacto cuando la distancia máxima entre residuos de un par de átomos pesados (C, O, S, or N) está dentro de un valor umbral (ej. 6 o 7 Å). Otras restricciones pueden ser una distancia euclidiana mínima entre los residuos (e.g. 3 Å); una separación secuencial de al menos 3 residuos entre sí; o una distancia entre átomos pesados específicos como C-α or C-β.

El número de contactos nativos se ha utilizado para evaluar la "semejanza nativa" de las conformaciones (cuánta estructura tiene una conformación) en varios estudios relacionados con el plegamiento de proteínas como la construcción de rutas utilizando métodos probabilísticos de hoja de ruta.[96, 6]; el estudio de las primeras fases del plegado mediante simulaciones de dinámica molecular [34]; y en la construcción del camino usando eventos de desplegado [110].

Se calcula contando los contactos de la conformación que también están en la estructura nativa. Como se ha definido anteriormente, dos átomos de dos aminoácidos diferentes a_i y a_j están en contacto, si su distancia entre residuos $\delta(a_i, b_j) \leq \delta^{max}$, donde

$$\delta(a_i, b_j) = \sqrt{(x_i - x_j)^2 + y_i - y_j)^2 + (z_i - z_j)^2}$$

es la distancia euclidiana entre dos átomos(ej. carbono C-α o C-β) de los dos residuos diferentes con coordenadas 3D x_i, y_i, z_i y x_j, y_j, z_j; y δ^{max} es algún umbral de distancia máxima permitida (ej. 6 o 7 Å) [110].

7.1.2. Orden de Contacto

La complejidad topológica de una proteína está relacionada con el tipo de interacciones, locales o no locales, que impulsan su plegamiento.[81]. El orden de contacto (CO) se define como una medida del número relativo de interacciones locales en comparación con las interacciones no locales [45]. Las estructuras dominadas por interacciones locales como las proteínas altamente α helicoidales corresponden a bajo CO y se pliegan más rápidamente que las estructuras dominadas por interacciones de largo alcance -necesarias para su estabilidad–como α/β y β proteínas con CO más alto.

Esta medida ayuda a cuantificar la topología y la estabilidad de las proteínas en sus estados nativos.[92]. Se ha correlacionado bien con las propiedades de la cinética de plegado de proteínas como las tasas de plegado

y la colocación del estado de transición de proteínas. [81]. Además, se ha utilizado en programas de estructura de proteínas *ab initio* para ayudar a seleccionar los mejores candidatos para una etapa de refinamiento posterior. [12]. En proteínas con cinética simple -plegamiento en dos estados- ha descrito el tipo de interacciones proteicas [81, 45].

Se calcula a partir de las coordenadas de la estructura de una proteína de la siguiente manera[81]:

$$CO = \frac{1}{LN}\sum_{i,j}^{N} \triangle S_{i,j} \, |i - j|$$

Donde L es el número total de residuos en la proteína, N es el número total de contactos, y $\triangle S_{i,j}$ es la separación secuencial de residuos entre i y j.

7.1.3. Radio de giro

El radio de giro (RG) es una medida de la forma o dimensiones de la conformación de una proteína que describe su tamaño y compacidad midiendo cuánto se extiende la estructura de la proteína desde su centro. En el análisis de proteínas, la RG ayuda a inferir cuán plegada o desplegada está una proteína, ya que es indicativa del nivel de compactación de la estructura[65]. Ayuda a trazar los cambios de conformación de una proteína durante su trayectoria de plegado trazando su RG en función del tiempo. [33]. Además, se ha demostrado que las arquitecturas de las proteínas ($\alpha, \beta, \alpha/\beta,$ y$\alpha + \beta$) tienen un RG característico [65].

El radio de giro de una conformación se calcula como la distancia entre su centro de gravedad y sus extremos.[104]:

$$R_g = \sqrt{\left(\frac{\sum_i \|r_i\|^2 m_i}{\sum_i m_i}\right)}$$

donde m_i es la masa del átomo i y r_i la posición del átomo i con respecto al centro de masa de la molécula.

7.1.4. Enlaces de Hidrógeno

Los enlaces de hidrógeno (HB) juegan un papel central en el plegamiento de proteínas, influyendo en la forma final en 3D de una proteína al contribuir en la formación de sus estructuras secundarias y terciarias. Un enlace de hidrógeno es un tipo de fuerza atractiva que existe entre un hidrógeno unido

a un átomo electronegativo de una molécula y un átomo electronegativo de la misma molécula o de una molécula diferente. Los enlaces de hidrógeno participan tanto en la estabilización de hélices αcomo en la interacción entre hebras β para format hojas β, así que son parcialmente responsables de la estabilidad de la estructura final de la proteína plegada [106]. Varios estudios han utilizado el HB para evaluar la estabilidad [36] y para indicar cierta organización en proteínas [38]. Generalmente, este tipo de enlace es más fuerte que otras fuerzas intermoleculares como el van der Waals pero más débil que los enlaces iónicos y covalentes. [98].

Para determinar el número de enlaces de hidrógeno formados en una estructura proteica, se cuentan los pares de átomos -donantes y aceptantes- que están dentro de un valor umbral.[95]. El umbral puede variar de 2.6 a 3.6 Å, que es la distancia entre átomos pesados. [106].

7.1.5. Área de superficie Accesible

La superficie accesible (ASA) de una proteína contribuye a determinar su tamaño y forma. El ASA es un concepto que corresponde a la superficie de una proteína accesible al solvente.[57]. Desempeña un papel importante en varias propiedades de las proteínas como la compacidad, el plegamiento de proteínas y la estabilidad conformacional. [62]. El ASA se ha utilizado para caracterizar la compacidad de una estructura proteica definiendo una relación entre su ASA y la superficie de la esfera ideal del mismo volumen. [65].

La idea general para calcular el ASA es utilizar una "sonda", correspondiente a una esfera del disolvente, que se hace rodar sobre la superficie de la proteína explorando el área accesible por la sonda. [93]. El radio de la sonda corresponde al radio de una molécula del disolvente, por ejemplo 1.4 Å para una molécula de agua.

7.1.6. Vacíos

Los vacíos son espacios vacíos o cavidades vacías, dentro de una proteína, que no son accesibles al disolvente. Permanecen vacías aunque pueden ser lo suficientemente grandes como para ser llenadas por otros átomos o moléculas como el agua. Se han relacionado con la estabilidad de las proteínas, ya que al rellenar las cavidades -en la ingeniería de proteínas- en algunos tetrámeros se puede estabilizar su estructura en general[75]. Además, los vacíos están relacionados con la evaluación de la calidad de empaquetamiento de las proteínas. [61]. Menos cavidades de vacíos dentro de una

proteína implican que está más empaquetada, y las altas densidades de empaquetado en los núcleos de proteínas a menudo se consideran más como sólidos que como líquidos. [61].

