UNIVERSIDAD NACIONAL DE COLOMBIA

Un método híbrido para la predicción de la estructura terciaria de las proteínas a partir de su secuencia de aminoácidos

Diego Felipe Carvajal Patiño

Universidad Nacional de Colombia

Facultad de Ingeniería, Departamento de Ingeniería de Sistemas e Industrial

Bogotá, D.C., Colombia

2015

Un método híbrido para la predicción de la estructura terciaria de las proteínas a partir de su secuencia de aminoácidos

Diego Felipe Carvajal Patiño

Tesis de investigación presentada como requisito parcial para optar al título de:

Magister en Ingeniería de Sistemas y Computación

Director:

Luis Fernando Niño Vásquez, Ph.D.

Línea de Investigación:

Bioinformática

Grupo de Investigación:

Laboratorio de Investigación en Sistemas Inteligentes - LISI

Universidad Nacional de Colombia

Facultad de Ingeniería, Departamento de Ingeniería de Sistemas e Industrial

Bogotá, D.C., Colombia

2015

A mis padres y hermanos.

Es el seguir arando lo que te hará ganar la jornada. ¡Así que no te me hagas el remolón, mi viejo amigo! Continúa esforzándote. ¡Es tan fácil dejarlo todo! Es mantener el mentón alzado y trabajar con persistencia lo que nos da la victoria.

Robert W. Service

Per aspera ad astra «A través del esfuerzo, el triunfo»

Agradecimientos

Agradezco a Dios por permitirme finalizar esta investigación, al profesor Luis Fernando Niño por su apoyo, paciencia, comprensión y orientación durante el desarrollo de este trabajo de grado. A Juvenal Yosa por guiar con su conocimiento y experiencia la ejecución de esta tesis. A mis padres y hermanos por apoyarme incondicionalmente, por comprenderme, por animarme en la consecución de mis metas y hacer de mí una persona íntegra. Así mismo, al resto de mi familia y a mis amigos por hacerme sentir una persona especial en sus vidas. Finalmente, agradezco a los compañeros del Laboratorio de Investigación en Sistemas Inteligentes (LISI) por los comentarios y las sugerencias realizadas para que este trabajo de investigación sea una realidad.

Resumen

En esta tesis se abordó el problema de la predicción de la estructura de las proteínas, lo cual es un problema fundamental de la biología debido a que la función de una proteína está determinada por su estructura terciaria. Por tanto, son de vital importancia los métodos que ayuden a identificar la estructura terciaria de las proteínas. En este trabajo se propone un método de inteligencia computacional para la predicción del plegamiento de las proteínas, que se compone de dos etapas: en la primera se genera un modelo tridimensional de la estructura de la proteína utilizando PyRosetta, mediante un procedimiento de ensamblaje de fragmentos de tamaño 3 y 9; y en la segunda etapa, se realiza un refinamiento de la estructura obtenida en la primera fase utilizando el algoritmo multiobjetivo de búsqueda dispersa AbYSS. Se efectuaron experimentos con las proteínas 1CRN, 2KBQ, 1ROP y 2MQL, en las cuales la estructura terciaria fue determinada experimentalmente, con el objetivo de comparar los resultados obtenidos mediante la superposición estructural entre las estructuras obtenidas y las estructuras reportadas en el *Protein Data Bank*.

Palabras clave: proteínas, conformaciones, matriz interna, ensamblaje de fragmentos, optimización multiobjetivo, búsqueda dispersa.

Abstract

Protein structure prediction is a fundamental problem in biology, because the function of a protein is determined by its tertiary structure. Therefore, the methods to help to identify the protein tertiary structure are vital. To solve this problem was the main goal of this thesis. This work presents a computational intelligence method for predicting a protein tertiary structure, which consists of two stages: in the first one, a three dimensional structure of a protein is built using PyRosetta by assembling fragments of size 3 and 9; and in the second stage, a refinement of the structure obtained in the first phase is performed using the multi-objective scatter search algorithm AbYSS. Several experiments were carried out with the proteins 1CRN, 2KBQ, 1ROP and 2MQL, whose tertiary structure was obtained through an experimental process, in order to compare the results obtained by structural superposition between generated structures and reported structures in the Protein Data Bank.

Keywords: protein, conformation, internal matrix, fragment assembly, multiobjective optimization, scatter search.

Contenido

Lista de figuras

Lista de tablas

Lista de abreviaturas

Abreviatura	Término
ADN	Ácido desoxirribonucleico
ARN	Ácido ribonucleico
PDB	Protein Data Bank
RMN	Resonancia Magnética Nuclear
BLAST	Basic Local Alignment Search Tool
PSI-BLAST	Position-Specific Iterative BLAST
HMM	Hidden Markov Model
ACO	Ant Colony Optimization
CASP	Critical Assessment of Techniques for Protein Structure Prediction
SS	Scatter Search
AbYSS	Archive-based hYbrid Scatter Search
RMSD	Root Mean Square Deviation
PSIPRED	PSI-BLAST Based Secondary Structure Prediction
DSSP	Define Secondary Structure of Proteins
CATH	Class Architecture Topology Homologous superfamily
SCOP	Structural Classification of Proteins
PSVP	Protein Structure Validation Suite

Introducción

Las proteínas son macromoléculas muy versátiles en los sistemas vivos y cumplen funciones cruciales en prácticamente todos los procesos biológicos. Se caracterizan por tener funciones como catálisis, transporte, producción de energía, regulación, también en proporcionar soporte mecánico y protección inmune, generar movimiento, transmitir los impulsos nerviosos, controlar el crecimiento y la diferenciación, entre otras. Con el descubrimiento de la secuencia de aminoácidos se determinó que esta era única para cada proteína y que determinaba su estructura terciaria, ya que es el enlace entre la información genética en el ADN y la estructura tridimensional que establece la función de una proteína. A través del análisis de relaciones entre la secuencia de aminoácidos y la estructura tridimensional se han descubierto reglas que gobiernan el plegamiento de los polipéptidos [1]. Alteraciones en la secuencia de aminoácidos pueden producir anormalidades en la función de la proteína, y enfermedades como la anemia de células falciformes y la fibrosis quística [1].

El problema del plegamiento de las proteínas es uno de los grandes retos de la biología estructural molecular, el cual consiste en conocer cómo se pliega una proteína desde su estructura primaria (secuencia) hasta su estructura terciaria. Así mismo, es un problema que interesa a diversas áreas del conocimiento como la biología, la física, la química y las ciencias de la computación, entre otras [2]. La importancia de conocer la estructura de las proteínas radica en que la mayoría de las funciones de las células están determinadas por las proteínas, y para ello es necesario conocer la estructura terciaria de las proteínas debido a que ésta conduce a la función de las mismas [3].

La búsqueda de las funciones de las proteínas ha proporcionado la investigación en mecanismos que puedan resolver la estructura tridimensional a través de experimentos o simulaciones computacionales.

Los métodos que se utilizan para determinar la estructura de las proteínas son la cristalografía de rayos X y la resonancia magnética nuclear (RMN), los cuales a pesar de

tener el apoyo de diversas organizaciones en el mundo presentan dificultades como: algunas proteínas no pueden ser fácilmente cristalizadas, la cristalografía puede tomar varios meses o años para determinar la estructura de una proteína, la resonancia magnética nuclear es un poco más rápida que la cristalografía pero no puede ser aplicada en proteínas con más de 100 residuos, actualmente se conocen cerca de 100.000 estructuras de proteínas, pero solo 2000 han sido determinadas por cristalografía de rayos X, lo que corresponde a una tasa muy baja anual; y finalmente, asumiendo que se pudieran resolver las estructuras de todas las proteínas por medio de estos métodos, tomaría cerca de 500 años para alcanzar esta meta. También es importante anotar que son procesos con un alto costo económico [2].

Por los inconvenientes que presentan los métodos experimentales surgen como alternativa los métodos computacionales, que son algoritmos desarrollados con el objetivo de simular el plegamiento de las proteínas para obtener la estructura terciaria de las mismas. Los referidos métodos computacionales se caracterizan por predecir una proteína en menor tiempo y reducir el gasto económico, sin embargo su precisión en proteínas grandes (más de 150 aminoácidos en la secuencia) no es la deseable.

Predecir la estructura de una proteína a través de los métodos computacionales se puede llevar a cabo a través del modelamiento por homologías, el reconocimiento del plegamiento o *threading*, y el modelado libre (*ab initio*). El primero parte de la afirmación “proteínas con secuencia de aminoácidos similares, posiblemente tendrán estructuras tridimensionales similares”, es decir, que tiene en cuenta la cercanía evolutiva de una proteína con respecto a otras que tienen la estructura resuelta; el segundo se centra en la observación de que las proteínas usualmente adoptan pliegues similares sin importar su similitud de secuencia; y el tercero está compuesto por los métodos que no usan información existente para predecir la estructura terciaria de una proteína, sino que lo hacen con base, por ejemplo, en la dinámica molecular del proceso de plegamiento de una proteína.

El objetivo general de esta investigación consiste en predecir la estructura terciaria de una proteína a partir de su secuencia de aminoácidos mediante el uso de un algoritmo multiobjetivo de búsqueda dispersa y ensamblaje de fragmentos.

Para ello, el problema se divide en dos etapas: en la primera se genera un modelo inicial a través del uso de bibliotecas de fragmentos de tamaño tres y nueve (3-meros y 9-

meros), mientras que en la segunda se realiza un refinamiento de las estructuras generadas en la primera etapa.

En la fase inicial es necesario incluir las bibliotecas de fragmentos generadas para la secuencia de aminoácidos mediante el servidor de predicción de estructuras de proteínas Robetta, se especifican los parámetros de ejecución del algoritmo y la representación computacional de las proteínas. Posteriormente, los resultados obtenidos en la fase anterior son utilizados como información primaria en la siguiente etapa, en la cual se define la representación computacional de las proteínas, se aplica un algoritmo de búsqueda de soluciones mediante el uso de una función de puntuación encargada de guiar la exploración sobre el espacio de soluciones.

A continuación se presenta la estructura de este documento:

En el capítulo uno se presentan los fundamentos biológicos y computacionales básicos que permiten comprender el desarrollo de la investigación, y también contiene la definición del problema. En lo que respecta a los fundamentos biológicos se presentan los conceptos relacionados con los aminoácidos y las estructuras de las proteínas. En lo referente a los conceptos computacionales, se exponen los fundamentos de algoritmos de búsqueda, optimización multiobjetivo y la noción de un problema NP-completo.

En el capítulo dos se presentan los métodos computacionales que han sido propuestos para la resolución de la estructura de las proteínas, como el modelamiento por homologías, el reconocimiento del plegamiento y el modelado libre o *ab initio*. Así mismo, se describe la metodología y los componentes de cada uno de los métodos que se aplicaron.

En el capítulo tres se presenta el modelo propuesto para la predicción de la estructura de las proteínas a partir de su secuencia de aminoácidos, mediante el uso de un método de búsqueda dispersa y ensamblaje de fragmentos. Igualmente, se describe el proceso realizado para la generación del modelo inicial, utilizando el paquete de software PyRosetta, modelo que es refinado posteriormente. Luego se expone el sistema de representación de las proteínas utilizado en la segunda etapa; el algoritmo de búsqueda dispersa (SS) usado como método de exploración del espacio de conformaciones y la función de puntuación Talaris2013 que se calcula mediante el paquete de software PyRosetta, que permite evaluar la energía potencial de las proteínas.

En el capítulo cuatro se muestran los experimentos realizados con el conjunto de proteínas 1CRN, 2KBQ, 1ROP y 2MQL, las cuales son proteínas con estructura terciaria determinada a través de métodos experimentales como la RMN y la cristalografía de rayos X. Luego se realiza un análisis de los resultados obtenidos en las predicciones mediante la superposición estructural RMSD (del inglés *root-mean-square deviation*) global y local, y adicionalmente se evalúa que los ángulos de torsión de las estructuras generadas se encuentren en regiones permitidas a través de un gráfico de Ramachandran. También se realiza una comparación del método propuesto con otros servidores de predicción de estructuras de proteínas, como son QUARK y I-TASSER.

Finalmente, en el capítulo cinco se presentan las conclusiones y recomendaciones de la investigación y se dan algunas sugerencias para trabajo futuro sobre la misma.

1. Fundamentos y definición del problema

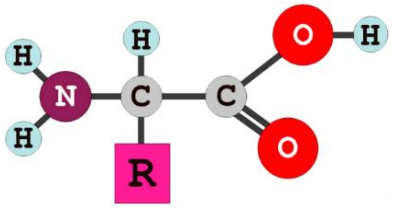
En este capítulo se presentan los conceptos biológicos y computacionales que enmarcan el desarrollo de la tesis. Inicialmente se describen los fundamentos biológicos que soportan la predicción de la estructura tridimensional de las proteínas y, posteriormente, se exponen los conceptos computacionales necesarios para comprender los antecedentes del problema objeto de investigación, el modelo propuesto y la implementación realizada. Finalmente, se explica la definición del problema objeto de investigación.

1.1 Proteínas

1.1.1 Aminoácidos y sus propiedades

Los aminoácidos son moléculas conformadas por átomos de carbono (C), hidrógeno (H), oxígeno (O) y nitrógeno (N). Se caracterizan por tener un grupo amino ($-\mathrm{NH_2}$), un grupo carboxilo ($-\mathrm{COOH}$) y un grupo variable denominado R, que corresponde a la cadena lateral; en la Figura 1-1 se observa la estructura general de un aminoácido.

Figura 1-1: Estructura general de un aminoácido.



Estructura de un aminoácido compuesta por un grupo amino (izquierda), un grupo carboxilo (derecha), una cadena lateral o residuo (R) y un átomo de carbono localizado entre el grupo amino y el grupo carboxilo llamado carbono alfa ($C\alpha$). (imagen tomada de [4]).

En la naturaleza se identifican cerca de 300 aminoácidos, pero sólo 20 de ellos pueden constituir una proteína, por tal razón son conocidos como aminoácidos proteinogénicos.

Estos aminoácidos se caracterizan por tener enlazados los grupos funcionales (amino y carboxilo) y la cadena lateral por medio de un átomo de carbono denominado alfa ($C\alpha$). Para caracterizar un aminoácido se utiliza su punto isoeléctrico, dado que es el valor de pH al que una molécula tiene carga neta igual a cero, es decir, tiene la misma cantidad de cargas negativas y positivas; en este punto la sustancia es insoluble. De acuerdo con las propiedades fisicoquímicas de la cadena lateral, los aminoácidos se clasifican en las siguientes categorías [4]:

Ácidos: son aminoácidos que se pueden cargar negativamente debido a la pérdida de un protón.

Básicos: son aminoácidos que se pueden cargar positivamente gracias a la ganancia de un protón.

Aromáticos: son aminoácidos que tienen en su cadena lateral un anillo aromático.

Azufrados: son aminoácidos que contienen azufre en su cadena lateral.

Hidrofílico o polar sin carga: son aminoácidos en los que su cadena lateral está compuesta por un grupo funcional hidrofílico, es decir, que tiene afinidad por el agua. Se caracterizan por estar involucrados en la formación de puentes de hidrógeno con el agua.

Hidrofóbico inactivo: son aminoácidos que tienen una cadena lateral de hidrocarburos sin grupos funcionales.

Estructura especial: son aminoácidos que poseen una conformación diferente a los demás, como es el caso de la prolina, en la cual el grupo R está directamente relacionado con el grupo amino.

Alifáticos: son aminoácidos no polares e hidrofóbicos, así mismo son cadenas de hidrógeno y carbono.

1.1.2 Péptidos y proteínas

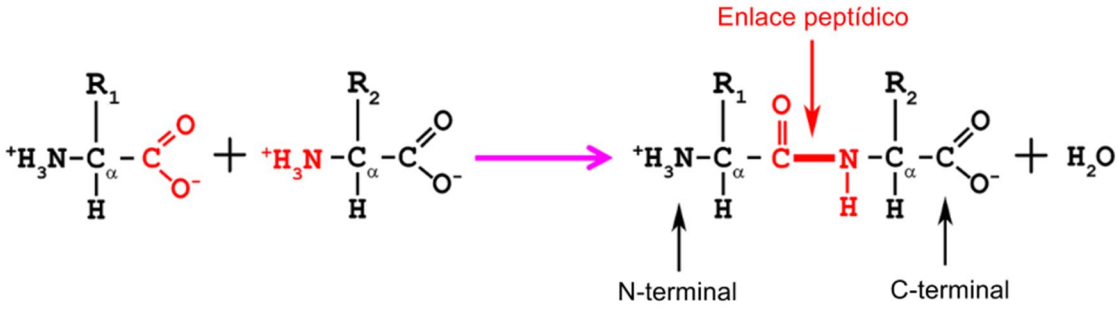
Una proteína se define como una cadena lineal de aminoácidos, así mismo la unión de varios aminoácidos (mayor a 50) conforma una proteína. Los aminoácidos en una cadena de polipéptidos están unidos por enlaces peptídicos, cada aminoácido de esta cadena se

llama residuo, y la serie enlazada de átomos de carbono, nitrógeno y oxígeno se conoce como la cadena principal o esqueleto de la proteína (*backbone*) [5] [4].

Enlaces peptídicos

Las proteínas están conformadas por uno o más polipéptidos ordenados de una forma biológicamente funcional y a menudo se unen con otras proteínas. Los enlaces peptídicos son originados en una reacción de condensación, a través de un enlace amida formado entre el nitrógeno del grupo α-amino y el carbono del grupo α-carboxilo de otro aminoácido, como se observa en la Figura 1-2. El enlace peptídico es una estructura plana y rígida, mientras que los enlaces entre el α-carbono y un grupo amino o el grupo carboxilo tienen la posibilidad de girar [4].

Figura 1-2: Enlace peptídico entre dos aminoácidos.



Formación de un enlace peptídico por reacción de condensación, las cadenas de aminoácidos son unidas para formar un péptido o proteína. (imagen adaptada de [4]).

1.1.3 Estructura de las proteínas

Las proteínas presentan diferentes niveles de organización jerarquizados e interdependientes y son denominados las estructuras primaria, secundaria, terciaria y cuaternaria. La importancia de estas estructuras radica en la dependencia que tienen entre sí, ya que la secuencia de aminoácidos (estructura primaria) determina las demás estructuras, y del mismo modo la secuencia plegada en hélices o láminas (estructura secundaria) establece la conformación de dominios estructurales (estructura terciaria) que se pueden relacionar con otras proteínas formando complejos proteicos (estructura cuaternaria) [6].

Es necesario conocer la estructura de las proteínas, ya que de ellas depende la función que pueda realizar, entre las que se encuentran: la catálisis de reacciones químicas (enzimas), el flujo de pequeñas moléculas e iones (transporte), la detección y reacción con el medio ambiente (señalización), control de la actividad de la proteína (regulación), organización del genoma, los lípidos membrana bicapa, y el citoplasma (estructura), y la generación de la fuerza para el movimiento (proteínas motoras) [6] . A continuación se describen en más detalle las estructuras de las proteínas.

Estructura primaria

Es la forma más básica de organización de las proteínas. La secuencia de aminoácidos de la cadena proteica define este tipo de estructura de las proteínas, la cual, se determina por el número de aminoácidos presentes y por el orden en que están enlazados por medio de enlaces peptídicos. Por convención, las cadenas laterales de los aminoácidos se extienden a partir de una cadena principal, que coincide con el sentido de síntesis natural, en la estructura primaria el orden de escritura es siempre desde el grupo amino-terminal hasta el carboxilo-terminal [4].

Estructura secundaria

Esta estructura consiste en el plegamiento de la cadena peptídica sobre su propio eje para formar una hélice o alguna otra estructura tridimensional específica, que se produce mediante las interacciones de puentes de hidrógeno entre fragmentos contiguos de la cadena polipeptídica. A continuación se describen los componentes de esta estructura.

Puentes de hidrógeno: es una atracción electrostática, sin ionización en las moléculas. Son conformados por un átomo de hidrógeno y dos átomos con carga negativa, normalmente nitrógeno (N) u oxígeno (O). El átomo de hidrógeno es unido con uno de los átomos de carga negativa mediante un enlace covalente y la carga eléctrica es distribuida en los tres átomos que forman el puente. Los enlaces de hidrógeno son un factor importante en la estabilización de la estructura de macromoléculas como las proteínas y los ácidos nucleicos [4].