Los vacíos se calculan utilizando un método de búsqueda de vacíos basado en cuadrículas que mapea una proteína en una cuadrícula 3D y asigna cada punto de cuadrícula a la proteína, al solvente o a la cavidad de acuerdo con su ubicación en la estructura de la proteína.[22]. Dos tipos de sondas con un radio pequeño (ej. 0.0 Å) y un radio grande (ej. 1.4 Å) se usan para detectar vacíos y delimitar las regiones accesibles con solventes, respectivamente. La sonda más grande delimita las regiones accesibles al disolvente, mientras que la más pequeña identifica los vacíos sin fusionarlos con el disolvente en bruto.

7.1.7. Error cuadrático Medio

El error cuadrático medio(*Root-Mean Square Deviation* en inglés, y más conocido como RMSD)refleja la distancia media en posiciones entre los átomos de dos proteínas superpuestas. Generalmente, estructuras similares tienen valores bajos de RMSD, en el orden de 1-3 Å, mientras que los valores RMSD más grandes corresponden a valores más disímiles [10].

El RMSD ha sido ampliamente utilizado como una medida del grado de similitud entre dos estructuras de proteínas. El experimento CASP ha utilizado este método -además de otras medidas- para evaluar los protocolos de predicción utilizados por los expertos para predecir la estructura 3D entre proteínas.[72].

Se calcula a partir de las coordenadas del objetivo i and una referencia j como sigue:

$$RMSD(i,j) = \sqrt{\frac{1}{N}\sum_{k=1}^{N}\left||r_k^{(i)} - r_k^{(j)}|\right|^2}$$

donde i y j son dos estructuras de proteínas óptimamente superpuestas de igual tamaño . N es el número de átomos en cada estructura, k es un índice sobre estos átomos, y $r_k^{(i)}, r_k^{(j)}$son las coordenadas cartesianas del átomo k en conformaciones i, j. El valor de RMSD resulta mínimo porque está determinado por la superposición de las estructuras que minimiza su distancia. En nuestro proyecto, el RMSD se utiliza siempre para comparar una determinada conformación proteica con la nativa.

7.1.8. Error cuadrático Medio Local (Local RMSD)

El citado RMSD es una medida global incapaz de evaluar las similitudes en las estructuras locales, y considerada como una "medida imperfecta" para comparar estructuras como conformaciones intermedias donde los elementos estructurales pueden estar presentes pero no todavía organizados espacialmente.[36, 74].

En su proyecto, Garreta desarrolló el *local root-mean square deviation* (LRMSD). Compara una conformación de referencia i de una proteína con una conformación objetivo *j*, utilizando el RMSD sólo en cada elemento de estructura secundaria (EES) de la estructura de referencia y promediando los valores obtenidos. Se relaciona con la topología de una proteína, y contribuye a medir la similitud estructural entre dos conformaciones de proteínas comparando localmente sus elementos estructurales secundarios.

Se calcula extrayendo las posiciones de los aminoácidos de los SSEs desde la primera conformación (la eferencia), usando el programa DSSP[50]. Las mismas posiciones se extraen de la segunda conformación (objetivo). Se instala cada par de EESs y se calcula su valor RMSD, finalmente se determina el promedio de todas las SSEs en la estructura de referencia:

$$LRMSD(i,j) = \frac{1}{S}\sum_{s=1}^{S} RMSD(i_s, j_s)$$

donde *i*, es la conformación de referencia y *j* la conformación objetivo . S es el número de EESs en la estrucutra referencia. $RMSD(i_s, j_s)$ es el valor RMSD entre el número EES s en la conformación de referencia *i* y los átomos correspondientes en la conformación del objetivo.
En este trabajo, la conformación de referencia es siempre la nativa.

7.1.9. La energía potencial

Si una proteína se toma como una molécula estática con muchos átomos interactuando, su energía potencial puede ser definida como la energía dada por su estructura tridimensional.

La energía potencial está relacionada con la estabilidad termodinámica, considerando que una proteína se pliega al estado de menor energía al disminuir su energía potencial, y alcanza la estabilidad termodinámica en su estado nativo.[30].

Si las interacciones entre los átomos de la proteína son consideradas como fuerzas elásticas, entonces pueden ser escritas en términos de ecuaciones

conocidas como funciones de energía potencial (o campos de fuerza) en las que la energía total es dada por la suma de términos enlazados y no enlazados que describen átomos enlazados por enlaces covalentes e interacciones de largo alcance, respectivamente. A continuación se presenta una función general de energía potencial; para funciones más específicas, véase[82]:

$$E_{total} = E_{enlaces} + E_{ángulos} + E_{torsiones} + E_{van-der-waals} + E_{electrostatica}$$

donde las primeras tres energías corresponden a términos enlazados que describen enlaces, ángulos y rotaciones de enlaces en la proteína. Y, los dos últimos términos describen interacciones entre átomos no ligados o átomos separados por 3 o más enlaces covalentes, la energía de interacción de van der Waals y la energía electrostática.

7.1.10. El momento dipolar

El momento dipolar puede definirse como el producto de la magnitud de la carga y la distancia de separación entre las cargas de una molécula.

El momento dipolar está relacionado con las estructuras geométricas y eléctricas de una molécula; y su magnitud puede afectar la forma de una proteína, su interacción intermolecular, la solución biomolecular y sus actividades bioquímicas.[112]. Su magnitud también puede utilizarse para distinguir entre moléculas polares y no polares. En las proteínas, el momento dipolar se calcula como la suma vectorial del momento dipolar de tres componentes: cargas fijas en la superficie, grupos polares en las cadenas principales y laterales y las fluctuaciones móviles de iones. [99].

7.1.11. Residuos en estructuras secundarias correctas o en cualquiera

Hélices alfa, hebras beta formando hojas beta, y otros elementos de estructura secundaria *(*EES*s)* son contribuyentes esenciales para la estabilización del pliegue proteínico general. Estas estructuras son segmentos de la cadena de proteínas en los que los ángulos de torsión del backbone ϕ y ψ se repiten en patrones regulares. Hemos creado dos medidas para cuantificar la proporción de EES*s* en una conformación de proteína. La primera, *residuos en estructura secundaria correcta,* evalúa la proporción de residuos de una conformación objetivo que coinciden con en su EES con el de la nativa. La segunta, *residuos en cualquier estructura secundaria* evalúa la proporción de residuos formando cualquier EES*s* en la conformación objetivo.

Ambas medidas se calculan obteniendo y comparando los EES*s* asignados por el programa DSSP[50] para las conformaciones de la proteína: objetivo y nativo para la medida correcta de SSEs, y sólo objetivo para cualquier medida de EES*s*.

7.1.12. Puntaje estructural

Esta medida fue desarrollada por Garreta y está relacionada con una tesis doctoral de nuestro grupo de investigación. Se basa en la similitud estructural entre proteínas, pero las comparaciones se realizan utilizando motivos locales cortos o bloques de proteínas (BP) [53] en lugar de EES*s*. La representación de las proteínas con BP proporciona más información estructural local que los EESs porque los BP resaltan las variaciones sutiles y las conservaciones estructurales de la estructura. [68, 53]. La estructura 3D de la proteína fue representada como una cadena unidimensional de BPs a partir de una biblioteca de 40 que fueron construidos utilizando fragmentos de 5 residuos que consisten en las coordenadas 3D de sus átomos C-α. Para su cálculo, se codifican ambas estructuras de proteínas: objetivo y nativa, transformando su representación 3D en una secuencia 1D de BPs. Luego, la proporción de BPs del objetivo que se calcula que también están en el nativo.

7.1.13. Clusters rígidos, regiones estresadas y grados de libertad

Estas medidas se toman del análisis de rigidez que se ocupa de analizar regiones rígidas y flexibles de una proteína.[47, 90]. *Clusters rígidos* son las partes de la proteína que se comportan como cuerpos rígidos (es decir, las distancias entre todos los átomos permanecen fijas). [20]. *Grados de libertad* está relacionado con el número de restricciones necesarias para rigidizar la estructura de la proteína [90]. Y *regiones estresadas* son regiones que contienen restricciones excesivas [90], como restricciones de distancia entre los átomos impuestas por las fuerzas de enlace, que reducen el número total de grados de libertad disponibles para la proteína. [100]. La rigidez puede estar relacionada con la estabilidad mecánica de una proteína [47], que se explica como la resistencia de la proteína a desplegarse en respuesta a una fuerza externa [105].

Estas medidas se calculan utilizando el software FIRST que implementa el algoritmo del juego de los guijarros [83]. Éste genera un gráfico dirigido para modelar las fuerzas físicas de las restricciones debidas a la unión

presente dentro de la proteína. Luego FIRST comprueba cada enlace para determinar si forma parte de dos tipos de elementos de rigidez: un clúster rígido o una región sometida a tensión que forma una conexión flexible.

7.2. Toolkit para cálculo de métricas

Se trata de un conjunto de herramientas de software para el análisis de proteínas que ofrece un conjunto de funciones listas para usar para calcular las propiedades que se describen a continuación. Además, como estos cálculos se ejecutarán en rutas de plegado completas o trayectorias con cientos o miles de conformaciones de proteínas, por lo que usamos un marco de trabajo distribuido propio para evaluarlas de forma distribuida en ordenadores de escritorio sencillos.

7.2.1. Funciones

Las principales funciones del conjunto de herramientas se presentaron en la tabla 7.1. Aquí, presentamos en la Tabla 7.2.2 las propiedades con el nombre de la función implementada y un resumen de las diferentes herramientas y parámetros utilizados para calcular cada propiedad. El kit de herramientas ofrece un conjunto de funciones a través de programas y bibliotecas independientes de python que pueden utilizarse directamente desde una interfaz de línea de comandos o llamarse desde programas externos, respectivamente.

7.2.2. Implementación

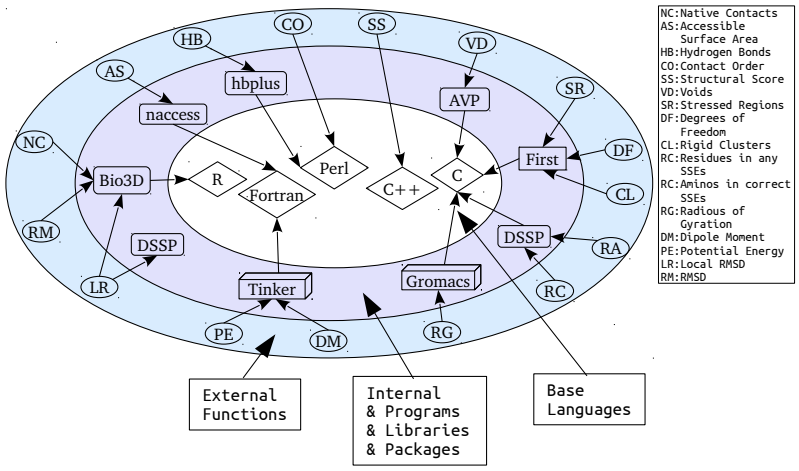
El kit de herramientas implementa diferentes algoritmos y sirve como *wrapper* de varias funcionalidades incluidas en otras herramientas de software para la manipulación de proteínas desarrolladas en los últimos años como programas, librerías, paquetes y frameworks. La figura 7.2.2 presenta una vista de las funciones implementadas en el kit de herramientas. En la capa externa están nuestras funciones utilizadas directamente para evaluar las propiedades de una estructura proteica. En el medio hay una capa de programas externos como bibliotecas y paquetes que están 'envueltos' por nuestras funciones para implementar su funcionalidad. Y en la capa interna están los diferentes lenguajes en los que se implementan algunas de nuestras funciones o programas externos.

Función	ID	Descripción
native_contacts	NC	El NC utiliza mapas de contacto creados con la función cmap del paquete bio3d del sistema R[39]. Define un umbral de contacto de 7 Å y un umbral de vecindad de 3 átomos.
contact_order	OC	OC uses the code developed by Baker and co-workers [81] con un umbral de 6 Å.
radius_gyration	RG	Utiliza el g_gyrate del paquete Gromacs para la simulación de dinámica molecular[103].
hydrogen_bonds	HB	Usa el programa *hbplus* [70], identificando enlaces de hidrógeno con una distancia máxima de 2.6 y 3.9 Å para el donante-aceptor y el receptor de hidrógeno, respectivamente.
asa	AS	Usa el programa *naccess [44]*, con una sonda de 1.4 Å para calcular el área total polar y no polar.
voids	VD	Usa el programa *AVP* [22]para calcular el volumen de los vacíos.
rmsd	RM	Usa la función *rmsd* del paquete *bio3d*.
local_rmsd	LR	Usa el programa DSSP [50]para obtener los elementos de estructura secundaria de las proteínas.
potential_energy	PE	Usa el programa *analyse* del paquete *tinker* para simulación de dinámica molecular [79] con un campo de fuerza *amber96*
dipole_moment	DM	También usa el programa *analyse* del paquete *tinker* para simulación de dinámica molecular.
correct_sse	RC	Usa DSSP para obtener los elementos de estructura secundaria asignados para cada residuo.
any_sse	RA	Como RC, usa DSSP para obtener los elementos de estructura secundaria asignados para cada residuo.
structural_score	SS	Codifica la estructura objetivo y la nativa como una secuenca 1D de bloques de proteína usando una versión modificada del métod PB-ALIGN [68].
rigid_clusters	CL	Usa la herramienta FIRST para análisis de rigidez[46].
degrees_freedom	DF	Usa la herramienta FIRST para análisis de rigidez[46].
stressed_regions	SR	Usa la herramienta FIRST para análisis de rigidez[46].