Ángulos de torsión: las cadenas polipeptídicas no permanecen rígidas debido a que existe una rotación libre alrededor de la mayoría de los enlaces que permiten a la cadena

cambiar su orientación. Un enlace peptídico presenta tres tipos de ángulos rotables: phi (φ), psi (ψ), y omega (ω). Se denomina ángulo φ al valor del ángulo formado en el enlace N-Cα; se llama ángulo ψ al constituido en el enlace Cα-C, y el ángulo ω es aquel que define la planaridad del enlace peptídico. De los anteriores ángulos presentes en el esqueleto de la proteína, solo φ y ψ poseen libre rotación. Por tal motivo una proteína podría adoptar un número ilimitado de conformaciones (formas), sin embargo, la mayoría de cadenas polipeptídicas se pliegan adoptando una conformación particular. Esto debido a que los átomos asociados al enlace peptídico se encuentran en el mismo plano y, por tanto, el enlace peptídico no puede girar libremente (ángulo ω). El ángulo de rotación de un enlace se conoce como diedral o de torsión por estar ubicado entre dos planos [7].

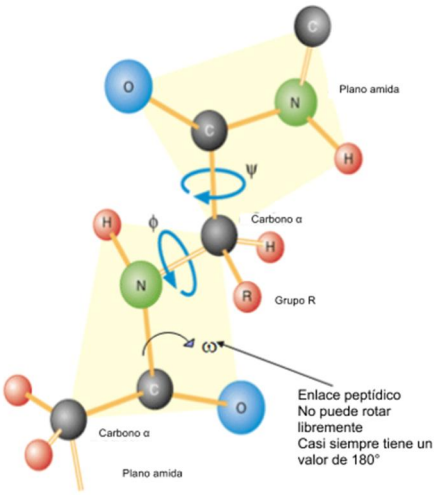
Tabla 1-1: Ángulos de torsión presentes en las proteínas.

Enlace	Rotación	Ángulo de torsión
$NH - \mathrm{C}\alpha$	Libre	φ
$\mathrm{C}\alpha$ a $\mathrm{C} = \mathrm{O}$	Libre	ψ
$\mathrm{C} = \mathrm{O}$ a NH (enlace peptídico)	Plano rígido	ω

(tabla tomada de [8])

En la naturaleza se observan tres tipos de orientaciones: helicoidal, laminar y aleatoria. Las dos primeras son disposiciones geométricas ordenadas que se deben a la repetición del mismo valor de φ y de ψ en los enlaces sucesivos. Los valores que pueden tener los ángulos de torsión ω son $0°$ y $180°$, mientras que los ángulos φ y ψ se extienden entre $-180°$ y $180°$; el signo negativo o positivo del ángulo depende de la dirección de la rotación, si el giro es menor a $180°$ en el sentido de las manecillas del reloj, es positivo; de lo contrario si la rotación es menor a $180°$ en sentido antihorario el signo será negativo [7]. En la Figura 1-3 se muestran los ángulos de torsión de una proteína.

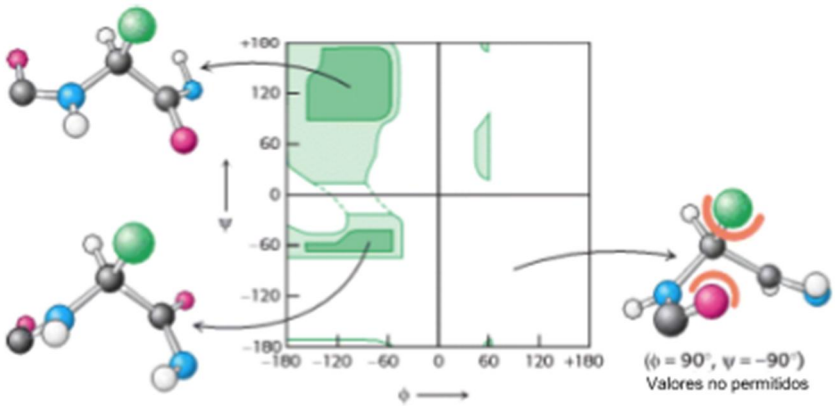
Figura 1-3: Ángulos de torsión alrededor de un enlace peptídico



(imagen adaptada de [9]).

Gráfico de Ramachandran: muestra las combinaciones de φ y ψ que son permitidas, es decir, que no todos los valores de φ y ψ son posibles sin que haya colisiones entre los átomos; en la Figura 1-4 se observa la gráfica de Ramachandran y se presenta el caso de las colisiones entre átomos [1].

Figura 1-4: Gráfico de Ramachandran



La región verde oscura exhibe las regiones más favorables para los ángulos φ y ψ, mientras que las regiones de valores límite se muestran en verde claro. (imagen adaptada de [1]).

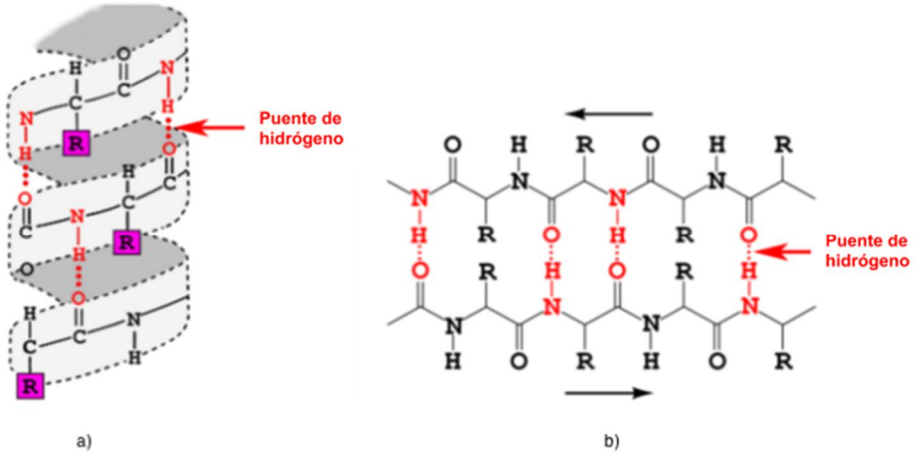
Las principales estructuras secundarias son las hélices alfa, las láminas beta y las estructuras aleatorias, las primeras dos se observan en la Figura 1-5. A continuación se presentan estas estructuras secundarias.

Hélice Alfa: se caracterizan por tener una forma helicoidal, es decir, su estructura geométrica está dada por una espiral, en la cual cada vuelta está constituida por 3,6 aminoácidos. La hélice se mantiene mediante puentes de hidrógeno entre el hidrógeno del grupo amino del enlace peptídico de un aminoácido y el grupo carboxilo del enlace peptídico de otro aminoácido [5]. Una hélice alfa puede estar dirigida hacia la derecha (sentido antihorario) o hacia la izquierda (sentido horario), siendo la más común hacia la derecha. Si un lado de la hélice contiene aminoácidos hidrofílicos, y el otro hidrofóbicos, la hélice alfa tiene propiedades anfipáticas (contiene porciones polares y no polares en la estructura). Entre los tipos de estructuras locales en las proteínas, la hélice alfa es la más regular y prevalente, así como la más predecible a partir de la secuencia de aminoácidos [4].

Lámina Beta: es otro tipo de estructura secundaria común, que se caracteriza por tener una forma aplanada y extendida; está compuesta de varias cadenas peptídicas que permanecen enfrentadas y se mantienen juntas por medio de puentes de hidrógeno en los dipolos C=O y N-H [5]. La formación de puentes de hidrógeno en las estructuras laminares puede ocurrir entre dos o más segmentos de la misma cadena o entre dos o más segmentos de distintas cadenas. De acuerdo con lo anterior se pueden presentar dos formas laminares, dada la orientación de las diferentes cadenas o segmentos; si estos se orientan en la misma dirección de una terminal a otra, la disposición recibe el nombre de lámina beta paralela; en caso contrario, si están en direcciones opuestas, la disposición es una lámina beta antiparalela. Aunque ambas ocurren en la naturaleza, la orientación antiparalela es un poco más estable porque los dipolos mencionados previamente se encuentran mejor orientados para tener una interacción óptima [7]. Una lámina beta se caracteriza por tener una longitud de 5 a 10 aminoácidos, por estar compuesta de dos o más láminas y porque los residuos de dos aminoácidos vecinos se encuentran en direcciones opuestas [4].

Estructuras aleatorias: cuando un segmento de una cadena polipeptídica carece de una combinación repetitiva, es decir, si no hay un patrón geométrico repetitivo se dice que es aleatoria. Las interacciones en las que intervienen los grupos R y las limitaciones impuestas a la posible presencia de uno o más puentes disulfuro impiden la formación de motivos como las hélices o las láminas [4].

Figura 1-5: Estructuras secundarias de las proteínas.



a) Hélice alfa b) Lámina beta (imágenes adaptadas de [4])

Estructura terciaria

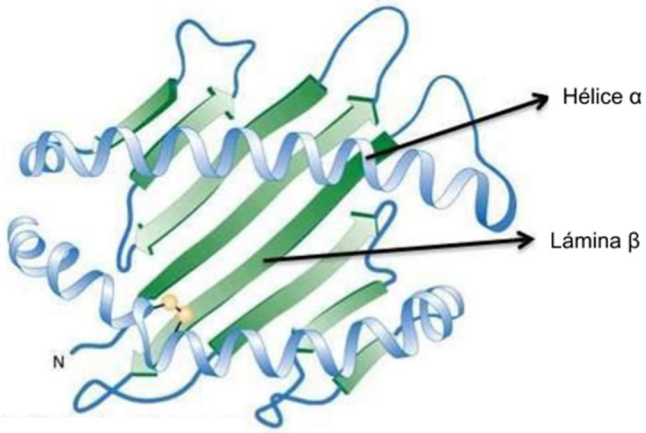
Describe la organización espacial de la proteína a un nivel superior a la estructura secundaria. También se define como la distribución espacial global y las interrelaciones de las cadenas plegadas de cada uno de los residuos de los aminoácidos de una cadena de polipéptidos. Se caracteriza por su alta compacidad y no tener vacíos en el interior de la molécula, así mismo, porque las cadenas laterales hidrofílicas se encuentran en la superficie de la molécula e interactúan con moléculas de agua; mientras que los grupos hidrofóbicos se hallan ocultos en el interior de la proteína [4].

La estructura terciaria se precisa como el plegamiento de los elementos estructurales secundarios, es decir, las posibles combinaciones de hélices alfa y láminas beta que se relacionan formando unidades globulares enlazadas de forma compacta, denominadas dominio proteico. En la Figura 1-6 se expone la estructura tridimensional, mientras que en la Figura 1-7 se muestran las interacciones fisicoquímicas presentes en esta estructura [5].

Los dominios son formados a partir de una región de la cadena polipeptídica con una longitud cercana a 50 y 350 aminoácidos. Una proteína está constituida por uno o más dominios. Existe un número limitado de combinaciones entre hélices alfa y láminas beta

para formar una estructura globular, varias de ellas se denominan motivos, ya que se presentan repetidamente en el núcleo de muchas proteínas no relacionadas entre sí [5].

Figura 1-6: Estructura terciaria de las proteínas.



(imagen adaptada de [10])

La estructura tridimensional de las proteínas es el resultado de un equilibrio entre fuerzas contrapuestas potentes, entre ellas se identifican las interacciones electrostáticas, las fuerzas de Van der Waals y los puentes de hidrógeno, que se explican a continuación:

Fuerzas electrostáticas: las moléculas son conjuntos de partículas con carga eléctrica, por tanto se utiliza la ley de Coulomb para medir la energía de asociación entre dos cargas eléctricas. Los efectos electrostáticos son uno de los factores más relevantes en las conformaciones de las proteínas, sin embargo, el cálculo de estas fuerzas está sujeta a errores debido a que su alcance es muy alto (atómico) [11].

Las interacciones electrostáticas ocurren cuando el exceso de carga negativa en una región es neutralizado por cargas positivas en otra región formando puentes salinos entre residuos de carga opuesta. Los enlaces de hidrógeno son un tipo de interacciones electrostáticas que involucran a un átomo de hidrógeno de un residuo y a un átomo de oxígeno de otro residuo. El hidrógeno con carga positiva se une parcialmente al oxígeno con carga negativa [11].

Fuerzas de Van der Waals: son asociaciones no covalentes entre moléculas neutras eléctricamente, proceden de las interacciones electrostáticas entre dipolos permanentes o dipolos inducidos. Las interacciones dipolo-dipolo son débiles pero estabilizan

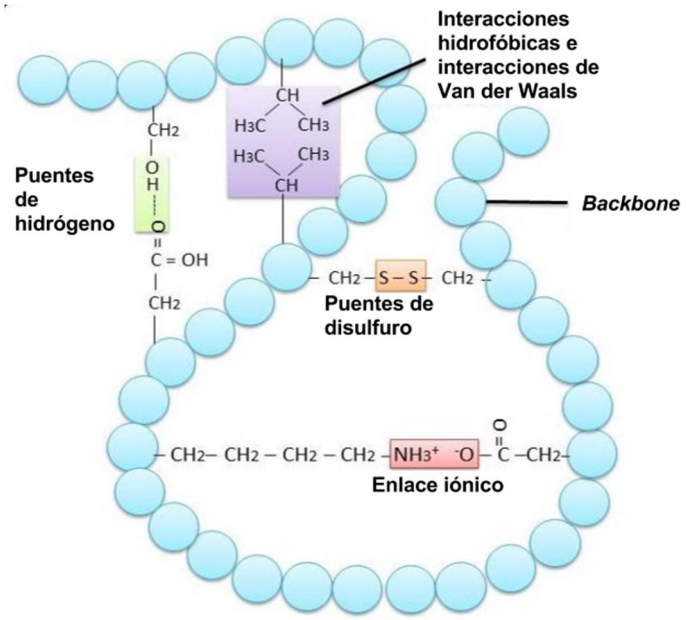
significativamente las estructuras de las proteínas. Estas fuerzas son responsables de numerosas interacciones entre átomos vecinos que no se hallan enlazados [11].

Interacciones iónicas: la asociación de dos grupos iónicos de una proteína con carga opuesta se conoce como par iónico o puente salino. Las interacciones iónicas son fuertes pero no estabilizan mucho a las proteínas, ya que los iones libres en disolución acuosa se hallan muy solvatados, de modo que la energía libre de solvatación de dos iones separados es casi igual a la energía libre de formación de su par iónico sin solvatar [11].

Puente de disulfuro: es un enlace covalente formado por dos grupos sulfidrilo (-SH), cada uno de ellos perteneciente a una cisteína. Los dos residuos que forman el puente, pueden estar separados por muchos aminoácidos en la secuencia o bien pueden pertenecer a diferentes cadenas polipeptídicas. Este enlace estabiliza la estructura tridimensional de las proteínas y se forma cuando una proteína se pliega a su conformación nativa [11].

Interacciones hidrofóbicas: son aquellas que se dan entre las cadenas laterales de los aminoácidos hidrofóbicos, estos aminoácidos suelen disponerse en el interior de la proteína, evitando de esta manera las interacciones con el agua. Las uniones hidrofóbicas suelen darse en el interior de la proteína, debido a que la mayoría de cadenas laterales puede asociarse estrechamente y se encuentran protegidas de las interacciones con el disolvente [11].

Figura 1-7: Interacciones fisicoquímicas presentes en la estructura terciaria.



Puentes de hidrógeno, interacciones hidrofóbicas, interacciones de Van der Waals, puentes de disulfuro y enlaces iónicos presentes en la estructura terciaria (imagen adaptada de [10])

Estructura cuaternaria

Es la estructura que se presenta cuando los polipéptidos se asocian con otros constituyendo subunidades de moléculas mayores, llamadas complejos proteicos, las cuales están unidas entre sí por un elevado número de débiles interacciones no covalentes [5].

1.2 Conceptos Computacionales

1.2.1 Complejidad computacional

Es la medida con la que se evalúan los recursos computacionales utilizados para solucionar un problema. Estos son el tiempo de cómputo y el espacio de memoria. En la mayoría de casos, el tiempo de cómputo es más alto que el espacio de memoria utilizado para solucionar problemas de bioinformática. El tiempo de cómputo es una aproximación a través de una función matemática en la cual se da una tendencia de su comportamiento en el tiempo. La complejidad computacional puede ser dividida en varias clases: los problemas que pueden ser resueltos en un tiempo polinomial (clase P),

aquellos problemas que son no deterministas y pueden ser verificados en tiempo polinomial (clase NP), es decir, P es una subclase de NP; y, por último, se encuentra la clase co-NP, es el conjunto de problemas complementarios a los de la clase NP, es decir, aquellos en que sus respuestas positivas o negativas se encuentran invertidas y del mismo modo son resueltos en un tiempo polinomial. Uno de los conceptos más importantes es la noción de completitud, cuando un problema es completo con respecto a una clase significa que el problema pertenece a los más difíciles de solucionar en la clase, ya que se considera que su tamaño es enorme. Casi todos los problemas de NP que no admiten tiempo polinomial se han demostrado de ser NP-completos, con algunas excepciones. También es considerado NP-completo cuando un algoritmo es NP y NP-*hard* (problemas que pueden ser reducidos a un problema NP en un tiempo polinomial) [12].

La predicción de la estructura de las proteínas es un problema NP completo, por tanto, la búsqueda exhaustiva no es factible. De este modo, las técnicas de búsqueda no determinísticas como la simulación de Monte Carlo, templado simulado, búsqueda tabú, algoritmos genéticos, colonia de hormigas, entre otras, han sido utilizadas para explorar el espacio de conformaciones [13].

1.2.2 Optimización multiobjetivo

La optimización multiobjetivo o multicriterio es aquella en la cual se busca optimizar un conjunto de funciones objetivo simultáneamente, es decir, encontrar una solución con los mejores valores para todas las funciones objetivo. Un problema de optimización se caracteriza por tener restricciones impuestas por la disponibilidad de un recurso y por el entorno, entre otras; las cuales deben ser satisfechas para que una solución pueda ser considerada aceptable. Una solución es el óptimo de Pareto si no existe otra solución factible, es decir, aquella que satisfaga las restricciones del problema sin afectar otros criterios de decisión [14].

1.2.3 Algoritmos de búsqueda

Son tipos especiales de algoritmos capaces de explorar el espacio de soluciones de un problema con el propósito de encontrar la solución óptima de acuerdo con unos criterios establecidos. Matemáticamente consisten en optimizar una función, es decir, encontrar

su máximo o mínimo global de acuerdo con las necesidades del problema. Se distinguen dos tipos de estrategias de búsqueda, estas son: la búsqueda informada y la búsqueda no informada. La primera es aquella que tiene información acerca de los posibles estados/ o solución, a diferencia de la segunda que es la búsqueda exhaustiva que explora todas las posibles soluciones para determinar cuál es la solución con mejor puntuación (no informada). Los métodos de búsqueda están compuestos por un conjunto de soluciones, una función de puntuación y un criterio de parada o finalización [15].

La búsqueda no determinística se caracteriza por producir distintas soluciones a partir de una misma entrada de ejecución; este tipo de búsqueda es utilizada para encontrar una solución aproximada, en caso de que sea muy costoso encontrar la solución exacta a través de una búsqueda exhaustiva o determinística [15].

Una búsqueda heurística busca reducir el espacio de búsqueda para encontrar una solución. También se caracteriza por no garantizar que se encuentre la mejor solución y por tener una función de evaluación que permite conocer si la solución es próxima a la óptima, es decir, se encarga de guiar el proceso de búsqueda [15].

Estos algoritmos, en general, están compuestos por cuatro pasos: el primero se refiere a la generación de una solución aleatoria, en segundo lugar a la obtención de una nueva solución del conjunto de soluciones usando un operador definido, luego, en el tercer paso se lleva a cabo la evaluación de la solución escogida en el paso anterior. Posteriormente, si el valor de la función de puntuación es mejor que el de la solución anterior se establece esa solución como la actual. Finalmente, se realizan varios ciclos o iteraciones de los pasos anteriores, sin incluir la fase inicial, hasta cuando una solución obtenga el valor óptimo en la función de puntuación [13].