Cuadro 7.2: **Principales funciones del toolkit de análisis de proteínas.**

7.3. Instalación y despliegue

El sistema consiste en un único archivo ('vm-eval.ova') que corresponde a un dispositivo virtual (también conocido como máquina virtual). Es un completo paquete de software listo para usar (sistema operativo con apli-



Estructura del Toolkit para análisis de proteínas Tres capas. En la primera capa, las funciones para evaluar las propiedades de las proteínas. En la segunda los programas externos 'envueltos' por nuestras funciones. En la tercera, los lenguajes de computación usados por las funciones y programas externos.

caciones) que no necesita configuración o instalación especial. El usuario sólo necesita descargar el dispositivo (`http://bioinformatica.univalle.edu.co/software/vm-eval`) e importarlo a un virtualizador preinstalado (un software de virtualización como VirtualBox, VMWare, Xen).

En este proyecto hemos utilizado el virtualizador VirtualBox(`http://www.virtualbox.com`) para construir, importar y ejecutar el artefacto. Se preconfiguró con requisitos básicos de hardware que los usuarios pueden cambiar de acuerdo con la capacidad del software del virtualizador y de la máquina, que en VirtualBox son la cantidad de RAM asignada al dispositivo, el número de núcleos de CPU virtual que debe ver el dispositivo y el porcentaje de uso de la CPU (por defecto es del 100 %).

Actualmente, el dispositivo se ejecuta en un sistema operativo Linux reducido (servidor Ubuntu 12.04) con una interfaz de línea de comandos. El usuario predefinido que ejecuta la aplicación es el usuario *fm* (*/home/fm*) en cuyo directorio hay dos subdirectorios, uno para ejecutar el procesamiento distribuido *(/home/fm/framework)* y otro para la aplicación del usuario *(/home/fm/application)*. El repositorio de nube creado para el servicio de almacenamiento en nube es el */home/fm/Dropbox*). Otros directorios son el repositorio de los resultados */home/fm/results*, y un sistema de archi-

vos RAM para una lectura y escritura rápida de archivos temporales */home/fm/ramdisk*.

7.3.1. Resultados

El *framework* se utilizó para evaluar distributivamente todas las vías y trayectorias, y para realizar los análisis necesarios para culminar el presente trabajo. El framework fue operado en los PCs de escritorio del Laboratorio de Bioinformática en Universidad del Valle`http://eisc.univalle.edu.co/index.php/grupos-investigacion/bioinformatica`. Específicamente, realizamos la evaluación de 16 propiedades de las proteínas en tres conjuntos de datos diferentes: (1) en las 16000 conformaciones de una trayectoria de plegado de proteínas del subdominio de la villin headpiece (HP-35NleNle) from the folding@home project generated by molecular dynamics (see chapter 5); (2) las 37 vías obtenidas del aprendizaje basado en PRM tomado del Servidor Parasol, que comprende alrededor de 6000 conformaciones (ver capítulo 6); y (3) en las conformaciones 630000 de una trayectoria de plegado de proteínas del subdominio de villina. (HP-35NleNle) obtenidas por el gruo de investigación de David E. Shaw [64] usando dinámica molecular.

El framework funcionó sin problemas; y los tiempos de ejecución disminuyeron considerablemente en comparación con nuestros intentos antes de desarrollar el framework.

Capítulo 8

Resultados

Los resultados de aplicar las definiciones anteriores a una trayectoria de Anton es presentada en la figura 8.1. Dicha figura muestra los niveles de plegado para las conformaciones de la trayectoria simulada de la villina [64]. El comportamiento de plegado que se observa es que la ruta de plegamiento comienza con un nivel desplegado y durante ese nivel la proteína presenta muchos saltos correspondientes a eventos de plegado en los que los elementos estructurales pueden aparecer y desaparecer tratando de alcanzar estados más estables, pero de repente el plegado alcanza el estado de plegado. Las conformaciones con un bajo grado de plegado se encuentran en los pasos iniciales de plegado, y las con un alto grado se encuentran en los pasos finales de plegado.

Este comportamiento es interesante ya que el modelo de la dinámica de los niveles de plegado captura que sólo hay dos estados con alta probabilidad: el desplegado y el plegado. Este comportamiento es muy diferente al encontrado en la obra de Luis Garreta, ya que en su análisis encontró caminos con conformaciones en cuatro estados: desplegado, intermedio temprano y tardío, y plegado como se muestra en la siguiente figura. (Figure 8.2)

La aparición de dos estados en el plegamiento de proteínas ya ha sido reportado anteriormente en la literatura[113, 85, 7]. Esto significa que para dichas proteínas, las transición del estado desplegado al nativo ocurre de manera más bien directa, sin la aparación de intermedios.

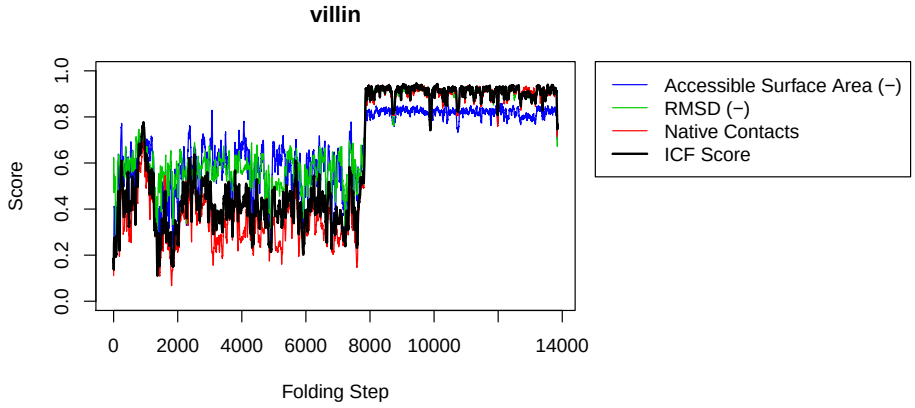


Figure 8.1: **Niveles de plegamiento de la trayectoria de la proteína de la vilina.** Conformaciones organizadas a lo largo del eje x, de la manera en la que aparecen en su ruta de plegamiento. Niveles desplegados en la ruta inicial (Paso de plegamiento < 8000 μs) y niveles plegados en los pasos finales de la ruta (pasos de plegamiento > 8000 μs) .