Búsqueda dispersa

Una búsqueda dispersa (Scatter Search -SS) es un método evolutivo que se ha utilizado en la resolución de problemas de optimización. Se basa en el principio de que la información sobre la calidad o el atractivo de un conjunto de reglas, restricciones o soluciones pueden ser utilizadas mediante la combinación de estas. De manera similar a los algoritmos genéticos mantiene una población con la que realiza combinaciones entre sus individuos, sin embargo, no está fundamentado en la aleatorización sobre un conjunto relativamente grande de soluciones, mientras que la búsqueda dispersa realiza

las elecciones sistemáticas y estratégicas sobre un conjunto pequeño. También se caracteriza por realizar una exploración sistemática sobre una serie de buenas soluciones llamadas conjuntos de referencia.

Los algoritmos de búsqueda dispersa tienen tres fases: la primera es inicializar el conjunto de soluciones mediante un proceso heurístico; la segunda es la generación de nuevos individuos a través de combinaciones lineales; y, finalmente, la extracción de los mejores individuos de la población [16]. La búsqueda dispersa ha sido utilizada con éxito para resolver problemas de optimización mono-objetivo, sin embargo, se han propuesto algoritmos multiobjetivo basados en esta técnica como AbYSS (*Archive-based hYbrid Scatter Search*) [17].

1.3 Plegamiento de las proteínas

El plegamiento de las proteínas es el proceso por el cual una proteína alcanza su estructura tridimensional, la cual determina su función. Si una proteína se pliega incorrectamente, ésta no será funcional. Desde principios de 1950 se ha intentado entender la estructura y la función de las proteínas, por tanto, es importante caracterizarlas. El primer avance en este tema consistió en la identificación de la secuencia de aminoácidos de las proteínas, lo cual se obtuvo en 1955 por Frederick Sanger. La relevancia de este avance radica en que la secuencia de aminoácidos de una proteína determina sus propiedades químicas, sin embargo, se desconocían los factores que determinaban la estructura de las proteínas. Posteriormente, Christian Anfinsen estudió la estructura terciaria de las proteínas y observó que una cadena de polipéptidos se plegaba espontáneamente en una única cadena tridimensional, que posteriormente fue llamada conformación nativa de las proteínas; esto es enunciado en la hipótesis de Anfinsen [18].

1.3.1 La hipótesis de Anfinsen

Los estudios de Anfinsen en la teoría del plegamiento de las proteínas lo llevaron a enunciar el paradigma como: “La conformación nativa es determinada por la totalidad de interacciones interatómicas y, por lo tanto, de su secuencia de aminoácidos, en un ambiente particular” [18], también llamada la hipótesis termodinámica. Lo anterior fue

deducido del experimento realizado sobre la proteína Ribonucleasa A, que se caracteriza por tener una secuencia con 124 aminoácidos y cuatro puentes de disulfuro entre ellos. Inicialmente se redujeron los puentes de disulfuro a ocho grupos sulfhidrilo (-SH), luego se desnaturalizó con el componente orgánico urea con concentración 8 M (molar). Bajo esas condiciones su función era inactiva con una estructura aleatoria y flexible. La siguiente fase del experimento consistió en remover lentamente la urea, por lo que los grupos sulfhidrilo se oxidan para volver a formar puentes de disulfuro, es decir, volvió a su estado inicial. Lo anterior demostró que la proteína recuperó espontáneamente su función y, por tanto, su estructura tridimensional [19], la cual es única, estable y cinéticamente accesible al mínimo de energía libre.

La teoría de Anfinsen se estableció como la base de estudio para realizar la predicción de la estructura tridimensional de las proteínas a partir de su secuencia de aminoácidos, ya que toda la información necesaria para predecir la conformación nativa se encuentra en esta secuencia [18].

1.3.2 La paradoja de Levinthal

La flexibilidad en las proteínas se debe a los grados de libertad de los ángulos de torsión presentes en una proteína; por tanto, son múltiples las conformaciones que se pueden generar mediante la combinación de estos. En 1968 se postuló la paradoja de Levinthal, que consiste en encontrar la estructura nativa de una proteína a través de una búsqueda aleatoria entre todas las posibles configuraciones, lo que puede tomar mucho tiempo. Sin embargo, las proteínas se pliegan en segundos o menos tiempo [18]. Por ejemplo, una secuencia de 101 aminoácidos tiene 100 enlaces peptídicos y, por cada uno, tres grados de libertad, es decir que se pueden generar 3^{100} conformaciones. Si la proteína es capaz de explorar las conformaciones en 10^{13} segundos, entonces el tiempo de explorar todo el espacio de conformaciones demoraría 10^{27} años, tiempo que gastaría la proteína en encontrar la estructura nativa haciendo una búsqueda exhaustiva por el espacio de conformaciones [20].

Levinthal también sugirió que la estructura estable no se encuentra necesariamente en el mínimo de energía absoluto, en caso de que ésta no sea cinéticamente accesible [21]. Ante esta situación se han propuesto alternativas de resolución a la paradoja que se explican brevemente a continuación.

Modelos clásicos de plegamiento proteico

Son modelos basados en la existencia de rutas de plegamiento, es decir, supone que todas las moléculas siguen el mismo camino pasando en su caso por los mismos intermediarios (mínimos locales de energía libre de Gibbs) y por el estado de transición (conformación de alta energía por la que atraviesa el polipéptido para plegarse). Este tipo de solución plantea los siguientes modelos [22]:

El modelo de nucleación - propagación: se caracteriza por la formación de uno o más núcleos estructurales, alrededor de los cuales se propaga la estructura.

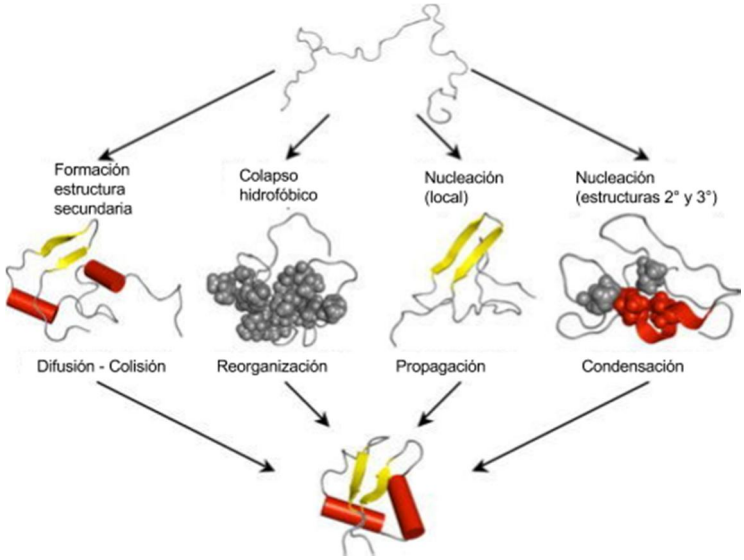
El modelo de armazón (*framework*): posterior a la creación de núcleos, se forman estructuras secundarias que a su vez se asociarán para formar estructuras supra-secundarias, y de este modo formar la estructura terciaria.

El modelo de difusión – colisión: la nucleación sucede simultáneamente en partes diferentes de la cadena polipeptídica generando microestructuras que tienen un tiempo de vida controlado por la difusión de un segmento.

El modelo del colapso hidrofóbico: inicialmente la proteína colapsa a través de interacciones hidrofóbicas de largo alcance y sucede previa a la formación de una estructura secundaria.

Una manera de superar la paradoja de Levinthal sería mediante la existencia de una ruta definida de plegamiento que pasara obligatoriamente por determinados estados (intermedios), donde cada uno debería tener menor energía que el estado inmediatamente anterior, y así hasta llegar al estado nativo. Algunos de esos estados intermedios presentan propiedades del estado nativo y del estado desnaturalizado, aquellos son denominados como el estado de glóbulo fundido; se caracterizan por tener una elevada cantidad de estructuras secundarias, pero que no tienen interacciones de un rango mayor (estructura terciaria) [21].

Figura 1-8: Modelos clásicos de plegamiento de proteínas.

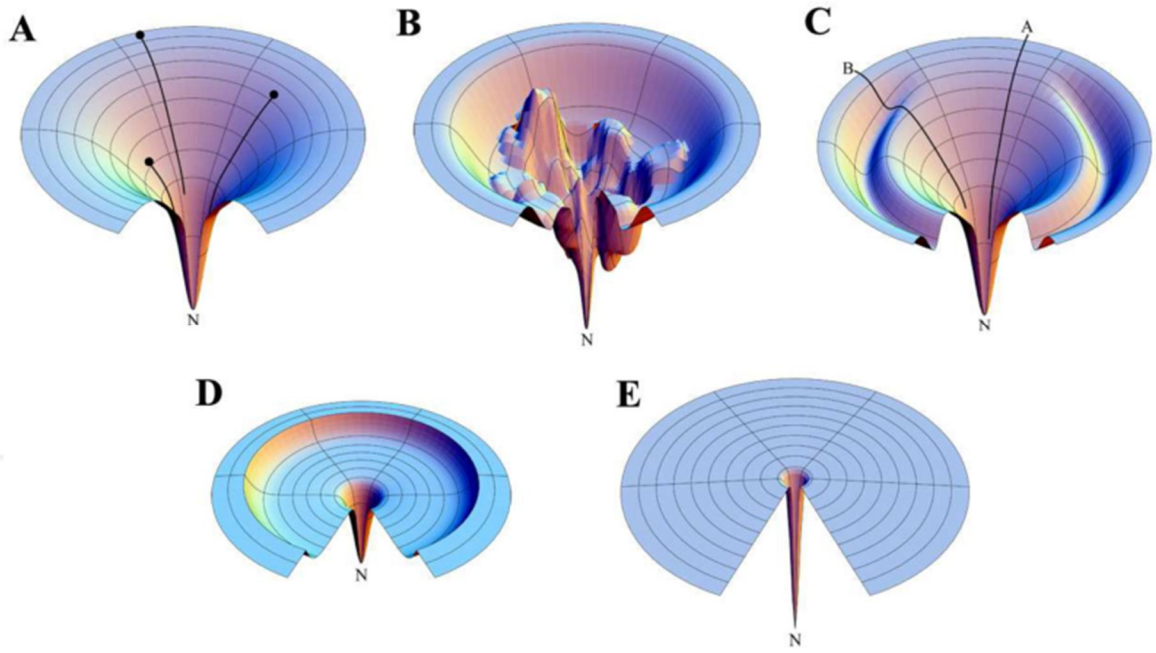


Modelos de difusión-colisión, colapso hidrofóbico, nucleación-propagación y nucleación - condensación. (imagen adaptada de [23])

La nueva visión (*energy landscape*) - *Folding Funnel*

Se centra en el paisaje de energía del plegamiento de las proteínas, el cual se comporta como un embudo, como se muestra en la Figura 1-9. El embudo expone el comportamiento termodinámico y cinético de las moléculas del plegamiento hasta alcanzar la estructura nativa. El ancho del embudo está relacionado con la entropía configuracional de la cadena polipeptídica, mientras que la profundidad representa una función de energía libre que no incluye los grados de libertad internos de la proteína. El embudo implica que en el plegamiento de las proteínas disminuye la energía libre y la entropía de la estructura [24]. Es considerado como el modelo más cercano a la realidad del plegamiento ya que abarca el espacio de conformaciones, lo que incluye varios estados como el nativo, no plegado, desnaturalizado, conformaciones cercanas o lejanas a este, e intermediarios como el estado de glóbulo fundido. El embudo de energía se caracteriza por presentar diferentes formas que son conducidas por la termodinámica que lleva a la formación de estados, y por la cinética que define las barreras de energía entre los estados de conformaciones y las transiciones entre ellos [25].

Figura 1-9: Embudo de energía del plegamiento de las proteínas.



El embudo A representa el paisaje de energía ideal en el cual la proteína se pliega a gran velocidad; el embudo B presenta colinas y valles de energía en los que se encuentran mínimos y máximos locales; el embudo C presenta dos posibles caminos de plegamiento, uno lento debido a una barrera energética y otro rápido sin obstáculos; el embudo D exhibe como la entropía conformacional puede generar barreras de energía a través de una caminata aleatoria por la meseta plana. Finalmente, el embudo E muestra la búsqueda aleatoria de Levinthal la cual demora en encontrar el estado nativo. (imagen tomada de [25])

2. Métodos computacionales para la predicción de la estructura de las proteínas

La predicción de la estructura terciaria de las proteínas ha sido un problema que se ha tratado de solucionar a través de métodos experimentales como la cristalografía de rayos X y la resonancia magnética nuclear, y de métodos computacionales, los cuales son objetivo de estudio en este capítulo.

Los métodos computacionales se dividen en dos, los que utilizan información previa para realizar la predicción de la estructura tridimensional, también llamado modelado basado en plantillas y aquellos que predicen la estructura terciara sin información previa o modelado libre (*ab initio)*.

2.1 Modelado basado en plantillas

2.1.1 Modelado comparativo

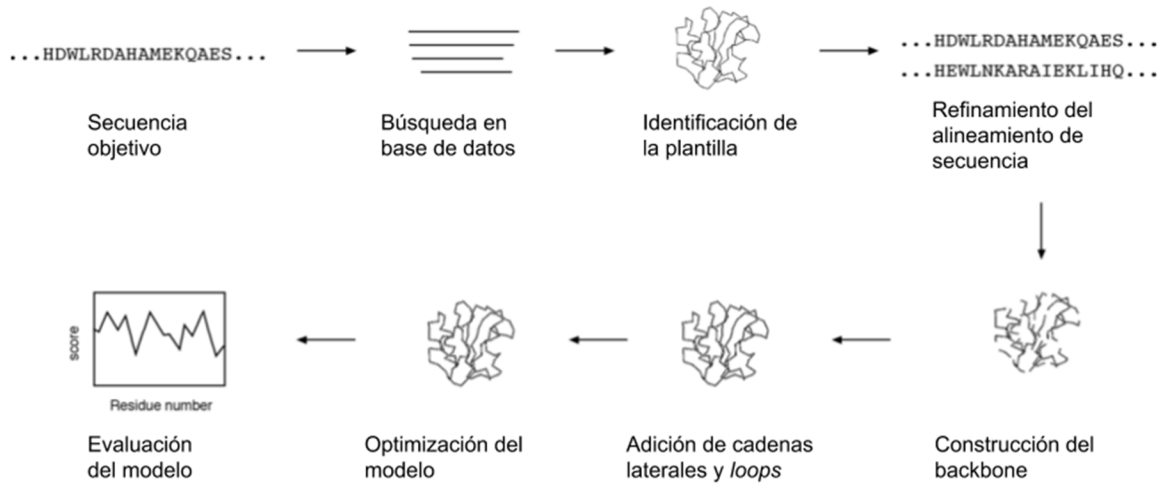
El modelado comparativo, también llamado modelado por homología, consiste en construir un modelo de resolución atómica de una proteína objetivo a partir de su secuencia de aminoácidos y de la estructura tridimensional de una proteína plantilla, de la cual se conoce su estructura terciaria a través de métodos experimentales. Se fundamenta en la siguiente afirmación: “las proteínas que tienen secuencia similar usualmente tienen estructura similar debido a que evolutivamente provienen de un antecesor común”, y en que la estructura de las proteínas se conserva más que sus secuencias [18] [26]. Las proteínas que son descendientes de un antecesor común pueden adoptar formas distintas, sin embargo algunas regiones de la secuencia de la proteína se conservan y del mismo modo la estructura en esas regiones [18].

Las proteínas se pueden agrupar en familias, es decir, un grupo de proteínas que comparten un origen evolutivo común, que se refleja en sus funciones y en la similitud de secuencia o estructura. Estas se clasifican en una superfamilia cuando el grupo de

proteínas es distante evolutivamente, pero mantiene una relación entre sí; mientras que una subfamilia es un grupo pequeño de proteínas estrechamente relacionadas (evolutivamente cercanas) entre sí [27].

Los métodos basados en homologías se caracterizan por tener un conjunto de etapas durante el proceso de predicción de la estructura tridimensional de una proteína, como se observa en la Figura 2-1. A continuación se explica cada una de ellas.

Figura 2-1: Pasos del modelamiento por homologías.

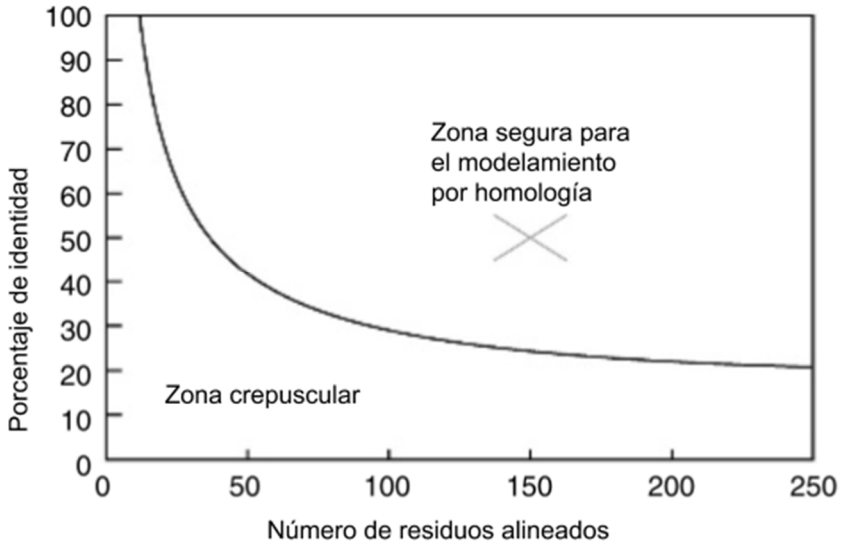


(imagen adaptada de [28])

1. Selección de la proteína plantilla y alineamiento inicial.

Constituye la base del modelamiento por homología, que consiste en la selección de una proteína plantilla, lo cual involucra la búsqueda de proteínas homólogas con estructuras determinadas en el *Protein Data Bank* (PDB). Esta exploración se puede realizar a través de un software de búsqueda de alineamiento por pares como BLAST o FASTA [28]. El resultado de los alineamientos muestran un porcentaje de similitud, el cual si es menor a 30% indica que encontrar una homología no es confiable [18]. Ocasionalmente, un 20% de identidad puede ser usado como umbral, siempre que el porcentaje de identidad de la secuencia se encuentre en la “zona segura”, es decir, que el porcentaje de identidad supera el umbral establecido para encontrar homologías [3], como se muestra en la Figura 2-2.

Figura 2-2: Zonas del alineamiento de secuencias.



Las zonas del alineamiento de secuencias. La zona oscura o crepuscular (*twilight*) representa aquella en la cual se encuentran homologías con dificultad, a diferencia de la zona segura (imagen adaptada de [28])

Los resultados de un alineamiento pueden mostrar varias estructuras que superan el umbral de homología establecido, en este caso se recomienda elegir como plantilla las estructuras con el mayor porcentaje de identidad, con mayor resolución y con los cofactores más apropiados. Por otro lado, se puede dar el caso en el que no se encuentra similitud de secuencia con la base de datos de estructuras, lo que dificulta la selección de plantillas. Para detección de homólogos remotos se utiliza un método de búsqueda más sensible basado en perfiles como PSI-BLAST o un modelado por reconocimiento del plegamiento (*threading*) [28].

Para elegir la plantilla apropiada se recomienda que esta tenga alta similitud de secuencia y que al realizar el alineamiento con la secuencia objetivo se encuentre la menor cantidad de huecos (*gaps*), es decir, inserciones o deleciones, ya que estos tienen un impacto alto en la construcción del modelo final. Por lo tanto, es mejor elegir un alineamiento que tenga pocos huecos e inserciones cortas [18].