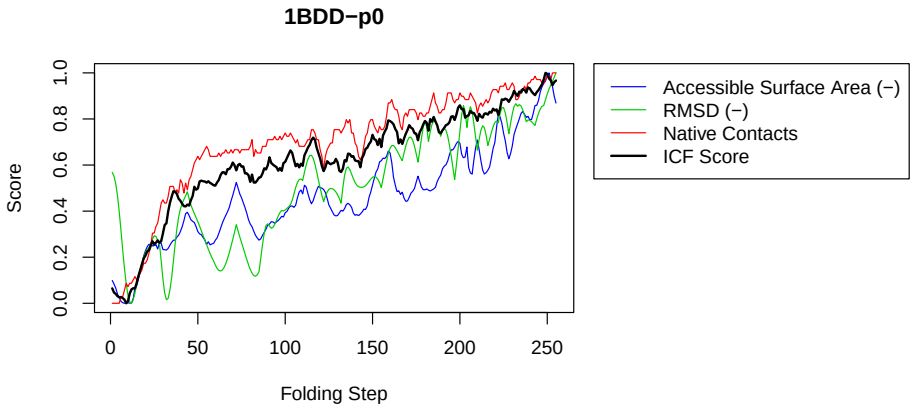


Figure 8.2: **Niveles de plegamiento para una ruta de 4 estados.** El gráfico muestra que el plegado comienza en el estado de despliegue, y después de pasar por niveles intermedios de plegado, alcanza el estado plegado.

Chapter 9

Conclusiones

En este trabajo presentamos un algoritmo de reducción de trayectorias que visualmente produce reducciones que logran preservar la dinámica de la trayectoria original en cuanto a los eventos principales y la relación de tiempo en la que estos ocurren. El algoritmo tiene cuatro fases: particionamiento, reducción local, y reducción global. El algoritmo es altamente configurable, se puede escoger el número de conformaciones de estructuras de proteínas por partición, el umbral de comparación entre dos conformaciones, y el número K para seleccionar las más representativas por partición. Además, el enfoque de particiones que tiene el algoritmo lo vuelve altamente paralelizable ya que cada reducción (local y global) se aplica de forma independiente, tanto local como, sobre cada una de ellas.

Usamos la métrica de TM-score en vez del RMSD para comparar las estructuras de proteínas. Aunque tradicionalemente se ha usado el RMSD, se conoce muy bien que esta métrica es muy sensible a pequeñas diferencias (grupos de átomos) entre las estructuras. Esas pequeñas diferencias dan como resultado grandes valores de RMSD que sugieren que las estructuras comparadas son muy diferentes. El TM-score es una métrica más robusta que el RMSD y produce mejores resultados a la hora de comparar estructuras de conformaciones muy cercanas, que es exactamente lo que se tiene cuando se comparan estructuras de conformaciones consecutivas en una línea de tiempo.

La implementación del algoritmo se realizó en el lenguaje R y Fortran para las librerías de agrupamiento y la fácil paralelización de tareas. En R están implementados los tres módulos: particionamiento, clustering local, y clustering global, mientras que en Fortran está implementada la rutina de evaluación del TM-score, que es la que más se llama tanto en el agrupamiento rápido de la reducción local, como en el agrupamiento detallado de la reducción local.

En este trabajo se pudo observar que es posible la aplicación de técnicas usadas para reducción de secuencias de ADN en la reducción de secuencias de proteínas, lo que provee un mecanismo útil a la hora de manerjar grandes cantidades de información como es el caso de las trayectorias de

simulaciones de plegamiento de proteínas al ofrecer un subconjunto de la población total que resulta significativo en cuanto a las características de las trayectorias (como la semejanza de cada conformación respecto a la nativa). Esto, por otro lado, facilitó la aplicación de un trabajo anterior sobre rasgos conformacionales en un rango más amplio de proteínas, con muchas más conformaciones, lo que permitió observar resultados distintos a los reportados en dicho trabajo al encontrarse con que estos nuevos conjuntos de proteínas presentan 2 estados en lugar de 4. Estos resultados fueron contrastados con los encontrados en la literatura y pudimos concluir que en ciertas proteínas, la presencia de unicamente dos estados es normal. Esto sugiere que por tratarse de proteínas pequeñas (30∼40 aminoácidos) y para las cuales el proceso de plegamiento es muy rápido, estas pasan del estado no plegado al plegado sin alcanzar estado intermedios claramente visibles.

Pudimos comprobar además que el uso de aplicaciónes virtuales es muy util sobre todo en el campo de la bioinformática. Esto porque muchas veces se tienen funcionalidades que son codificadas en distintos lenguajes de programación que pueden ser dependientes en su compilación de determinados sistemas operativos lo que dificulta su portabilidad, despliegue y uso. Al empaquetar todas estas funcionalidades en una máquina virtual, podemos abstraer todos estos detalles y concentrarnos más en la aplicación de dichas funcionalidades sobre el problema actual que estemos analizando.

Bibliografía

[1] Modelling sequential protein folding under kinetic control.

[2] Virtual Box. `https://www.virtualbox.org/`, 2018.

[3] Vmware. `https://www.vmware.com/`, 2018.

[4] Xen project. `https://www.xenproject.org/`, 2018.

[5] Osman Abul, Anthony Lo, Reda Alhajj, Faruk Polat, and Ken Barker. Cluster validity analysis using subsampling. In *Systems, Man and Cybernetics, 2003. IEEE International Conference on*, volume 2, pages 1435–1440. IEEE, 2003.

[6] Mehmet Serkan Apaydin, Douglas L Brutlag, C Guestrin, D Hsu, and J Latombe. Stochastic roadmap simulation: an efficient representation and algorithm for analyzing molecular motion. *J. Comput. Biol.*, 10:257–281, 2003.

[7] Audun Bakk, Johan S. HÃžye, Alex Hansen, Kim Sneppen, and Mogens H. Jensen. Pathways in two-state protein folding. *Biophysical Journal*, 79(5):2722 – 2727, 2000.