2. Alineamiento de la secuencia objetivo

Una vez establecida la plantilla, las secuencias de las proteínas plantilla y de la proteína objetivo deben ser realineadas usando algoritmos de alineamiento refinados para obtener un alineamiento óptimo. Este reajuste es el paso más relevante en el modelado de homología, ya que afecta directamente la calidad del modelo final, debido a que una alineación incorrecta conduce a la asignación incorrecta de homólogos y, por lo tanto, a modelos estructurales incorrectos. Para llevar a cabo esta etapa se utilizan algoritmos de alineamiento múltiple como PRALINE y T-Coffee. Sin embargo, estas herramientas de *software* de alineación están sujetos a errores y deben ser supervisados manualmente para inspeccionar que los principales residuos conservados están correctamente alineados [28].

3. Construcción del esqueleto de la proteína (*backbone*)

Después de lograr una alineación óptima, los residuos de las regiones alineadas de la proteína objetivo pueden asumir una estructura similar a la de las proteínas plantilla, lo que significa que las coordenadas de los residuos correspondientes con las proteínas de la plantilla se pueden copiar en la proteína objetivo. Si los dos residuos alineados son idénticos, las coordenadas de los átomos de la cadena lateral se copian junto con los átomos de la cadena principal. Si los dos residuos son diferentes, solo los átomos del esqueleto se pueden copiar. En el modelado del esqueleto, se suele utilizar solamente una plantilla de estructura [28].

4. Modelamiento de los *loops*

Después de realizar un alineamiento de secuencias, usualmente se generan regiones con gaps o huecos producidos por inserciones o deleciones. Los huecos en el alineamiento producen los ciclos o *loops*, que deben ser modelados con precisión ya que puede ser una fuente de error para el modelo final. En la actualidad no se pueden modelar los ciclos confiablemente. Sin embargo, se distinguen dos técnicas que buscan la solución a este problema: los métodos de búsqueda en base de datos y los métodos *ab initio*. Los primeros métodos consisten en encontrar "piezas de recambio" de proteínas con estructuras conocidas en una base de datos, para ajustarlas a la región del *loop*, la cual se encuentra entre los tallos (*stem*) de la proteína objetivo. Los tallos son la cadena principal de átomos que preceden y siguen el *loop* a modelar. El procedimiento inicia con la medición de la orientación y la distancia de las regiones de anclaje de los tallos y la

búsqueda en el PDB para los segmentos de la misma longitud que también coinciden con el punto final de la conformación. Por lo general, muchos segmentos tienen varias alternativas que se ajustan a los criterios de valoración de los tallos que están disponibles. La mejor selección se puede hacer con base en la similitud de secuencia, así como en la menor cantidad de choques estéricos con las partes vecinas de la estructura. La conformación con los mejores fragmentos coincidentes se copia en los puntos de anclaje de los tallos.

Mientras que el segundo método, *ab initio*, genera bucles y realiza búsquedas aleatorias para hallar uno que no choque con cadenas laterales cercanas, que tenga energía baja y sus ángulos φ y ψ se encuentren en regiones permitidas, según la gráfica de Ramachandran [28].

5. Modelado de las cadenas laterales

Una cadena lateral se puede construir mediante la búsqueda de cada conformación posible según la combinación de todos los ángulos de torsión de la cadena lateral y se seleccionaría el de menor energía. Sin embargo, este enfoque es computacionalmente costoso y, por tanto, se utilizan los rotámeros, que son los posibles ángulos de torsión de la cadena lateral extraídos de proteínas con estructuras conocidas. Una colección de conformaciones de la cadena lateral es una biblioteca de rotámeros, en la que los rotámeros se clasifican por su frecuencia de ocurrencia. Tener una biblioteca de rotámeros reduce el tiempo computacional, ya que se utilizan los ángulos de torsión reportados [28].

6. Refinamiento y evaluación del modelo

Al generar un modelo por homología no se puede garantizar que esté libre de irregularidades estructurales, que se pueden corregir en caso de presentarse mediante la minimización de energía del modelo completo, ya que se sustituyen las colisiones estéricas sin alterar la estructura global. Sin embargo, se debe tener en cuenta que una minimización excesiva de la energía puede desplazar los residuos de sus posiciones correctas y, por consiguiente, se recomienda usar un número limitado de iteraciones [28].

Finalmente, se evalúa que las características estructurales del modelo sean consistentes con las reglas fisicoquímicas, las cuales involucran la verificación de anomalías en los

ángulos de torsión y en la longitud de los enlaces, entre otras [28]. La evaluación del modelo constituye un paso crítico en el modelado por homología, ya que se utiliza una función de puntuación que debe distinguir entre un modelo bueno y uno malo. Las funciones de puntuación se clasifican en estadísticas (basadas en las propiedades de los aminoácidos de proteínas con estructuras conocidas) y en funciones energéticas (se fundamentan en la cantidad de energía libre en la proteína) [18].

Existen cuatro métodos para la construcción del modelo y se distinguen de acuerdo con la forma en que se transfiere la información de las estructuras conocidas a la secuencia de la proteína objetivo; entre ellos se encuentran: modelado por ensamblaje de cuerpos rígidos, modelado por armonización de segmentos y modelado por satisfacción de restricciones espaciales [29].

- Modelado por ensamblaje de cuerpos rígidos: es el método más simple y utilizado. Inicia con la identificación de las regiones conservadas y de las regiones variables de las plantillas, a través de la superposición de estructuras. Las regiones conservadas se reconocen fácilmente a través de un alineamiento múltiple de estructuras, es decir, la distancia entre los fragmentos (RMSD: *Root Mean Square Distance*) es relativamente pequeña, y las regiones variables se encuentran normalmente en *loops* con huecos frecuentes en la alineación estructural. En las regiones no conservadas, o *loops*, la región es construida a través de un enfoque *ab initio* o mediante la búsqueda en una base de datos de estructuras que se adapten a las regiones y que del mismo modo tengan una secuencia compatible. Las cadenas laterales son modeladas de acuerdo con preferencias conformacionales y en las conformaciones de las cadenas laterales equivalentes en la estructura de la plantilla. Para secuencias con homólogos de alta identidad, una sola plantilla es suficiente; pero para plantillas con homólogos de baja identidad, se suele utilizar la media ponderada a través de múltiples plantillas, lo cual resulta más confiable [18].
- Armonización de segmentos: es un método basado en encontrar la mayor cantidad de hexapéptidos que pueden ser agrupados en 100 clases estructurales. El modelo de homología se construye de acuerdo a posiciones aproximadas de átomos conservados de las plantillas como "posiciones guías" para calcular las coordenadas de otros átomos. Las posiciones guía generalmente corresponden a los átomos de

los segmentos que se conservan en el alineamiento de las estructuras de la plantilla y de la secuencia objetivo [18].

- Modelado por satisfacción de restricciones espaciales: consiste en la generación de restricciones para la proteína objetivo basadas en el alineamiento estructural con la proteína plantilla. Las restricciones son obtenidas al suponer que la distancia entre dos residuos del modelo objetivo es similar a la distancia de dos residuos alineados en la plantilla. Las restricciones son complementadas con restricciones estereoquímicas en los ángulos de enlace, la longitud del enlace y los enlaces peptídicos, entre otros. Para homologías débiles, se adicionan restricciones de experimentos (si están disponibles) para incrementar la precisión del modelo. Del mismo modo, se establece un intervalo de valores para cada restricción, que pueden ser estimados por información estereoquímica o mediante un análisis estadístico de las relaciones entre las estructuras de las proteínas similares [18].

  Mediante la exploración de un conjunto de estructuras relacionadas entre sí, se pueden cuantificar correlaciones como las distancias o los ángulos diedros de la cadena principal. Estas relaciones se expresan como funciones de densidad de probabilidad condicional y se pueden utilizar como restricciones espaciales. Las probabilidades para las distancias equivalentes y los ángulos diedros de la cadena principal se calculan a partir del tipo de residuo, desde la conformación de la cadena principal de un residuo equivalente y la similitud de secuencia entre las dos proteínas [18].

  Finalmente, las restricciones espaciales y los campos de fuerza se combinan en una función objetivo que debe cumplir con las condiciones estereoquímicas adecuadas. El modelo se obtiene mediante la optimización (minimizando las restricciones) de la función de energía. La ventaja de este método radica en el uso de varios tipos de información de la secuencia objetivo, incluyendo la estructura secundaria. Sin embargo, para las secuencias altamente homólogas, la información es almacenada en las estructuras de la plantilla, mientras que introducir información derivada de otros miembros de la familia puede degradar el modelo [18].

2.1.2 Reconocimiento del plegado (*Threading*)

Es un método de predicción de la estructura tridimensional de las proteínas que se fundamenta en la observación de que hay pliegues de proteínas que se presentan en varias familias en las que no hay una relación evolutiva cercana. Las posibles explicaciones a este fenómeno se exponen a continuación [3]:

- Evolución divergente: Las proteínas tienen un plegamiento similar y están relacionadas entre sí. Los algoritmos como PSI-BLAST, los modelos ocultos de Markov (HMM) y los alineamientos perfil - perfil son utilizados para identificar homologías remotas.
- Evolución convergente: proteínas con funciones comunes usualmente poseen estructuras terciarias similares.
- Número limitado de plegamientos: proteínas no relacionadas evolutivamente tienen pliegues similares debido a que el espacio de posibles pliegues es pequeño.
- Análisis erróneo: Una aparente similitud estructural puede ser el resultado de deficiencias en las herramientas de análisis y no por una similitud real entre las estructuras de las proteínas.

Este método se caracteriza por usar una biblioteca de estructuras únicas, en la cual se buscan estructuras análogas para la secuencia objetivo y se fundamenta en la teoría de la existencia de un número limitado de pliegues en las proteínas [30].

El reconocimiento del plegado consiste en colocar los aminoácidos de la secuencia de la proteína objetivo, siguiendo su orden secuencial y permitiendo huecos, en las posiciones estructurales de la estructura de la proteína plantilla de una manera óptima, de acuerdo con unos puntajes de calidad asignados. Este procedimiento se repite contra una colección de estructuras tridimensionales de proteínas previamente resueltas para una proteína objetivo.

Los alineamientos secuencia-estructura se evalúan utilizando medidas estadísticas o energéticas para determinar la probabilidad global de los pliegues estructurales que la proteína objetivo pueda adoptar. El mejor alineamiento de secuencia-estructura permite establecer los átomos de la cadena principal de la proteína objetivo, de acuerdo con sus ubicaciones en la estructura de la proteína plantilla [12].

Los pasos que se realizan para predecir una estructura son similares a los que usa el modelado comparativo o por homología, y se diferencia en la etapa de identificación de plegamientos. En primer lugar debe estar definida una biblioteca de estructuras, la cual puede incluir: cadenas, dominios, o núcleos de proteínas conservadas. Después la secuencia objetivo se ajusta a cada entrada de la biblioteca y se utiliza una función de energía para evaluar la concordancia entre la secuencia objetivo y las entradas de la biblioteca con el propósito de determinar las mejores plantillas. En general, se utilizan cuatro tipos de métodos para predecir la estructura de proteína por *threading*, estos son: (1) los métodos de reconocimiento del plegamiento que utilizan el entorno de cada residuo en la estructura como la función de energía y la programación dinámica para evaluar el ajuste y la alineación; (2) métodos que usan estadísticas derivadas de la interacción de potenciales entre pares de residuos o pares de átomos que se pueden usar para evaluar los mejores ajustes posibles entre la secuencia objetivo y los pliegues de la biblioteca; (3) son métodos que no utilizan una función de energía explícita, en lugar de eso emplean estructuras secundarias y la accesibilidad a cada residuo, las cuales son predichas anteriormente; la secuencia objetivo y los pliegues de la biblioteca son codificados en cadenas para el alineamiento secuencia-estructura; y (4) pertenecen aquellos métodos que combinan la similitud de secuencia con el reconocimiento del plegamiento, en estos se dispone la similitud de secuencia de los alineamientos iniciales, que son evaluados por los métodos de *threading* [30]. De estos métodos se identifica la comparación secuencia-secuencia, que detecte homólogos cercanos; la comparación perfil-secuencia o secuencia-perfil, para encontrar motivos conservados, y la comparación perfil-perfil que tiene una alta sensibilidad para encontrar similitudes en la zona crepuscular.

2.2 Modelado libre

El modelado libre o *ab initio* es un modelo de predicción de la estructura tridimensional de las proteínas que no utiliza información previa, a diferencia de los métodos basados en plantillas. La única información que utiliza este tipo de modelado es la obtenida en la secuencia de aminoácidos, ya que ésta determina la estructura terciaria. Este tipo de modelado se basa en estudios biofísicos que determinaron que la mayoría de proteínas se pliegan en una estructura estable en la cual se encuentra el mínimo de energía libre y

se denomina el estado nativo de la proteína [28]. Los modelos *ab initio* se fundamentan en la termodinámica y en la mecánica estadística, por lo tanto, sus métodos son simulaciones de dinámica molecular, simulaciones de Monte Carlo, entre otros [8].

El modelado *ab initio* es un reto debido a que la precisión de las funciones de potencial es limitada, ya que se puede hallar una conformación energéticamente baja pero biológicamente inactiva; también, el espacio de conformaciones es gigantesco y, por tanto, el costo computacional es alto. Sin embargo, varios métodos han utilizado representaciones reducidas, potenciales simplificados y diversas estrategias de búsqueda ante los mencionados inconvenientes [3].

Se identifican dos tipos de métodos en el modelado *ab initio*, los primeros son aquellos que utilizan potenciales físicos para dirigir el plegamiento y son llamados *ab initio* estrictos; mientras que los segundos se caracterizan por obtener información previa del PDB para generar una restricción en el espacio conformacional, estos son nombrados como métodos mixtos, como es el caso del método Rosetta que consiste en la reducción del espacio conformacional mediante el ensamblaje de segmentos, en el cual se toman fragmentos de estructuras de proteínas conocidas para generar una biblioteca de fragmentos, con el objetivo de reemplazar los ángulos de torsión y el segmento seleccionado para generar conformaciones con base a los mencionados fragmentos. Los fragmentos pueden ser cortos, tener origen en varias fuentes y no es necesario transferir el plegamiento a la proteína objetivo. Se identifican dos fases en su construcción, la primera es la elección de fragmentos a utilizar y la segunda es la minimización de las conformaciones generadas mediante la sustitución de fragmentos. Las ventajas de utilizar estos métodos consisten en que los fragmentos pueden transferir información de una posible homología y del mismo modo el uso de fragmentos derivados de plegamientos de proteínas con estructura conocida contienen información de una estructura estable, es decir, no hay *gaps*, y posiblemente una buena geometría. La sustitución de fragmentos aleatorios reduce el espacio de búsqueda de conformaciones [12].

2.2.1 Representación geométrica de las proteínas

Los métodos *ab initio* se caracterizan por tener una forma de representar computacionalmente las proteínas. Normalmente, se han utilizado los modelos reducidos de las proteínas, que se caracterizan por ser una representación de bajo costo computacional. Estos incluyen los modelos de celda, los modelos de espacio continuo (una proteína es reducida a sus C_α y al centroide de sus cadenas laterales), y los modelos híbridos (en los cuales algunos grados de libertad de la conformación son localmente discretizados).

- Modelos basados en celdas: son aquellos que consideran dos tipos de residuos, los hidrofóbicos y los polares (hidrofílicos) de acuerdo con su aversión o afinidad con el agua. Por tanto, una proteína es una cadena de caracteres definida sobre el alfabeto H (hidrófobo) y P (polar – hidrófilo). Todos los aminoácidos de la cadena ocupan una posición en una celda cuadrada. El ángulo entre aminoácidos consecutivos es de 90° grados [31]. Esta representación supone que los residuos de los aminoácidos tienen el mismo tamaño, los residuos de los enlaces químicos tienen la misma fuerza y su ubicación es estricta en la rejilla [32].

  Este modelo se basa en el concepto de que la mayor contribución de energía libre de una conformación nativa de una proteína se debe a la interacción entre aminoácidos hidrofóbicos. Estos aminoácidos tienden a agruparse en el núcleo de la proteína plegada, dejando a los aminoácidos polares en la parte externa de la proteína, en contacto con el medio acuoso, en el caso de las proteínas globulares. La cantidad de energía libre es inversamente proporcional a la cantidad de contactos hidrófobo – hidrófobo (H-H) no locales, es decir, aminoácidos que no son adyacentes en la secuencia, pero tienen posiciones contiguas en la rejilla (2D – 3D, si es de dos o tres dimensiones), entonces para minimizar la energía libre es necesario maximizar el número de contactos H-H [31].

  La representación basada en celdas se caracteriza por tener las siguientes restricciones [32]:

  1) Cada residuo de la secuencia debe ser localizado en el sistema de coordenadas enteras.

2) La malla es finita (condición de frontera).
3) Dos residuos contiguos en la cadena de aminoácidos deben ser vecinos en la rejilla y la distancia entre los residuos es 1.
4) Los residuos no se pueden superponer.
5) Las conformaciones compactas tienen menor energía que las no compactas.
6) Diferentes tipos de aminoácidos tienden a estar separados.

La calidad de la configuración es medida a través de una función de energía, que sólo considera la fuerza hidrófoba. La energía de una conformación está definida como el número de contactos entre aminoácidos hidrófobos que no son vecinos en la secuencia dada.

Para el caso tridimensional se construye un cubo que se caracteriza porque cada residuo tiene 6 vecinos (uno por cada cara) y por estar compuesto de tres capas, que son el núcleo interno (*H-Core*), un núcleo externo y un núcleo mixto. La capa interna contiene residuos hidrofílicos, la capa externa residuos polares y la capa mixta se conforma de residuos hidrofílicos y polares unidos por enlaces covalentes [33]. Las limitaciones de esta representación radican en su incapacidad de reproducir hélices [34].

- Modelos no basados en celdas: son modelos que buscan representar los aminoácidos de una proteína con mayor detalle, como se muestra a continuación:

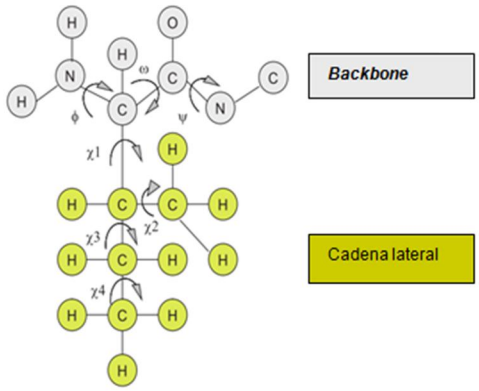
  Representación 3D del átomo: cada átomo es representado por medio de tres variables, por tanto, un aminoácido necesita aproximadamente 50 variables. Una proteína puede tener más de 50 aminoácidos, lo que dificulta trabajar con esta representación, además de que no todas las configuraciones son factibles [35].

  Representación parcial del átomo en 3D: es la misma representación que la anteriormente expuesta, pero se diferencia en que sólo se representan los principales átomos de cada aminoácido [35].

  Representación 3D de los átomos del *backbone* y los centroides de la cadena lateral: cada aminoácido tiene nueve átomos en el *backbone*, para un total de 27 coordenadas, además del centroide de la cadena lateral (tres coordenadas más) [35].

Representación del *backbone* y los ángulos de torsión de la cadena lateral: la longitud de los enlaces es un valor conocido y la única variable es el ángulo de torsión entre dos enlaces consecutivos. La representación provee enlaces factibles, y cada aminoácido contiene tres variables para el *backbone* (ω, ψ, φ); y entre cero y cuatro variables para la cadena lateral (χ_i) [35].

Figura 2-3: Representación del *backbone* y los ángulos de torsión de la cadena lateral



En color gris se presenta el *backbone* con sus respectivos ángulos de torsión, mientras que en verde se muestra la cadena lateral de la proteína (imagen adaptada de [29])

2.2.2 Función de puntuación

Es una medida que permite guiar la exploración al espacio de conformaciones y evaluar la calidad de las distintas conformaciones exploradas, con el propósito de seleccionar una de éstas [18]. Hay dos tipos de potenciales que pueden ser empleados en la evaluación de la energía libre de una cadena de péptidos y su entorno solvente: funciones basadas en campos de fuerza empíricos y funciones de energía basadas en el conocimiento.