[8] Robert L Baldwin and George D Rose. Is protein folding hierarchic ? II . Folding intermediates and transition states. volume 98, pages 77–83. 1999.

[9] Robert L Baldwin and George D Rose. Molten globules, entropy-driven conformational change and protein folding. *Current opinion in structural biology*, 23(1):4–10, February 2013.

[10] Bryan Bergeron. *Bioinformatics Computing*. Prentice Hall PTR, 2002.

[11] Jacob Bien and Robert Tibshirani. Hierarchical Clustering With Prototypes via Minimax Linkage. *Journal of the American Statistical Association*, 106(495):1075–1084, September 2011.

[12] Richard Bonneau, Ingo Ruczinski, Jerry Tsai, and David Baker. Contact order and ab initio protein structure prediction. *Protein Science*, pages 1937–1944, 2002.

[13] I. Borg and P. Groenen. *Modern Multidimensional Scaling: Theory and Applications*. Springer Series in Statistics, Berlin, 2 edition, 2005.

[14] David J Brockwell and Sheena E Radford. Intermediates : ubiquitous species on folding energy landscapes ? *Current opinion in structural biology*, 17:30–37, 2007.

[15] Igor Burstyn. Principal component analysis is a powerful instrument in occupational hygiene inquiries. *The Annals of occupational hygiene*, 48(8):655–61, November 2004.

[16] Christopher Bystroff and Yu Shao. Modeling Protein Folding Pathways. 2003.

[17] T Chalikian and K Breslauer. Compressibility as a means to detect and characterize globular protein states. *Proc. Natl. Acad. Sci. USA*, 93(3):1012–1014, 1996.

[18] Hue Sun Chan and Ken A. Dill. The Protein Folding Problem. *Physics Today*, 1993.

[19] Yiwen Chen, Feng Ding, Huifen Nie, Adrian W Serohijos, Shantanu Sharma, Kyle C Wilcox, Shuangye Yin, and Nikolay V Dokholyan. Protein folding: then and now. *Archives of biochemistry and biophysics*, 469(1):4–19, January 2008.

[20] Mykyta V Chubynsky. Characterizing the intermediate phases through topological analysis. pages 1–37, 2008.

[21] Anna B. Costello and Jason W. Osborne. Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation*, 10:173–178, 2005.

[22] Alison L Cuff and Andrew C R Martin. Analysis of void volumes in proteins and application to stability of the p53 tumour suppressor protein. *Journal of Molecular Biology*, 344(5):1199–1209, 2004.

[23] Valerie Daggett and A Fersht. Is there a unifying mechanism for protein folding? *Trends Biochem Sci.*, 28(1):18–25, January 2003.

[24] Payel Das, Mark Moll, and H Stamati. Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction. *Proceedings of the . . .* , 103(26), 2006.

[25] Xavier Daura, Karl Gademann, Bernhard Jaun, Dieter Seebach, Wilfred F. van Gunsteren, and Alan E. Mark. Peptide Folding: When Simulation Meets Experiment. *Angewandte Chemie International Edition*, 38(1-2):236–240, 1999.

[26] James Dean. Choosing the Right Type of Rotation in PCA and EFA. *Shiken: JALT Testing & Evaluation SIG Newsletter*, 13(November):20–25, 2009.

[27] Ken A. Dill and Justin L. MacCallum. The protein-folding problem, 50 years on, 2012.

[28] Ken A. Dill, S. Banu Ozkan, M. Scott Shell, and Thomas R. Weikl. Review The protein folding problem. *Annual review of biophysics*, 2008.

[29] Ken A. Dill, S. Banu Ozkan, Thomas R. Weikl, John D. Chodera, and Vincent A. Voelz. The protein folding problem: when will it be solved?, 2007.

[30] Nikolay V Dokholyan. What is the protein design alphabet? *Proteins*, 54(4):622–8, March 2004.

[31] Idilio Drago, Marco Mellia, Politecnico Torino, Maurizio M Munafò, and Anna Sperotto. Inside Dropbox : Understanding Personal Cloud Storage Services.

[32] Mojie Duan, Jue Fan, Minghai Li, Li Han, and Shuanghong Huo. Evaluation of Dimensionality-reduction Methods from Peptide Folding-unfolding Simulations. *Journal of chemical theory and computation*, 9(5):2490–2497, may 2013.

[33] Yong Duan and Peter A Kollman. Pathways to a Protein Folding Intermediate Observed in a 1-Microsecond Simulation in Aqueous Solution. *Science*, 282(October):740–744, 1998.

[34] Yong Duan, Lu Wang, and Peter A Kollman. The early stage of folding of villin headpiece subdomain observed in a 200-nanosecond fully solvated molecular dynamics simulation. *Proceedings of the National*

Academy of Sciences of the United States of America, 95(17):9897–9902, 1998.

[35] Robert C. Edgar. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26(19):2460–2461, 2010.

[36] P Fleming, H. Gong, and G Rose. Secondary structure determines protein topology. *Protein Sci*, 15(8):1829–1834, August 2006.

[37] Luis Garreta and Irene Tischer. Evaluation of structural and energetic protein properties on the villin folding simulation. In *Computing Congress (CCC), 2011 6th Colombian*, pages 1–6. IEEE, 2011.

[38] Haipeng Gong, Patrick J Fleming, and George D Rose. Building native protein conformation from highly approximate backbone torsion angles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(45):16227–16232, 2005.

[39] B.J. Grant, A.P.C. Rodrigues, K.M. ElSawy, J.A. McCammon, and L.S.D. Caves. Bio3D: An R package for the comparative analysis of protein structures. *Biinformatics*, 22:2695–2696, 2006.

[40] E James Harner. *Modeling Multivariate Data*. 2012.

[41] L. Hatcher. *A step-by-step approach to using the SAS system for factor analysis and structural equation modeling*. 1994.

[42] B M Hespenheide, A J Rader, Michael F Thorpe, and L A Kuhn. Identifying protein folding cores from the evolution of flexible regions during unfolding. *J Mol Graph Models*, 21:195–207, 2002.

[43] Uwe Hobohm, Michael Scharf, Reinhard Schneider, and Chris Sander. Selection of representative protein data sets. *Protein Science*, 1(3):409–417, mar 1992.

[44] S.J. Hubbard, S.F. Campbell, and J M Thornton. Conformational analysis of limited proteolytic sites and serine proteinase protein inhibitors. *J Mol Biol.*, 220:507–530, 1991.

[45] S Jackson. How do small single-domain proteins fold? *Folding and Design*, 3(4):R81–R91, August 1998.

[46] D J Jacobs, a J Rader, L a Kuhn, and Michael F Thorpe. Protein flexibility predictions using graph theory. *Proteins*, 44(2):150–65, August 2001.

[47] Donald J Jacobs, Leslie A Kuhn, and Michael F Thorpe. Flexible and Rigid Regions in Proteins. In M F Thorpe and P M Duxbury, editors, *Rigidity Theory and Applications*, Fundamental Materials Research, pages 357–384. Springer US, 2002.

[48] Anil JAIN, Karthik NANDAKUMAR, and Amn ROSS. Score normalization in multimodal biometric systems. *Pattern recognition*, 38(12):2270–2285, 2005.

[49] Leili Javidpour. Computer Simulations of Protein Folding. *Computing in Science & Engineering*, 14(2):97–103, 2012.

[50] W. Kabsch and C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22:2577–2637, 1983.

[51] Hirak Kashyap, Hasin Afzal Ahmed, Nazrul Hoque, Swarup Roy, and Dhruba Kumar Bhattacharyya. Big data analytics in bioinformatics: architectures, techniques, tools and issues. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 5(1):28, Sep 2016.

[52] Avita Katal, Mohammad Wazid, and R. H. Goudar. Big data: Issues, challenges, tools and Good practices. In *2013 6th International Conference on Contemporary Computing, IC3 2013*, 2013.

[53] R Kolodny and M Levitt. Protein decoy assembly using short fragments under geometric constraints. *Biopolymers*, 68(3):278–285, March 2003.

[54] Irina Kufareva and Ruben Abagyan. Methods of protein structure comparison. pages 231–257, 2015.

[55] Abel Lajtha, Naren Banik, A. Szilágyi, J. Kardos, S. Osváth, L. Barna, and P. Závodszky. *Handbook of Neurochemistry and Molecular Neurobiology*. Springer US, Boston, MA, 2007.

[56] T Lazaridis and M Karplus. New View of Protein Folding Reconciled with the Old Through Multiple Unfolding Simulations. *Science*, 278:1928–1931, 1997.

[57] B Lee and F M Richards. The interpretation of protein structures: Estimation of static accessibility. *J. Mol. Biol.*, 55:379–400, 1971.

[58] Erel Levine and Eytan Domany. Resampling Method For Unsupervised Estimation Of Cluster Validity. 2000.

[59] C Levinthal. Are there pathways for protein folding? *J Chem Phys*, 65:44–45, 1968.

[60] W. Li, L. Jaroszewski, and A. Godzik. Tolerating some redundancy significantly speeds up clustering of large protein databases. *Bioinformatics*, 18(1):77–82, 2002.

[61] J Liang and K a Dill. Are proteins well-packed? *Biophysical journal*, 81(2):751–66, August 2001.

[62] Jie Liang, Herbert Edelsbrunner, Ping Fu, Pamidighantam V Sudhakar, and Shankar Subramaniam. Analytical shape computation of macromolecules: II. Inaccessible cavities in proteins. *Proteins: Struct., Funct., Genet.*, 33(1):18–29, 1998.

[63] K Lindorff-Larsen, S Piana, R O Dror, and D E Shaw. How Fast-Folding Proteins Fold. *Science*, 334(5):517–520, oct 2011.

[64] Kresten Lindorff-Larsen, Stefano Piana, Ron O Dror, and David E Shaw. How fast-folding proteins fold. *Science (New York, N.Y.)*, 334(6055):517–20, October 2011.

[65] M Lobanov, Natalya S Bogatyreva, and Oxana V Galzitskaya. Radius of gyration as an indicator of protein structure compactness. *Molecular Biology*, 42(4):623–628, 2008.

[66] Peter S Lomdahl and David M Beazley. State-of-the-Art Parallel Computing: molecular dynamics on the connection machine. *Los Alamos Science*, (22):45–57, 1994.

[67] A. Marsden. M. Lougher, M. Lücken, T Machon, M. Malcomson. Computational Modelling of Protein Folding. Technical report.

[68] N Srinivasan M. Tyagi A. de Brevern and B Offmann. Protein structure mining using a structural alphabet. *Proteins*, 71(2):920–937, 2008.

[69] Martin Maechler, Peter Rousseeuw, Anja Struyf, Mia Hubert, and Kurt Hornik. *cluster: Cluster Analysis Basics and Extensions*, 2013.

[70] I K McDonald and J M Thornton. Satisfying Hydrogen Bonding Potential in Proteins. *Journal of Molecular Biology*, 238:777–793, 1994.

[71] Seonwoo Min, Byunghan Lee, and Sungroh Yoon. Deep learning in bioinformatics, 2017.

[72] John Moult. Rigorous performance evaluation in protein structure modelling and implications for computational biology. *Philos Trans R Soc Lond B Biol Sci*, 361:453–458, 2006.

[73] John Moult, Krzysztof Fidelis, Andriy Kryshtafovych, Torsten Schwede, and Anna Tramontano. Critical assessment of methods of protein structure prediction (CASP)–round x. *Proteins*, 82 Suppl 2(0 2):1–6, feb 2014.

[74] A Mucherino, Susan Costantini, D di Serafino, M D'Apuzzo, Angelo Facchiano, and Giovanni Colonna. Understanding the role of the topology in protein folding by computational inverse folding experiments. *Comput. Biol. Chem.*, 32(4):233–239, 2008.

[75] Takao Nomura, Rui Kamada, Issaku Ito, Koichi Sakamoto, Yoshiro Chuman, Koichiro Ishimori, Yasuyuki Shimohigashi, and Kazuyasu Sakaguchi. Probing phenylalanine environments in oligomeric structures with pentafluorophenylalanine and cyclohexylalanine. *Biopolymers*, 95(6):410–9, June 2011.

[76] S Ozkan, G Wu, J Chodera, and K Dill. Protein folding by zipping and assembly. *Proc Natl Acad Sci USA.*, 104:119:87–92, 2007.

[77] V S Pande, A.Yu Grosberg, D Rokshar, and T Tanaka. Pathways for protein folding: is a new view needed? *Curr. Opin. Struct. Biol.*, 8:68–79, 1998.

[78] Vijay S Pande, Eric J Sorin, Christopher D Snow, and Young Min Rhee. Chapter 8 Computer Simulations of Protein Folding. In *Protein Folding, Misfolding and Aggregation: Classical Themes and Novel Approaches*, pages 161–187. The Royal Society of Chemistry, 2008.