Las funciones de energía basadas en campos de fuerza empíricos son aquellas que utilizan la mecánica molecular, que consiste en la aplicación de las leyes de la mecánica clásica o Newtoniana a nivel molecular. En la mecánica molecular las moléculas son tratadas como un conjunto de esferas (átomos), con masa m, radio r, volumen V y unidos por muelles (enlaces) con una constante K (fuerza de enlace). También se considera que el valor de energía molecular varía de acuerdo con la geometría de la molécula.

Para la predicción de la estructura tridimensional de las proteínas se han utilizado diversas funciones de energía basadas en la mecánica molecular, ya que su cálculo es sencillo y se distinguen las energías requeridas para determinar los campos de fuerza. Algunos prototipos de esos campos de fuerza son AMBER, CHARMM, y ENCAD. Los parámetros utilizados en estos campos de fuerza son obtenidos mediante el ajuste experimental de los datos [12].

La energía potencial de una proteína se expresa como la suma de energías de enlace y energías de no enlace (ver Ecuación 2-1). En otras palabras, es la variación de energía relacionada con los cambios de longitud, ángulo y torsión de los enlaces, junto con la energía resultante de las interacciones entre átomos (van der Waals) y la energía electrostática.

$$E_{Total} = E_{Enlace} + E_{Ángulo} + E_{Torsión} + E_{Electrostática} + E_{van\ der\ Waals} \quad (2\text{-}2)$$

Por otro lado, se han empleado los potenciales basados en el conocimiento, que son funciones de energía derivadas del análisis de estructuras de proteínas conocidas almacenadas en el PDB, es decir, que se originan a partir de propiedades observadas en proteínas con plegamiento conocido, las cuales no se presentan en péptidos no plegados o desplegados [12].

Se identifican dos tipos de funciones de energía basadas en conocimiento; el primero es derivado del análisis estadístico de una base de datos de estructuras de proteínas. Mientras que el segundo está basado en la optimización, en el que un conjunto de parámetros para una función potencial son optimizados de acuerdo a un criterio, maximizando los huecos (*gap*s) de energía entre la estructura nativa conocida y el conjunto de conformaciones candidatas [18].

2.2.3 Método de búsqueda

Con el desarrollo de algoritmos que permiten la exploración y explotación del espacio de conformaciones de la estructura tridimensional de las proteínas se ha tratado de solucionar este problema fundamental de la biología, para ello se han utilizado procedimientos de diferente naturaleza. Estos se pueden agrupar así: en aquellos que parten de conformaciones aleatorias y simulan el proceso de plegamiento de una

proteína; en los que realizan el ensamblaje de fragmentos; y otros que combinan los dos anteriores [18].

Dentro del grupo de métodos que realizan la simulación del proceso de plegamiento se encuentran aquellos que buscan minimizar la energía hasta encontrar el mínimo global. Entre ellos se distinguen los algoritmos evolutivos, los cuales tienen como base una población que avanza a través de generaciones sucesivas en las que los individuos evolucionan por operaciones genéticas (tales como cruce y mutación) con el objetivo de garantizar diversidad durante la exploración. Se caracterizan porque el tamaño de la población es invariable durante la ejecución del algoritmo y por seleccionar a los mejores individuos del conjunto de soluciones de acuerdo con su nivel de aptitud o *fitness* [36].

Del mismo modo se han usado los algoritmos evolutivos multiobjetivo, que buscan optimizar más de un objetivo en la función *fitness*, para ello se tiene en cuenta la eficiencia de Pareto. Se ha comprobado que su desempeño es alto debido a que el plegamiento de las proteínas relaciona más de una variable de decisión como son las energías de enlace y las energías de no enlace, entre otras, como se muestra en [37].

También se han utilizado algoritmos de inteligencia computacional como la optimización por colonia de hormigas (ACO), que es un proceso iterativo en el que una población de agentes (hormigas) construye repetidamente una solución candidata. Se caracteriza por el uso de una memoria compartida en la cual se almacena la experiencia de las hormigas (caminos visitados) en las iteraciones anteriores con el propósito de tener los rastros de feromonas. Este procedimiento ha mostrado buen rendimiento en la predicción de la estructura terciaria de las proteínas, pero depende de la función heurística y la actualización de las feromonas, la cual puede llegar a un mínimo local y mostrarlo como el óptimo global [38].

Por otro lado se encuentran los métodos que reducen el espacio de conformaciones a través de predicciones de estructuras locales y que modelan la estructura de las proteínas mediante el ensamblaje de fragmentos como Rosetta, o mediante el reconocimiento de plegamientos en el PDB para luego ser ensamblados como lo hace I-TASSER. Estos métodos se caracterizan por utilizar un método de Monte Carlo para construir los modelos tridimensionales, es decir, integrar los fragmentos para generar una

conformación [39]. Algunos de estos métodos utilizan una representación reducida inicialmente con el objetivo de disminuir el espacio de conformaciones; posteriormente, usan una representación completa (todos los átomos) y realizan agrupaciones de acuerdo a las similitudes entre las conformaciones [40].

2.3 Critical Assessment of Techniques for Protein Structure Prediction (CASP)

Es una competencia de predicción de estructuras de proteínas que se creó en 1994 y se ha realizado cada dos años desde ese momento, actualmente se encuentra en su versión CASP11. Se originó con el propósito de motivar a los grupos de investigación del mundo dedicados en este tema a mostrar sus avances a través de la evaluación de la precisión del método propuesto como solución del problema [28].

La competencia se clasifica en diversas categorías de acuerdo con el problema que se quiere abordar: predicción de dominios, predicción de la estructura secundaria, predicción de los mapas de contactos, reconocimiento del plegamiento, predicción de la estructura terciaria y predicción de la función de una proteína.

En la predicción de la estructura terciaria de las proteínas, esta competencia consiste en proveer a los concursantes de unas secuencias de aminoácidos de unas proteínas a las cuales se les ha encontrado la estructura tridimensional a través de cristalografía de rayos X o resonancia magnética nuclear, pero cuyas estructuras no han sido publicadas. Posteriormente, se comparan las estructuras generadas por los métodos contra las estructuras reportadas por medio de un alineamiento estructural. Esta evaluación se lleva a cabo teniendo en cuenta la naturaleza del método, es decir, si es de homología, *threading* o *ab initio*. Finalizada la evaluación de la calidad de cada uno de los métodos propuestos, se genera una tabla con los resultados obtenidos, dando de este modo a un grupo como el ganador del evento [3].

El problema de la predicción de la estructura terciaria de las proteínas cuenta con un conjunto de datos que puede ser utilizado para comparar los diferentes métodos que solucionan el problema. Este conjunto de datos puede ser descargado de la siguiente página web: http://predictioncenter.org/.

3. Metodología propuesta

En este capítulo se presentan el modelo y la metodología propuestos para la búsqueda de conformaciones de la proteína objetivo. La combinación de métodos de inteligencia computacional que se utilizan para el desarrollo del modelo propuesto es un algoritmo híbrido multiobjetivo de búsqueda dispersa.

El modelo propuesto consiste de dos etapas, en la primera se genera un modelo inicial para la secuencia objetivo mediante el uso de bibliotecas de fragmentos, mientras que en la segunda se realiza un refinamiento de esta conformación.

A continuación se presenta en detalle cada una de las etapas mencionadas anteriormente y se hace un análisis de la complejidad computacional del algoritmo AbYSS usado.

3.1 Generación del modelo inicial

Con el objetivo de reducir el espacio de conformaciones para encontrar la estructura nativa de una proteína, se utiliza la técnica de ensamblaje de fragmentos a través de PyRosetta, la cual es una herramienta de *software* de modelamiento molecular que permite generar estructuras de proteínas a partir de su secuencia de aminoácidos mediante el uso de la técnica usada por Rosetta, que está fundamentada en que una región local del *backbone* de una conformación con baja energía es el resultado de una secuencia local.

De acuerdo con lo anterior, un fragmento es un segmento de secuencia que reemplaza los valores de los ángulos de torsión para un conjunto consecutivo de residuos. Los fragmentos son derivados de una base de datos de fragmentos de baja energía para esa secuencia de residuos, conocida como biblioteca de fragmentos.

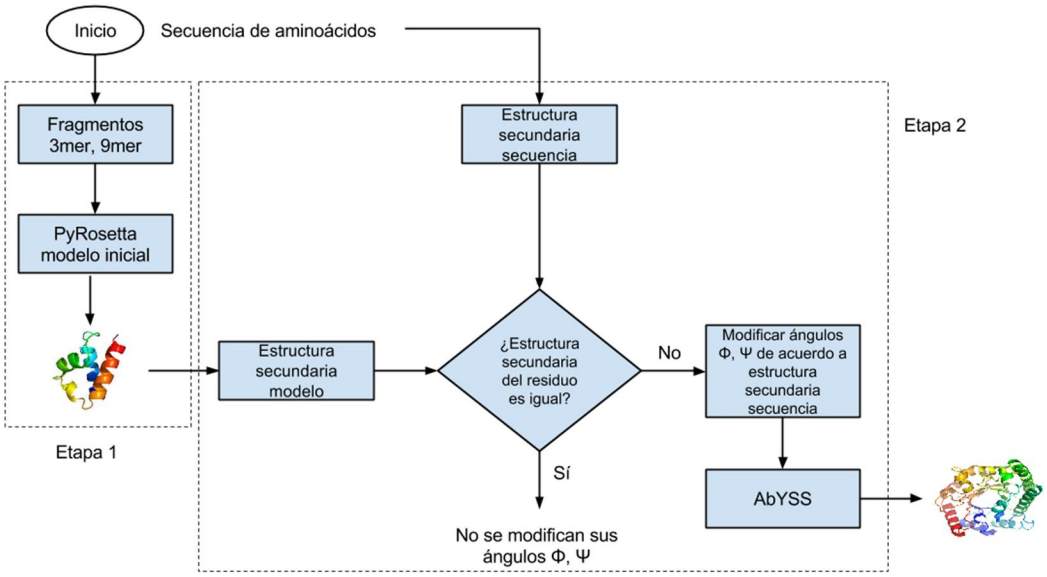
Se utilizan dos tipos de fragmentos de acuerdo con su tamaño; estos son 3-meros (tamaño 3) y 9-meros (tamaño 9).

El método para la generación de la biblioteca de fragmentos involucra la búsqueda de un conjunto no redundante en el PDB de las 100 más altas frecuencias de fragmentos que contienen un perfil de secuencia similar con la secuencia de búsqueda (secuencia objetivo de estudio) de fragmentos.

Para este caso se utilizó el servidor de predicción de estructuras de proteínas Robetta, que se encarga de generar las bibliotecas de fragmentos de tamaño 3 y 9 para una secuencia de aminoácidos.

Posteriormente, se genera un modelo a partir de la secuencia de aminoácidos, ejecutando un programa (*script*) que se ejecuta en PyRosetta, en el cual se realizan perturbaciones cortas (3-meros) y grandes (9-meros) en el *backbone* de acuerdo a las bibliotecas de fragmentos. A partir de esos movimientos en el *backbone* se optimiza el modelo a través de un método de Monte Carlo mediante la función de energía Talaris2013 que se explica en secciones posteriores. Finalmente, se obtiene un modelo que será insumo de la etapa dos. En la Figura 3-1 se observa el esquema general de la metodología propuesta.

Figura 3-1: Metodología propuesta



Esquema general de la metodología propuesta

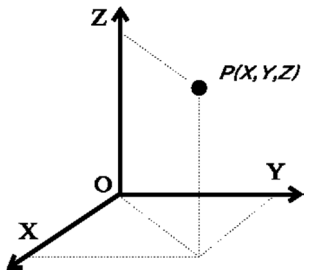
3.2 Refinamiento y optimización del modelo generado

Para finalizar el proceso de predicción de la estructura de una proteína objetivo se realiza el refinamiento de la estructura generada en la etapa anterior a través de un algoritmo híbrido de búsqueda dispersa que se encarga de explorar e intensificar la búsqueda en el espacio de conformaciones. Esta etapa se caracteriza por tener un sistema de representación de las proteínas, un algoritmo de búsqueda y una función de puntuación.

3.2.1 Sistema de representación de las proteínas

Para modelar computacionalmente una proteína se debe tener un sistema de representación. Como resultado de la etapa anterior se genera un modelo preliminar de la proteína a través de un archivo PDB, que contiene las coordenadas cartesianas (x, y, z) de cada uno de los átomos de los aminoácidos presentes en la proteína.

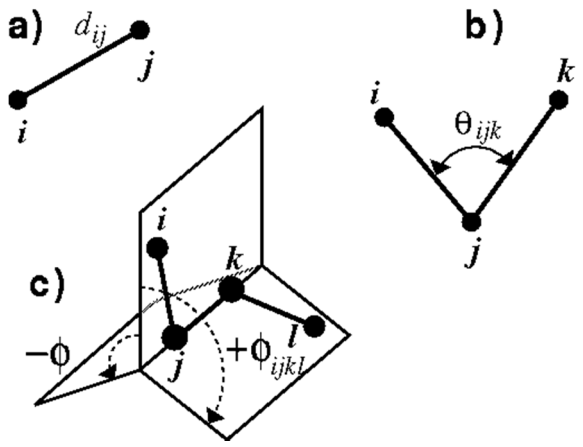
Figura 3-2: Representación cartesiana a nivel atómico de las proteínas.



Los átomos son representados como puntos en el espacio (imagen tomada de [41])

Sin embargo, para realizar modificaciones estructurales en la proteína se utiliza otra representación, debido a la dificultad que se presenta al realizar cambios en los ángulos de torsión de la estructura. Por tal motivo se utiliza una representación trigonométrica de las proteínas, también llamadas coordenadas internas o Matriz Z, que definen la posición de un átomo como la distancia de enlace con otro átomo, el ángulo que se forma entre dos enlaces y el ángulo de torsión con un átomo presente en otro plano. En la Figura 3-3 se muestra esta representación.

Figura 3-3: Representación trigonométrica a nivel atómico de las proteínas



a) Distancia del enlace conformado entre los átomos *i, j*. b) Ángulo formado entre los enlaces de los átomos *i, j, k*. c) Ángulo de torsión formado entre el plano conformado por los átomos *i, j, k* y el plano en el que se encuentra el átomo *l*. (imagen tomada de [41])

La representación trigonométrica es computacionalmente tratada a través de una matriz de coordenadas internas con la siguiente información:

Secuencia de átomos	Tipo de átomo	Átomo enlace	Longitud del enlace	Átomo ángulo	Ángulo del enlace	Átomo de torsión	Ángulo de torsión

Sin embargo, los cálculos de energía se realizan a través de las coordenadas cartesianas, por lo que se llevan a cabo las conversiones respectivas entre ambos tipos de representaciones.

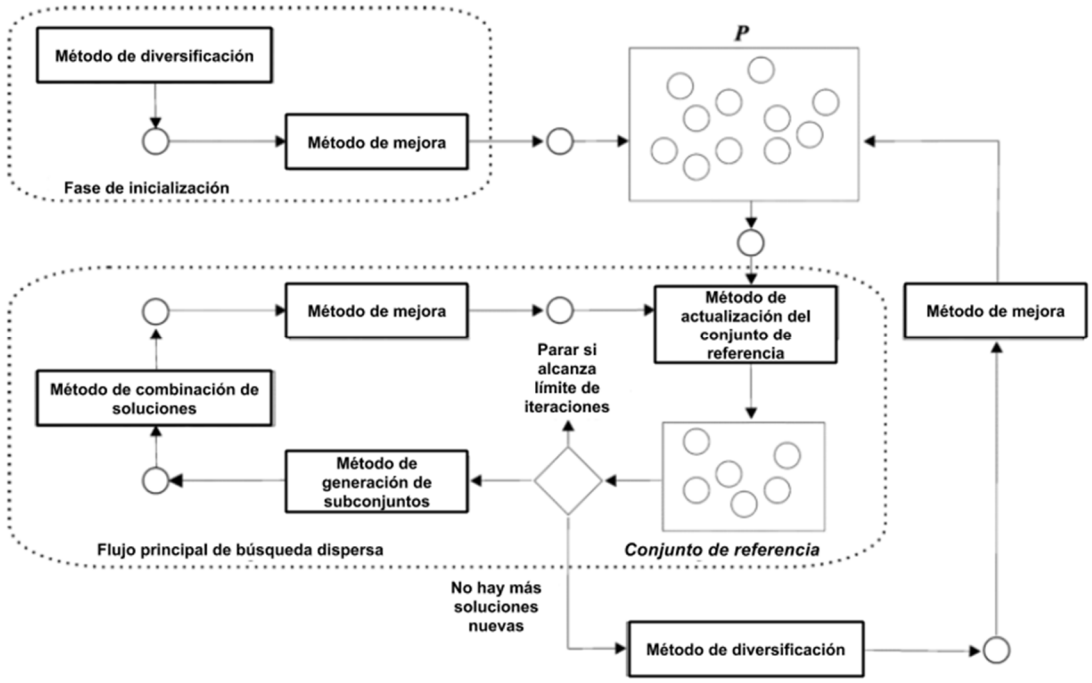
3.2.2 Algoritmo de búsqueda

El método de búsqueda utilizado es una implementación del algoritmo de búsqueda dispersa multiobjetivo híbrido llamado AbYSS (*A Hybrid Scatter Search*), el cual usa operadores de cruce y mutación durante la evolución de las soluciones; es decir, es un híbrido entre búsqueda dispersa y algoritmos genéticos. En la Figura 3-4 se observa el funcionamiento del algoritmo, el cual está compuesto por una población P y un conjunto de referencia por un conjunto de métodos que se encargan de realizar la exploración y la

intensificación en el espacio de búsqueda, y un archivo externo que contendrá las mejores soluciones halladas en cada iteración [42].

Los métodos que describen el algoritmo son: el método para la generación de soluciones diversas, el método de mejora, el método de actualización del conjunto de referencia, el método de generación de subconjuntos, y el método de combinación de soluciones, los cuales se explican a continuación [42].

Figura 3-4: Esquema del algoritmo AbYSS



(imagen adaptada de [42])

Método de generación de soluciones diversas: es un método basado en la división del intervalo de las variables de decisión en un número de subintervalos de igual tamaño. El valor de cada variable de decisión es generada en dos pasos, primero se elige un subrango de la variable aleatoriamente, la probabilidad de seleccionar ese subrango es inversamente proporcional a la frecuencia (número de veces que el subrango ha sido seleccionado). Luego, se genera un valor aleatorio con el intervalo de selección [42].

Método de mejora: es un método de búsqueda local al cual se le adiciona el operador mutación. Se hace una copia de la solución a la cual se le realiza una mutación, luego se compara la solución copia contra la original con el propósito de determinar cuál es mejor.

Para ello se realiza una prueba de violación de restricciones y una prueba de dominancia de Pareto, con el objetivo de elegir una solución factible y mejorada [42].

Método de actualización del conjunto de referencia: el conjunto de referencia es una colección que contiene soluciones diversas y de alta calidad. Está conformado por dos subconjuntos llamados RefSet1 y RefSet2, de tamaños p y q, respectivamente. El subconjunto RefSet1 contiene las soluciones con mejor calidad de P, es decir, individuos que tienen un mejor *fitness*; mientras que el subconjunto RefSet2 tiene las soluciones con mayor diversidad de P, es decir, que la distancia euclidiana con otras soluciones es alta [42].

Una solución es incluida en el conjunto de referencia si se satisfacen las siguientes condiciones: el nuevo individuo debe tener un mejor valor en la función objetivo con respecto al peor valor del RefSet1, también debe tener mejor valor de distancia con el conjunto de referencia que el individuo con peor valor de distancia del RefSet2 [42].

Cuando una solución no es dominada por el RefSet1, esta se inserta en el conjunto de referencia sólo si este no está completo. Es decir, esta solución debe dominar al menos a un individuo del conjunto de RefSet1. Si esta condición no se cumple, el individuo es insertado en el archivo externo [42].

Método de generación de subconjuntos: es el encargado de generar subconjuntos de individuos con el objetivo de crear nuevas soluciones mediante el método de combinación. Este método genera pares de individuos que pertenecen a RefSet1 y, por otro lado, pares de individuos que pertenecen a RefSet2 [42].