[79] R. V. Pappu, R. K. Hart, and J.W. Ponder. Tinker: a package for molecular dynamics simulation. *J. Phys. Chem. B*, 102:9725–42, 1988.

[80] Hae-Sang Park and Chi-Hyuck Jun. A simple and fast algorithm for K-medoids clustering. *Expert Systems with Applications*, 36(2):3336–3341, mar 2009.

[81] K W Plaxco, K T Simons, and David Baker. Contact order, transition state placement and the refolding rates of single domain proteins. *J Mol Biol*, 277(4):985–994, April 1998.

[82] J W Ponder and D A Case. Force fields for protein simulations. *Adv Protein Chem*, 66:27–85, 2003.

[83] A. J Rader, Brandon M Hespenheide, Leslie A Kuhn, and Michael F Thorpe. Protein unfolding: rigidity lost. *Proceedings of the National Academy of Sciences of the United States of America*, 99(6):3540–5, March 2002.

[84] Aruna Rajan, Peter L Freddolino, and Klaus Schulten. Going beyond clustering in MD trajectory analysis: an application to villin headpiece folding. *PloS one*, 5(4):e9890, January 2010.

[85] Hospital Readmissions and Reduction Program. HHS Public Access. 131(20):1796–1803, 2016.

[86] H Roder and W Colón. Kinetic role of early intermediates in protein folding. *Current opinion in structural biology*, 7(1):15–28, February 1997.

[87] Heinrich Roder, Kosuke Maki, and Hong Cheng. Early Events in Protein Folding Explored by Rapid Mixing Methods. *Chemical Reviews*, 106:1836–1861, 2006.

[88] P Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(1):53–65, 1987.

[89] Karen Sargsyan, Cédric Grauffel, and Carmay Lim. How Molecular Size Impacts RMSD Applications in Molecular Dynamics Simulations. *Journal of Chemical Theory and Computation*, 13(4):1518–1524, apr 2017.

[90] Courtney Schirf. *Automated Protein Classification Using Rigidity Analysis*. PhD thesis.

[91] David E. Shaw, Jack C. Chao, Michael P. Eastwood, Joseph Gagliardo, J. P. Grossman, C. Richard Ho, Douglas J. Lerardi, István Kolossváry, John L. Klepeis, Timothy Layman, Christine McLeavey, Martin M. Deneroff, Mark A. Moraes, Rolf Mueller, Edward C. Priest, Yibing Shan,

Jochen Spengler, Michael Theobald, Brian Towles, Stanley C. Wang, Ron O. Dror, Jeffrey S. Kuskin, Richard H. Larson, John K. Salmon, Cliff Young, Brannon Batson, and Kevin J. Bowers. Anton, a special-purpose machine for molecular dynamics simulation. *Communications of the ACM*, 51(7):91, 2008.

[92] Yi Shi, Jianjun Zhou, David Arndt, David Wishart, and Guohui Lin. Protein contact order prediction from primary sequences. *BMC Bioinformatics*, 9(1):255, 2008.

[93] A Shrake and J Rupley. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *J. Mol. Biol.*, 79:351–371, 1973.

[94] Elena Sivogolovko and Boris Novikov. Validating cluster structures in data mining tasks. *Proceedings of the 2012 Joint EDBT/ICDT Workshops on - EDBT-ICDT '12*, page 245, 2012.

[95] Guang Song. *A Motion Planning Approach to Protein Folding*. PhD thesis, Dept. of Computer Science, Texas A&M University, December 2004.

[96] Guang Song and Nancy M Amato. Using Motion Planning to Study Protein Folding Pathways. *Journal of Computational Biology*, pages 287–296, 2001.

[97] Soumya Raychaudhuri and Joshua M. Stuart and Russ B. Altman. Principal Components Analysis to Summarize Microarray Experiments: Application to Sporulation Time Series. 463:452–463, 2000.

[98] D F Stickle, L G Presta, K A Dill, and G D Rose. Hydrogen bonding in globular proteins. *J Mol Biol*, 226:1143–1159, 1992.

[99] S Takashima. Use of protein database for the computation of the dipole moments of normal and abnormal hemoglobins. *Biophysical journal*, 64(5):1550–8, May 1993.

[100] Michael F Thorpe, M Lei, a J Rader, D J Jacobs, and L a Kuhn. Protein flexibility and dynamics using constraint theory. *Journal of molecular graphics & modelling*, 19(1):60–9, January 2001.

[101] Jayant B Udgaonkar. Multiple Routes and Structural Heterogeneity in Protein Folding. *Annual Review of Biophysics*, 37:489–510, 2008.

[102] Vladimir N Uversky. What does it mean to be natively unfolded? *The Federation of European Biochemical Societies Journal*, 269(1):2–12, 2002.

[103] D van der Spoel, E Lindahl, B Hess, G Groenhof, A E Mark, and Herman J C Berendsen. GROMACS: Fast, Flexible and Free. *J. Comp. Chem.*, 26:1701–1719, 2005.

[104] David van der Spoel, Erik Lindahl, Berk Hess, Aldert R van Buuren, Emile Apol, Pieter J Meulenhoff, D. Peter Tieleman, Alfons L T M Sijbers, K Anton Feenstra, Rudi van Drunen, and Herman J C Berendsen. Gromacs User Manual version 3.3, 2005.

[105] Hui-chuan Wang. *STUDIES ON THE MECHANICAL STABILITY OF A PROTEIN BY SINGLE-MOLECULE ATOMIC FORCE MICROSCOPY*. PhD thesis, 2009.

[106] David Whitford. *Proteins: Structure and Function*. John Wiley and Sons, Ltd, 2005.

[107] Florian Wickelmaier. *An introduction to MDS*. Aalborg Universitetsforlag, 2003.

[108] Jinrui Xu and Yang Zhang. How significant is a protein structure similarity with TM-score = 0.5? 26(7):889–895, 2010.

[109] Hui Yang, Srinivasan Parthasarathy, and Duygu Ucar. A spatio-temporal mining approach towards summarizing and analyzing protein folding trajectories. *Algorithms for molecular biology : AMB*, 2(Md):3, 2007.

[110] M Zaki, V Nadimpally, D Bardhan, and C Bystroff. Predicting protein folding pathways. *Bioinformatics*, 20:386–393, 2004.

[111] J. Zhang Y.,Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins*, 57:702–710, 2004.

[112] Feng-Yu Li Zhao and Ji-Jun. Quantum chemistry PM3 calculations of sixteen mEGF molecules.

[113] Robert Zwanzig. Two-state models of protein folding kinetics. *Proceedings of the National Academy of Sciences*, 94(1):148–150, 1997.