Método de combinación de soluciones: se realiza una combinación pareada de individuos a través del operador de cruce [42].

Archivo externo: contiene el registro histórico de las soluciones no dominadas encontradas durante la búsqueda, intentando, a la vez mantener aquellas que producen una mejor distribución en el frente de Pareto [42].

Descripción del algoritmo AbYSS

Inicialmente, el método de generación de soluciones diversas es invocado para generar soluciones iniciales, cada una de ellas ejecuta el método de mejora; como resultado se obtiene el conjunto inicial P. Posteriormente se realiza el siguiente proceso durante un número determinado de iteraciones: primero se crea el conjunto de referencia, se llama el método de generación de subconjuntos, y se ejecuta el bucle principal hasta cuando no se generen más subconjuntos. Después, hay una fase de reinicio que consta de tres pasos. Primero, los individuos en RefSet1 se insertan en P; segundo, los n mejores individuos del archivo externo, se mueven también a P; y, tercero, se usan el método de generación de soluciones diversas y el método de mejora para producir nuevas soluciones hasta completar P [42]. Los parámetros de ejecución del algoritmo son los siguientes:

- Tamaño de la población
- Tamaño del conjunto de referencia 1 (RefSet1)
- Tamaño del conjunto de referencia 2 (RefSet2)
- Tamaño del archivo
- Número total de evaluaciones
- Probabilidad de cruce
- Distribución de probabilidad del cruce
- Probabilidad de mutación
- Distribución de probabilidad mutación
- Archivo PDB generado en la etapa anterior

Para el problema de la predicción de la estructura de las proteínas, primero se lee el archivo PDB generado en la etapa anterior con el propósito de generar las coordenadas internas de la proteína; para ello se utilizan los programas PDBXYZ y XYZINT del software TINKER. Para la ejecución de estos programas es necesario partir de un archivo PDB para generar un archivo .XYZ, que posteriormente será utilizado para crear un archivo .INT con las coordenadas internas de la proteína.

Luego se asignan las variables de decisión del problema, que son los valores que pueden modificarse para obtener distintos valores de la función de puntuación. En este caso los ángulos de torsión del *backbone* de la proteína son las variables de decisión, ya

que los cambios estructurales se van a realizar en el *backbone* y no se tendrán en cuenta las cadenas laterales.

El número total de variables de decisión se determina de la manera que se describe a continuación; así mismo, en la Figura 3-5 se muestra un ejemplo de este proceso.

1. Predecir la estructura secundaria de la secuencia objetivo a través del software de predicciones de estructuras secundarias PSIPRED.

2. Conocer las estructuras secundarias formadas en el modelo de la etapa anterior mediante el software PyRosetta.

3. Comparar las estructuras secundarias de cada residuo obtenidas en PSIPRED contra las conseguidas en el modelo de PyRosetta. En caso de que las estructuras secundarias sean iguales para ese residuo, los ángulos φ y Ψ de esos residuos no serán modificados (sólo aplica cuando la estructura secundaria es una hélice α o una lámina β); en otro caso, se varían de acuerdo con los rangos de valores establecidos para las estructuras secundarias como se muestra en la Tabla 3-1.

4. El número de variables de decisión es el siguiente:

$$NumVar = 2 \times (NumResiduos - NumResSS - 2) \quad (3\text{-}1)$$

donde numVar es el número de variables de decisión, NumResiduos es el total de residuos de la secuencia objetivo, NumResSS es la cantidad de residuos con estructura secundaria idéntica obtenida de la comparación hecha en un paso anterior.

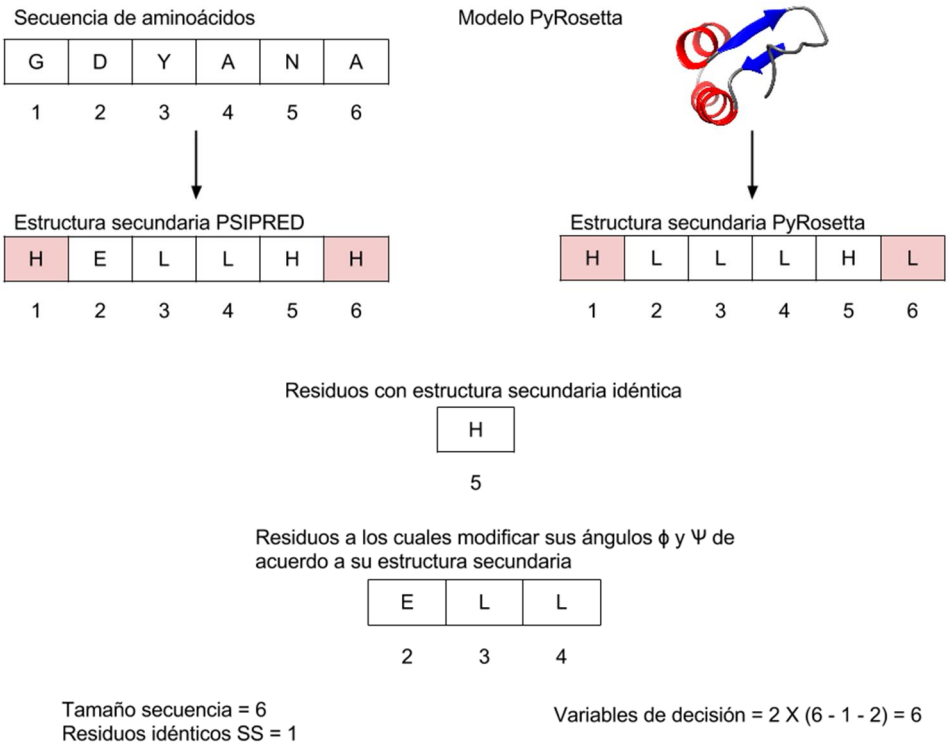
Al tamaño de la secuencia se le restan los residuos que contienen estructura secundaria idéntica y también se le restan dos residuos que corresponden al primero y al último de la secuencia, ya que estos sólo poseen un ángulo de torsión. El resultado de esta operación es multiplicado por dos, ya que por cada residuo hay un ángulo φ y Ψ.

Tabla 3-1: Rango de los ángulos φ y Ψ en las estructuras secundarias

Estructura Secundaria	**φ**	**Ψ**
Hélice Alfa	$[-67, -39]$	$[-57, -16]$
Lámina Beta	$[-130, -110]$	$[110, 130]$
Loop	$[-180, 180]$	$[-180, 180]$

(tabla tomada de [43])

Figura 3-5: Variables de decisión



3.2.3 Función de evaluación

La función de energía que se escogió para evaluar la calidad de la estructura fue Talaris2013, la cual es la función de energía utilizada por el predictor de estructuras de proteínas Rosetta. Esta función se caracteriza por ser una combinación lineal de términos que modelan la interacción de fuerzas entre átomos, efectos solventes y energías de torsión [44]. En la Ecuación 3-2 se observa que la función de energía asigna un peso a cada uno de los términos de la misma.

$$E(C|w,\theta) = \sum_j^{|T|} w_j T_j(C|\theta_j) \qquad (3\text{-}2)$$

La energía total del sistema para una conformación molecular C está compuesta por términos de energía T_j con parámetros θ_j y un peso w_j.

Los pesos de los términos son establecidos en el archivo de parámetros de la función de energía y varía de acuerdo con el tipo de residuo.

Los términos de la función de puntuación se organizaron en dos grupos para constituir los objetivos del algoritmo AbYSS, ya que este método es multiobjetivo. En el primero se agruparon los términos de no enlace, los cuales son los relacionados con puentes de hidrógeno, los puentes de disulfuro, las atracciones y repulsiones de Lennard-Jones, la energía de solvatación, el potencial electrostático y los enlaces de disulfuro. Por otro lado, el segundo está compuesto por términos relacionados con los ángulos de torsión y la geometría de la conformación como son las preferencias de Ramachandran, preferencias del ángulo omega, energía interna de los rotámeros, probabilidad de los ángulos Φ y Ψ, y energía de referencia de cada aminoácido.

Las ecuaciones de los términos mencionados anteriormente se encuentran en el Anexo A.

3.3 Complejidad computacional del algoritmo AbYSS

Para el análisis de la complejidad computacional del algoritmo AbYSS se tiene en cuenta el peor de los casos y se analiza a través de la cota superior asintótica, también llamada O Grande, que se define como [47]:

$$O(g(n)) = \left\{ \begin{matrix} f(n): existen\ constantes\ positivas\ c\ y\ n_0\ \ tales\ que \\ 0 \leq f(n) \leq cg(n)\ para\ toda\ n \geq n_0 \end{matrix} \right\}$$

donde $f(n)$ es una función que pertenece a $O(g(n))$ y se encuentra entre 0 y la función $g(n)$.

Dado que el algoritmo AbYSS está compuesto por dos fases (inicialización y bucle principal) se hizo el análisis de cada una, con el objetivo de escoger la de mayor valor como la complejidad del algoritmo.

Para determinar la complejidad se establecieron las siguientes variables, las cuales se usan en las fases de dicho algoritmo:

P: número de individuos de la población

I: número de iteraciones

S_1*:* tamaño del subconjunto de referencia 1

S_2*:* tamaño del subconjunto de referencia 2

a: número de átomos de una proteína

t: número de ángulos de torsión = (2 * cantidad total de aminoácidos) - 2

f(talaris): función de puntuación Talaris2013 (sujeta al orden de crecimiento de la codificación de la función)

A continuación se hace un análisis de la complejidad de los términos que componen la función de puntuación Talaris2013, cuyas fórmulas se encuentran en el Anexo A.

- Lennard Jones : $O(n^2)$
- Rama: $O(n)$
- Puentes de hidrógeno: $O(n^2)$
- Solvatación: $O(n^2)$
- Pair: $O(n^2)$
- Dun: $O(n)$
- Ref: $O(n)$

De acuerdo al análisis realizado, la función Talaris2013 es de orden cuadrático.

3.3.1 Fase de inicialización

Para determinar la complejidad de esta etapa se tuvo en cuenta la codificación de los métodos de diversificación y mejora, ya que estos son utilizados para generar la población inicial del algoritmo. Al revisar este proceso se estableció que depende directamente del tamaño de la población, ya que para cada individuo de la población se ejecutan los métodos mencionados anteriormente. Así mismo, se identificó que hay un

procesamiento de datos que se realiza en cada individuo, porque en este se almacenan las coordenadas cartesianas de los átomos que componen la proteína, los cuales son accedidos y modificados mediante la lectura y escritura de los archivos PDB, XYZ e INT, gracias a la modificación de los ángulos de torsión. También se evidenció que se realiza la ejecución de la función de puntuación en cada uno de los métodos de esta fase.

De acuerdo con lo anterior, la complejidad de esta fase está dada por:

$$O(P * (t + a + f(talaris))$$

Complejidad que es de orden cuadrático, ya que la complejidad de la función Talaris2013 es la de mayor crecimiento. Así mismo, se supone que la codificación de la misma está en dicho orden.

3.3.2 Fase del bucle principal

Para determinar la complejidad de esta etapa se tuvo en cuenta la codificación de los métodos de actualización del conjunto de referencia, de generación de subconjuntos, de combinación de soluciones y de mejora, los cuales son usados en el bucle principal del algoritmo AbYSS.

Al analizar esta etapa se observa que los métodos mencionados dependen del número de iteraciones y del tamaño de la población, ya que se ejecutan hasta determinado número de iteraciones y para cada uno de los individuos de la población.

Con respecto a la población del conjunto de referencia, esta se distribuye en los subconjuntos de referencia 1 y 2 mediante el ordenamiento de los individuos con mejor puntuación y los más diversos, respectivamente. Igualmente, se identificó que el procesamiento de datos y la ejecución de la función de puntuación mencionada en la fase anterior se llevan a cabo en el bucle principal en el método de generación de subconjuntos.

De acuerdo con lo anterior, la complejidad de esta fase está dada por:

$$O(I * (S^2 + P \log_2 P + f(talaris))$$

- Al igual que el caso de la fase anterior, los términos son de menor orden con respecto a la función Talaris2013; por tanto, se establece que la complejidad del

algoritmo utilizado es de orden cuadrático en función del número de átomos y de residuos de la proteína, en el cálculo de las funciones de Lennard Jones, puentes de hidrógeno, solvatación e interacciones electrostáticas.

4. Experimentación y resultados

En este capítulo se presentan la experimentación y los resultados obtenidos utilizando el modelo propuesto.

Las proteínas seleccionadas para la experimentación fueron tomadas del *Protein Data Bank,* con el propósito de tomarlas como punto de referencia para comparar los resultados obtenidos con las proteínas reportadas experimentalmente. En la Tabla 4-1 se describe el conjunto de datos utilizado.

Tabla 4-1: Conjunto de datos utilizado en la experimentación

Identificador PDB	Tamaño secuencia de aminoácidos	Método Experimental
1CRN	46	Cristalografía de rayos X
2KBQ	80	Resonancia Magnética Nuclear
1ROP	93	Cristalografía de rayos X
2MQL	105	Resonancia Magnética Nuclear

De las anteriores proteínas se obtuvo la secuencia de aminoácidos en formato FASTA publicada en el sitio web del *Protein Data Bank*.

El lenguaje de programación usado en la implementación del modelo propuesto fue Java.

Los parámetros utilizados en la ejecución del algoritmo se observan en la Tabla 4-2.

Tabla 4-2: Parámetros usados en la ejecución del procedimiento propuesto

Parámetro	Valor
Secuencia de aminoácidos	Secuencia ingresada
Biblioteca de fragmentos de tamaño 3	Archivo generado en Robetta
Biblioteca de fragmentos de tamaño 9	Archivo generado en Robetta
Estructura secundaria PSIPRED	Secuencia obtenida en PSIPRED
Tamaño del conjunto de referencia	20
Tamaño del conjunto de referencia 1	10
Tamaño del conjunto de referencia 2	5
Número total de evaluaciones	100000
Probabilidad de cruce	0.9
Probabilidad de mutación	1 / número de variables de decisión
Tamaño del archivo	5

Los valores de los parámetros de ejecución del algoritmo se tomaron con base en el artículo sobre el algoritmo AbYSS [42], en el cual se recomienda tener un conjunto de referencia de 20 individuos y el conjunto de referencia 1 de mayor tamaño con respecto al conjunto de referencia 2, con el propósito de intensificar la explotación y reducir la exploración; así mismo, sugieren una probabilidad de cruce de 0.9 y una probabilidad de mutación de 1/n, donde n es la cantidad de variables de decisión; el número total de evaluaciones se estableció en 100000, ya que en la etapa 1 se obtiene un modelo de la estructura de la proteína por medio de ensamblaje de fragmentos. Así mismo, el valor de estos parámetros fue validado mediante la ejecución del algoritmo con distintos valores en los que se comprobó el comportamiento de los mismos. El tamaño del archivo de

resultados de AbYSS contiene las mejores cinco soluciones halladas durante la ejecución del algoritmo.

Los resultados obtenidos fueron comparados con la estructura tridimensional experimental reportada en el PDB, y las estructuras predichas por los servidores de predicción de estructuras de proteínas I-TASSER y QUARK. Se escogieron estos servidores, teniendo en cuenta que el primero permite excluir de las plantillas las proteínas objetivo, es decir, aquellas que son objeto de estudio. Mientras que el segundo es un servidor *ab initio* que no usa plantillas.

Cada modelo obtenido se visualizó usando UCSF Chimera [48] y los resultados obtenidos se analizaron teniendo en cuenta las siguientes medidas de desempeño:

- RMSD C_α: valor obtenido al comparar los C_α de la estructura predicha con la estructura reportada experimentalmente en el PDB. El valor del RMSD se calculó con el servidor *SuperPose* [49]. Los valores del RMSD son calculados mediante la siguiente formula [12]:

$$RMSD = \sqrt{\frac{\sum_{i=0}^{n}(A_{ix} - B_{ix})^2 + \left(A_{iy} - B_{iy}\right)^2 + (A_{iz} - B_{iz})^2}{n}}$$

  donde A y B son las estructuras a comparar, *n* es el número de átomos equivalentes, A_{ix} es la posición en el eje x del átomo *i* de la estructura A, B_{ix} es la posición en el eje x del átomo *i* de la estructura B, A_{iy} es la posición en el eje y del átomo *i* de la estructura A, B_{iy} es la posición y del átomo *i* de la estructura B, A_{iz} es la posición z del átomo i de la estructura A, y B_{iz} es la posición z del átomo *i* de la estructura B.

- RMSD local: valor obtenido al comparar estructuras secundarias estables (hélices alfa y láminas beta) de la proteína predicha y la proteína determinada experimentalmente. El valor RMSD local se calculó con el programa *SuperPose* [49] teniendo en cuenta los residuos en los que se presentaron estas estructuras.

- Validación de la estructura generada mediante el software *Protein Structure Validation Suite* (PSVS) [50], que usa otras herramientas como PROCHECK, Molprobity, Verify3D, entre otras; para realizar los siguientes análisis:

- Análisis de estructuras secundarias.

- Gráfico de Ramachandran: permite comparar la cantidad de residuos ubicados en regiones permitidas. Los gráficos fueron generados mediante el servidor *Rampage* [51]. Este muestra cuatro gráficos de Ramachandran de acuerdo con el tipo de residuo; el primero es un esquema general en el cual se representan los ángulos de torsión de los residuos que no son glicina, ni prolina, ni residuos que son inmediatamente anteriores a la prolina; el segundo muestra los residuos tipo glicina; el tercero expone los pre-prolinas, es decir, residuos que preceden una prolina en la secuencia; y el cuarto presenta los residuos de tipo prolina.

  Puntuación de calidad global: se evalúa la compatibilidad del modelo atómico con respecto a su secuencia de aminoácidos mediante el software Verify3D; mientras que con el software Prosa II se modela una representación reducida de la energía de los pares de interacciones de la separación espacial de los átomos C_β de residuos locales; del mismo modo, con Procheck (phi-psi) y Procheck (all) se analiza la estereoquímica de la estructura para los ángulos de torsión, mediante el valor del G-Factor que provee una medida inusual de una propiedad, en el cual si un valor es menor a -0.5 se identifica como inusual, y si el valor es menor a -1.0 es altamente inusual; por tanto un valor positivo indica una mejor puntuación; así mismo, se hizo un análisis de los posibles golpes estéricos mediante el software MolProbity y su métrica Clashscore, la cual toma en cuenta superposiciones estéricas de 0.4Å o superiores entre átomos no enlazados. También se tienen en cuenta los contactos cercanos y desviaciones de la geometría ideal, mediante la distancia cuadrática media de distancias y ángulos entre enlaces; se define un contacto cercano como aquella distancia menor a 2.2Å para átomos pesados y 1.6Å para átomos de hidrógeno. Así mismo, la desviación cuadrática media de la distancia de enlaces covalentes se establece en 0.010Å y para ángulos en 0.4º; esto se estableció de acuerdo con el promedio de los valores RMSD de todos los enlaces y los ángulos relacionados con el valor estándar para aminoácidos (diccionario). Todo enlace covalente y ángulo que se extienda seis veces el valor del RMSD estándar del diccionario es considerado un valor atípico.

Con las anteriores medidas se obtienen dos tipos de puntación: el primero es el valor medio de los valores obtenidos en cada residuo, mientras que el segundo es el Z-score, mediante el cual se puede establecer que los datos obtenidos se sitúan por encima o por debajo de la media.

4.1 Resultados obtenidos con la proteína 1CRN

La proteína **1CRN** está compuesta por 46 residuos y se clasifica como una proteína de plantas. Su estructura fue predicha a través de cristalografía de rayos X. Es una proteína que se origina en el organismo *Crambe hispanica* subespecie abyssinica y se caracteriza por no tener asignada una función bioquímica; su proceso biológico es la respuesta de defensa, que consiste en reacciones a la presencia de cuerpos extraños o a la ocurrencia de una lesión, que dan como resultado daños en el organismo atacado o la prevención/restauración de la infección causada por el ataque. Su componente celular se ubica en la región extracelular, la cual corresponde al espacio externo de la estructura de la célula. Para células que no poseen una protección externa se refiere al espacio exterior de la membrana plasmática [52].

El dominio de la secuencia a la cual pertenece es Thionin y el dominio de su estructura es Crambin-like.

En la base de datos CATH, se clasifica de la siguiente manera:

- Clase: Alpha Beta
- Arquitectura: *sandwich* de 2 capas (*2-layer sandwich*)
- Topología: Crambin
- Homología: Crambin-like

Mientras que en la base de datos SCOP la clasificación es del siguiente modo:

- Clase: proteína pequeña
- Pliegue: Crambin-like
- Superfamilia: Crambin-like

En la Figura 4-1 se muestra el modelo experimental reportado en el PDB, y los modelos generados por los procedimientos mencionados previamente, mientras que en la Tabla 4-3 y en la Tabla 4-4 se muestran los valores RMSD de las estructuras generadas en

comparación con la reportada experimentalmente. La estructura terciaria de esta proteína se caracteriza por tener dos hélices alfa entre los residuos 7-17 y 23-30; mientras que en los residuos 2-3 y 33-34 se exhiben dos láminas beta.

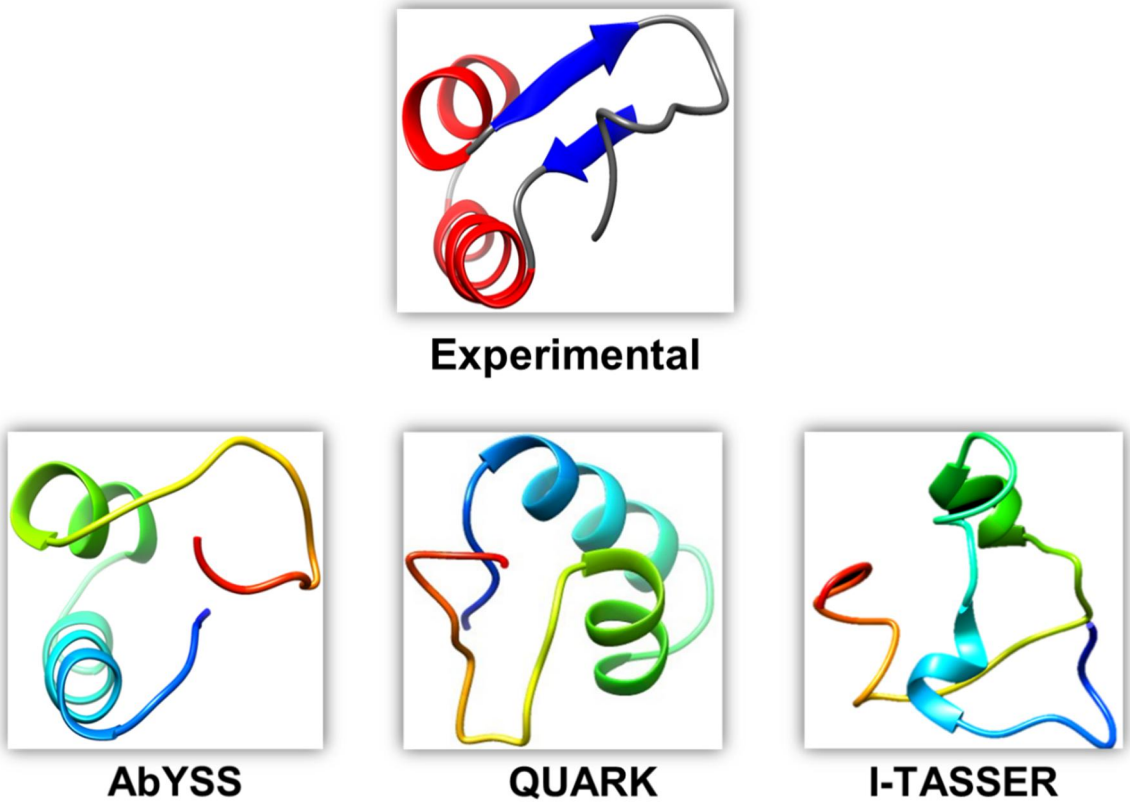
Tabla 4-3: Resultados obtenidos con la secuencia de la proteína 1CRN

Servidor	**RMSD Cα (Å)**
AbYSS	4.22
QUARK	3.33
I-TASSER	9.16

Tabla 4-4: RMSD local de la proteína 1CRN

Residuos (Å) / **Servidor**	**2-3**	**7-17**	**23-30**	**33-34**	**RMSD Local Cα (Å)**
ABYSS	0.49	0.21	0.3	0.16	1.16
QUARK	0.45	0.19	0.24	0.34	1.22
I-TASSER	0.68	2.5	1.82	0.55	5.55

Figura 4-1: Modelos generados para la proteína 1CRN



Datos de validación de la estructura generada:

Estructuras secundarias: se obtuvieron de la estructura generada dos hélices alfa entre los aminoácidos 7-17 y 23-30; no hay láminas beta, como se observa en la Figura 4-2.

Figura 4-2: Elementos de estructuras secundarias de la proteína predicha 1CRN

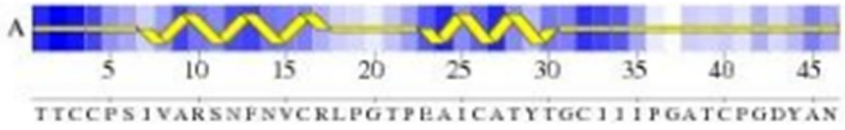
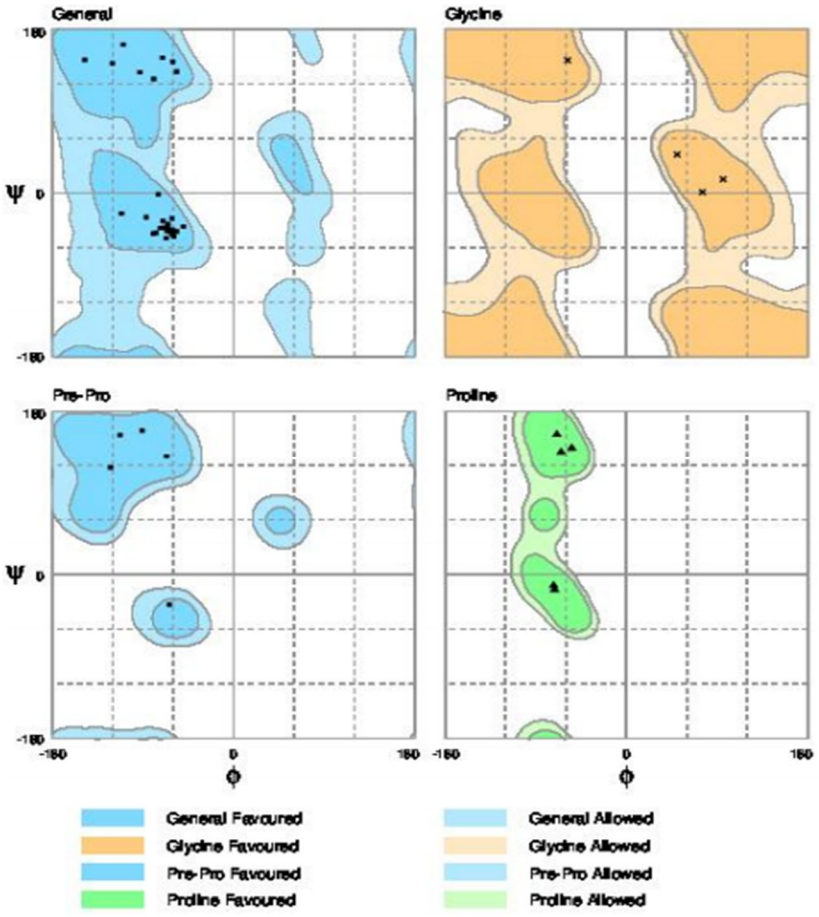


Gráfico de Ramachandran: en el que se muestra en la Figura 4-3 se observa que los ángulos de torsión de los residuos se encuentran en zonas destacadas, es decir, el 100% de los residuos están en zonas favorables. De acuerdo con las estructuras secundarias presentes en la estructura de esta proteína, se observa que los ángulos se ubican en el rango de ángulos permitido para las hélices alfa y las láminas beta. También se puede establecer que no se presentan valores atípicos en el gráfico.

Figura 4-3: Gráfico de Ramachandran para la estructura de la proteína 1CRN generada por el algoritmo AbYSS



Puntuación de calidad global de la estructura del modelo experimental reportado en el PDB:

Tabla 4-5: Resultados obtenidos con el modelo experimental reportado para la proteína 1CRN

Programa	Verify3D	Prosall (-ve)	Procheck (phi-psi)	Procheck (all)	MolProbity Clashscore
Mean score	0.26	0.31	0.18	0.16	0
Z-score	-3.21	-1.41	1.02	0.95	1.53

Número de contactos cercanos: 0

RMSD para ángulos de torsión: 2.4°

RMSD para longitud de enlaces: 0.023 Á

Puntuación de calidad global de la estructura modelo generado mediante el método AbYSS:

Tabla 4-6: Resultados obtenidos con el modelo generado para la proteína 1CRN

Programa	Verify3D	Prosall (-ve)	Procheck (phi-psi)	Procheck (all)	MolProbity Clashscore
Mean score	0.12	-0.47	0.26	-0.42	313.08
Z-score	-5.46	-4.63	1.34	-2.48	-52.20

Número de contactos cercanos: 50

RMSD para ángulos de torsión: 0.5°

RMSD para longitud de enlaces: 0.008 Á

4.2 Resultados obtenidos con la proteína 2KBQ

La proteína 2KBQ está compuesta por 80 residuos y se clasifica como una proteína estructural. Su estructura fue predicha a través de RMN y, por tanto, posee varios modelos tridimensionales. Para realizar las respectivas comparaciones y superposiciones estructurales se utilizó el modelo 1. Es una proteína que se origina en el humano (*Homo sapiens*) y se caracteriza por no tener asignado un proceso biológico bioquímico, ni una componente celular [52].

El dominio de secuencia al cual pertenece es Harmonin y no presenta dominios de estructura.

No posee anotaciones en las bases de datos CATH y SCOP.

En la Figura 4-4 se muestran el modelo experimental reportado en el PDB, y los modelos generados a través de métodos computacionales. La estructura terciaria de esta proteína se caracteriza por tener cinco hélices alfa entre los residuos 5-16, 20-36, 39-50, 59-65 y 68-77.

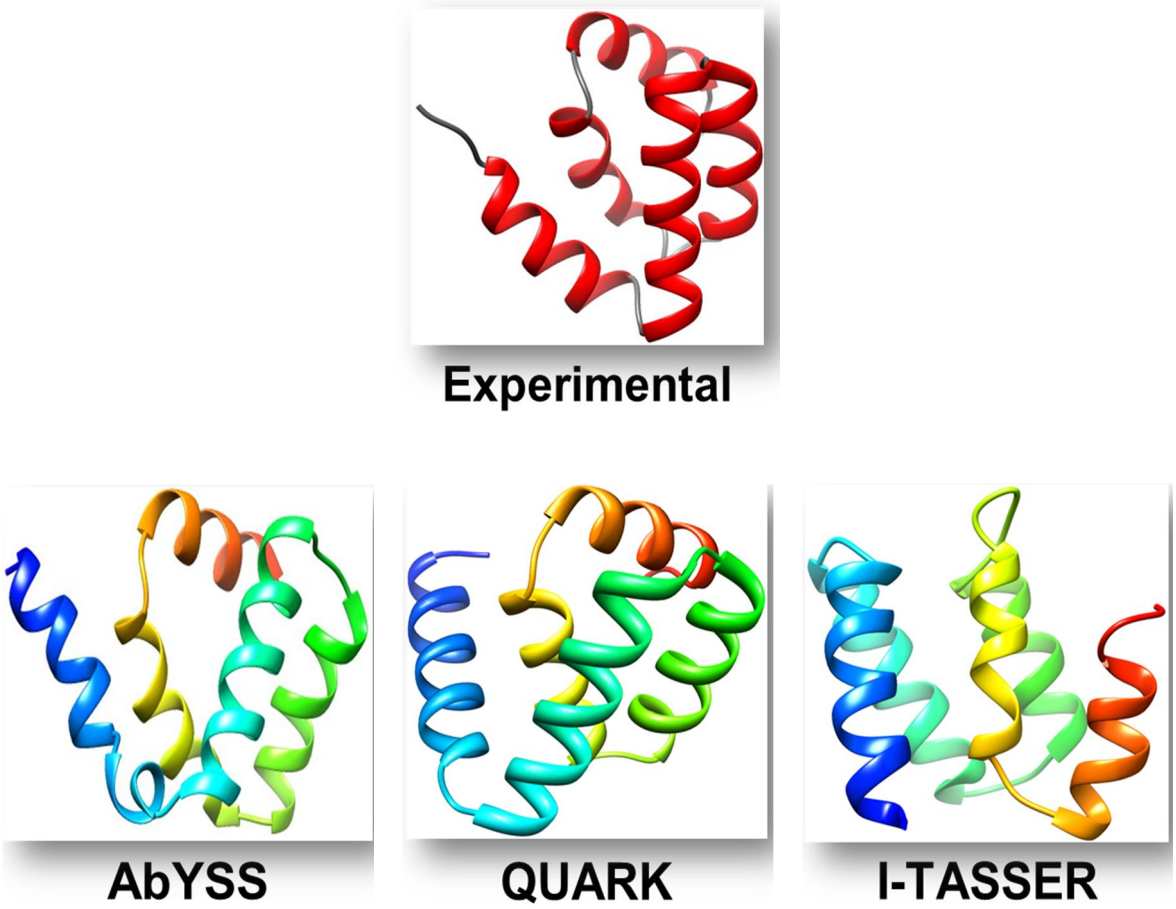
Tabla 4-7: Resultados obtenidos con la secuencia de la proteína 2KBQ

Servidor	RMSD Cα (Å)
AbYSS	4.42
QUARK	4.09
I-TASSER	1.95

Tabla 4-8: RMSD local de la proteína 2KBQ

Residuos (Å) / Servidor	5-16	20-36	39-50	59-65	68-77	RMSD Local Cα (Å)
ABYSS	1.69	0.57	0.85	0.39	0.66	4.16
QUARK	0.61	0.75	0.41	0.42	0.7	2.89
I-TASSER	0.43	0.41	0.42	0.27	0.3	1.83

Figura 4-4: Modelos generados para la proteína 2KBQ



Datos de validación de la estructura generada:

Estructuras secundarias: se obtuvieron de la estructura generada seis hélices alfa entre los aminoácidos 2-12, 14-17, 20-35, 39-51, 55-62 y 68-77 como se observa en la Figura 4-5.

Figura 4-5: Elementos de estructuras secundarias de la proteína predicha 2KBQ

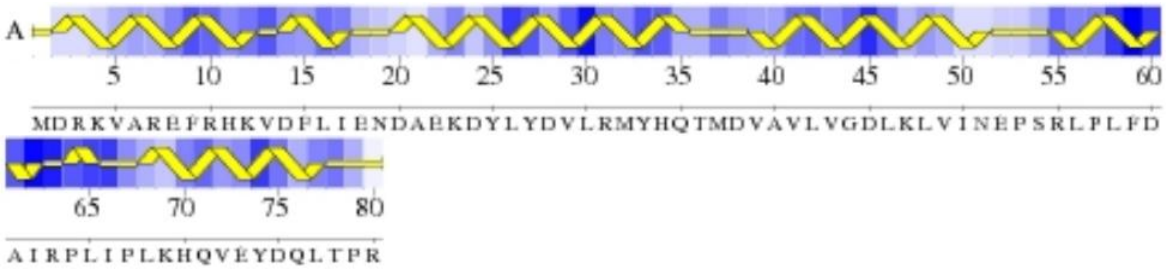
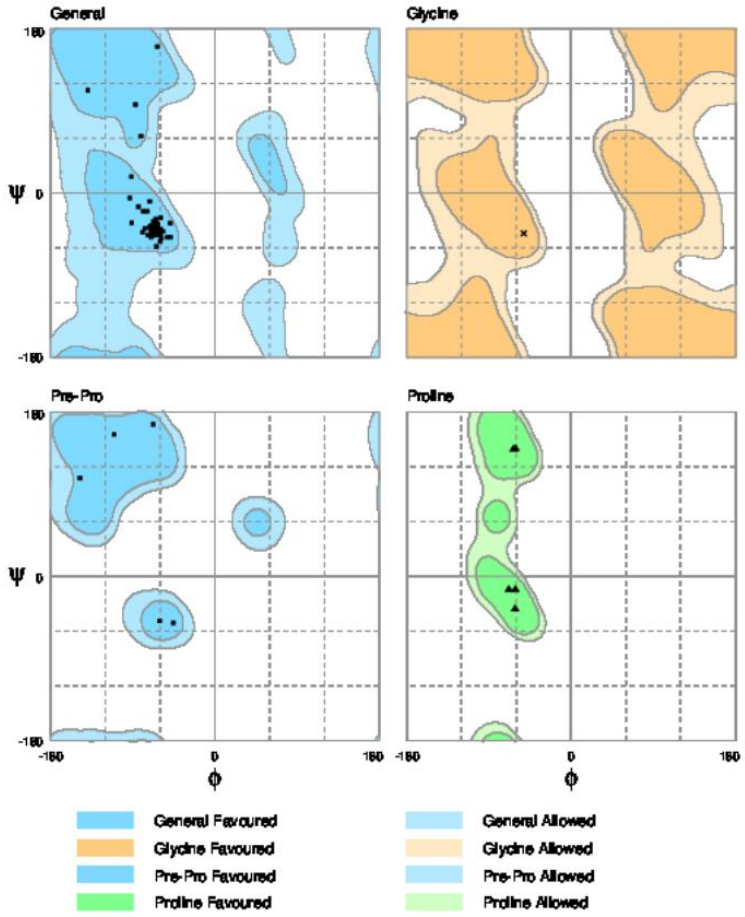


Gráfico de Ramachandran: de acuerdo con la Figura 4-6, se puede establecer que el 100% de los ángulos de torsión se encuentran en zonas favorecidas. De acuerdo con las estructuras secundarias presentes en la estructura de esta proteína, se observa que los ángulos se ubican en el rango de ángulos permitido para las hélices alfa, que es la estructura secundaria predominante en la proteína. También se observa que no se presentan valores atípicos en el gráfico.

Figura 4-6: Gráfico de Ramachandran para la estructura de la proteína 2KBQ generada por el algoritmo AbYSS



Puntuación de calidad global de la estructura del modelo experimental reportado en el PDB:

Tabla 4-9: Resultados obtenidos con el modelo experimental reportado para la proteína 2KBQ

Programa	Verify3D	Prosall (-ve)	Procheck (phi-psi)	Procheck (all)	MolProbity Clashscore
Mean score	0.27	0.49	0.03	-0.64	38.4
Z-score	-3.05	-0.66	0.43	-3.78	-5.06

Número de contactos cercanos: 0

RMSD para ángulos de torsión: 0.3°

RMSD para longitud de enlaces: 0.002 Á

Puntuación de calidad global de la estructura modelo generado mediante el método AbYSS:

Tabla 4-10: Resultados obtenidos con el modelo generado para la proteína 2KBQ

Programa	Verify3D	Prosall (-ve)	Procheck (phi-psi)	Procheck (all)	MolProbity Clashscore
Mean score	0.02	0.15	0.51	-0.98	581.13
Z-score	-7.06	-2.07	2.32	-5.8	-98.2

Número de contactos cercanos: 437

RMSD para ángulos de torsión: 0.4°

RMSD para longitud de enlaces: 0.010 Á

4.3 Resultados obtenidos con la proteína 1ROP

La proteína 1ROP está compuesta por 63 residuos y se clasifica como una proteína de regulación de la transcripción. Su estructura fue predicha a través de cristalografía de rayos X. Es una proteína que se origina en el organismo *Escherichia coli* y se caracteriza

por no tener asignada una función bioquímica ni una componente celular; su proceso biológico es la regulación de la transcripción [52].

El dominio de la secuencia a la cual pertenece es a las proteínas reguladoras Rop y el dominio de su estructura es *Helix Hairpins*.

En la base de datos CATH se clasifica de la siguiente manera:

- Clase: *Mainly Alpha*
- Arquitectura: *Orthogonal bundle*
- Topología: *Helix Hairpins*
- Homología: *Helix Hairpins*

Mientras que en la base de datos SCOP no tiene un dominio asignado.

En la Figura 4-7 se muestran el modelo experimental reportado en el PDB y los modelos generados a través de métodos computacionales. La estructura terciaria de esta proteína se caracteriza por tener dos hélices alfa entre los residuos 3-28 y 32-55.

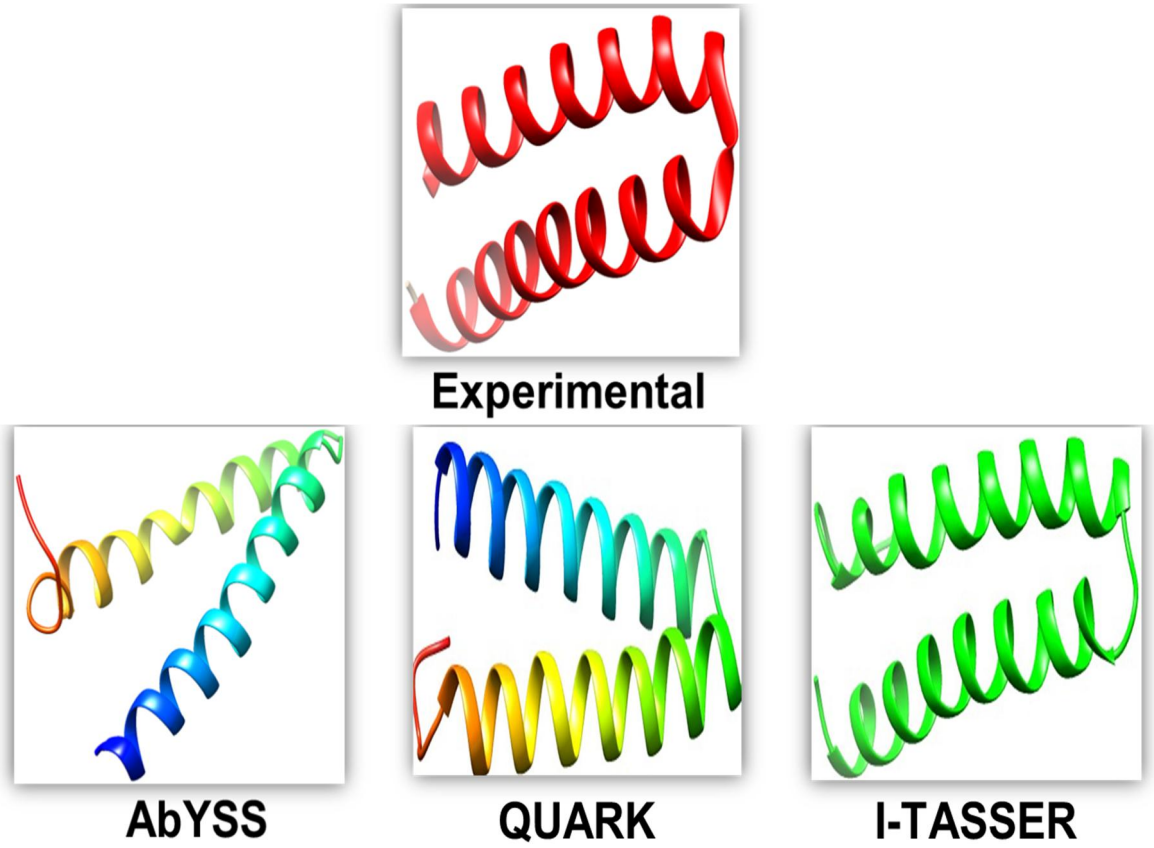
Tabla 4-11: Resultados obtenidos con la secuencia de la proteína 1ROP

Servidor	**RMSD Cα (Å)**
AbYSS	3.65
QUARK	1.46
I-TASSER	0.53

Tabla 4-12: RMSD local de la proteína 1ROP

Residuos (Å) / Servidor	3-28	32-55	RMSD Local Cα (Á)
ABYSS	1.14	1.27	2.41
QUARK	0.59	0.46	1.05
I-TASSER	0.27	0.28	0.55

Figura 4-7: Modelos generados para la proteína 1ROP



Datos de validación de la estructura generada:

Estructuras secundarias: se obtuvieron de la estructura generada tres hélices alfa entre los aminoácidos 2-27, 32-42 y 45-54; no hay láminas beta, como se observa en la Figura 4-8.

Figura 4-8: Elementos de estructuras secundarias de la proteína predicha 1ROP

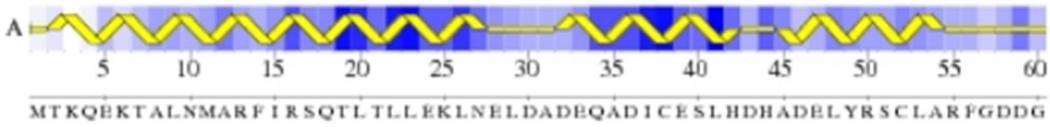
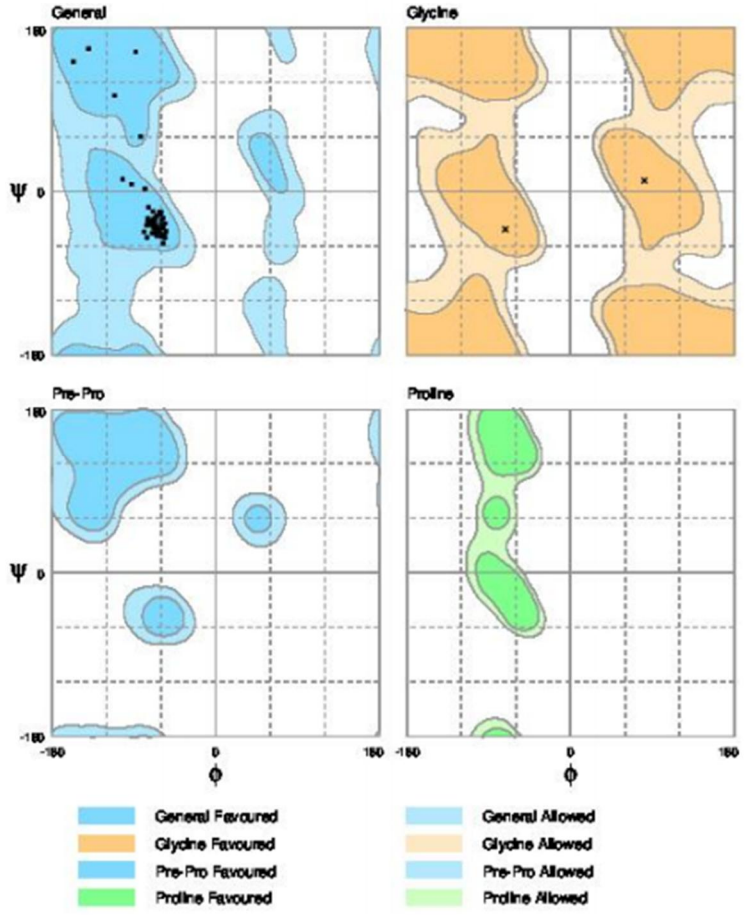


Gráfico de Ramachandran: como se muestra en la Figura 4-9, se puede establecer que los ángulos φ y ψ de los residuos se encuentran en zonas favorecidas y se hallan en zonas en las que se presentan hélices alfa. No se presentan valores atípicos en el gráfico y el 100% de los residuos están en zonas favorables.

Figura 4-9: Gráfico de Ramachandran para la estructura de la proteína 1ROP generada por el algoritmo AbYSS



Puntuación de calidad global de la estructura del modelo experimental reportado en el PDB:

Tabla 4-13: Resultados obtenidos con el modelo experimental reportado para la proteína 1ROP

Programa	Verify3D	Prosall (-ve)	Procheck (phi-psi)	Procheck (all)	MolProbity Clashscore
Mean score	0.21	0.55	0.83	0.52	6.88
Z-score	-4.01	-0.41	3.58	3.08	0.34

Número de contactos cercanos: 0

RMSD para ángulos de torsión: 3.6°

RMSD para longitud de enlaces: 0.033 Á

Puntuación de calidad global de la estructura modelo generado mediante el método AbYSS:

Tabla 4-14: Resultados obtenidos con el modelo generado para la proteína 1ROP

Programa	Verify3D	Prosall (-ve)	Procheck (phi-psi)	Procheck (all)	MolProbity Clashscore
Mean score	0.22	0.15	0.61	-0.72	509.32
Z-score	-3.85	-2.07	2.71	-4.26	-85.87

Número de contactos cercanos: 158

RMSD para ángulos de torsión: 0.3°

RMSD para longitud de enlaces: 0.011 Á

4.4 Resultados obtenidos con la proteína 2MQL

La proteína 2MQL está compuesta por 105 residuos y se clasifica como una proteína de unión de ARN. Su estructura fue predicha a través de RMN y, por tanto, posee varios modelos tridimensionales. Para este estudio se usó el modelo 1 con el objetivo de

realizar las respectivas comparaciones y superposiciones estructurales. Es una proteína que se origina en el organismo *Rattus norvegicus* y se caracteriza por no tener asignado un proceso biológico ni una componente celular; su función bioquímica es el enlace de ácidos nucleicos [52].

Los dominios de secuencia a los cuales pertenece son RNA *recognition motif domain* y *Nucleotide-binding alpha-beta plait domain*. No presenta dominios de estructura.

No posee anotaciones en las bases de datos CATH y SCOP.

En la Figura 4-10 se muestran el modelo experimental reportado en el PDB y los modelos generados a través de métodos computacionales. La estructura terciaria de esta proteína se caracteriza por tener tres hélices alfa entre los residuos 27-34, 59-63 y 66-69; mientras que entre los residuos 15-19, 40-46, 51-56, 74-75 y 80-85 se exhiben cinco láminas beta.

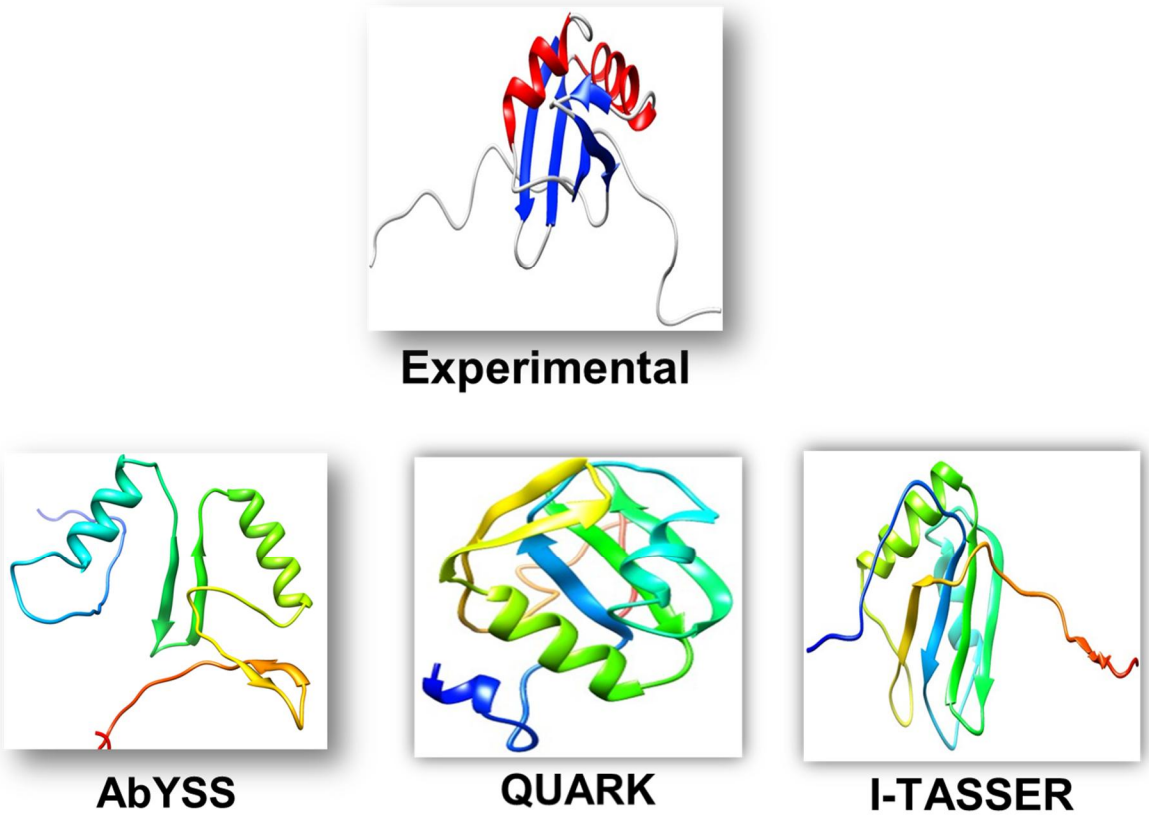
Tabla 4-15: Resultados obtenidos con la secuencia de la proteína 2MQL

Servidor	**RMSD Cα (Å)**
AbYSS	20,53
QUARK	10.36
I-TASSER	4.74

Tabla 4-16: RMSD local de la proteína 2MQL

Residuos (Å) / Servidor	**15-19**	**27-34**	**40-46**	**51-56**	**59-63**	**66-69**	**74-75**	**80-85**	**RMSD Local Cα (Å)**
ABYSS	1.14	0.38	2.21	0.95	0.22	0.43	0.30	0.89	6.55
QUARK	0.42	1.02	1.06	0.55	0.23	0.36	0.22	1.07	4.93
I-TASSER	0.43	0.53	0.5	0.21	0.23	0.37	0.22	0.48	2.97

Figura 4-10: Modelos generados para la proteína 2MQL



Datos de validación de la estructura generada:

Estructuras secundarias: se obtuvieron de la estructura generada dos hélices alfa entre los aminoácidos 27-36 y 59-69; y cuatro láminas beta entre los aminoácidos 43-46, 51-54, 82-84 y 89-91, como se observa en la Figura 4-11.

Figura 4-11: Elementos de estructuras secundarias de la proteína predicha 2MQL

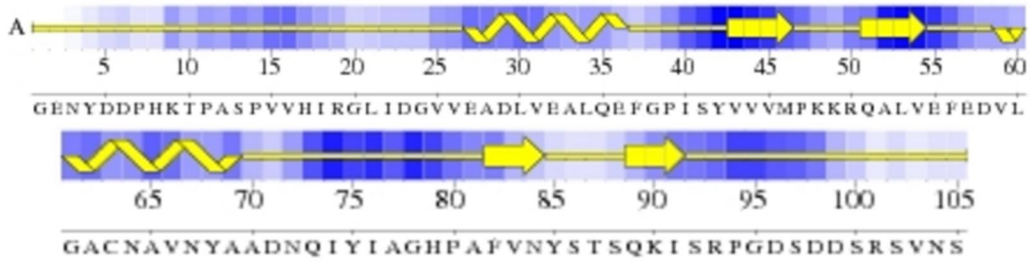
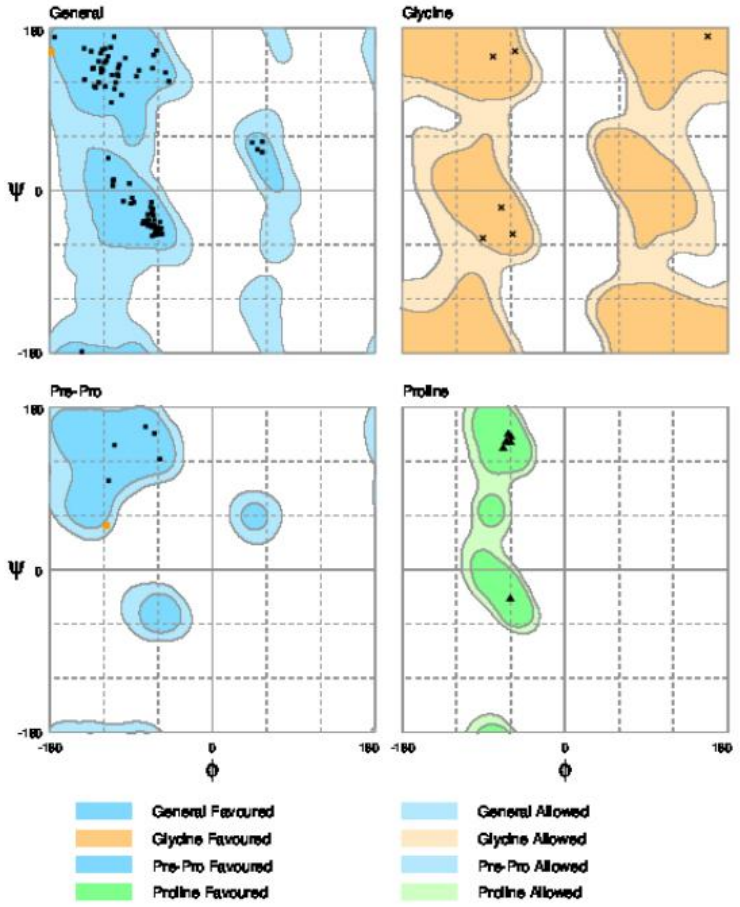


Gráfico de Ramachandran: de acuerdo con la Figura 4-12 se evidencia que los ángulos de torsión de 99 residuos se encuentran en zonas favorecidas, mientras que otros dos residuos se encuentran en zonas permitidas (puntos de color naranja en el esquema

gráfico general). Así mismo, no se presentan valores atípicos, lo cual indica que los ángulos están en el rango de las estructuras secundarias de las hélices alfa y láminas beta. De acuerdo con lo anterior, el 98.1% de los residuos se encuentran en regiones favorecidas, mientras que el 1.9% en regiones permitidas.

Figura 4-12: Gráfico de Ramachandran para la estructura de la proteína 2MQL generada por el algoritmo AbYSS



Puntuación de calidad global de la estructura del modelo experimental reportado en el PDB:

Tabla 4-17: Resultados obtenidos con el modelo experimental reportado para la proteína 2MQL

Programa	Verify3D	Prosall (-ve)	Procheck (phi-psi)	Procheck (all)	MolProbity Clashscore
Mean score	0.39	0.78	-0.68	-0.42	1.28
Z-score	-1.12	0.54	-2.36	-2.48	1.31

Número de contactos cercanos: 0

RMSD para ángulos de torsión: 1.4°

RMSD para longitud de enlaces: 0.009 Á

Puntuación de calidad global de la estructura modelo generado mediante el método AbYSS:

Tabla 4-18: Resultados obtenidos con el modelo generado para la proteína 2MQL

Programa	Verify3D	Prosall (-ve)	Procheck (phi-psi)	Procheck (all)	MolProbity Clashscore
Mean score	0.05	-0.21	-0.21	-0.98	248.23
Z-score	-6.58	-3.56	-0.51	-5.80	-41.07

Número de contactos cercanos: 145

RMSD para ángulos de torsión: 0.4°

RMSD para longitud de enlaces: 0.010 Á

4.5 Análisis de los resultados obtenidos

Al analizar los resultados obtenidos en esta investigación, se pudo establecer que los modelos generados con proteínas pequeñas (menor a 150 aminoácidos) fueron similares geométricamente con respecto a los modelos reportados experimentalmente, ya que su

RMSD global y local fueron menores a 4.5 Å, con excepción del modelo generado para la proteína 2MQL, en la cual estas medidas fueron superiores a dicho valor. Por tanto, el desempeño del método propuesto está relacionado con el tamaño de las proteínas, ya que de este depende el tamaño del espacio de búsqueda de conformaciones, así como de la cantidad de residuos en *loop*, los cuales son complejos de modelar debido a su flexibilidad en los ángulos de torsión. Así mismo, se validó con los gráficos de Ramachandran de cada modelo que los ángulos de torsión se movieron en regiones favorecidas y permitidas, como se estableció en el Capítulo 3 en la asignación de las variables de decisión del algoritmo AbYSS. De igual manera, se identificó que se presentan superposiciones estéricas desfavorables en las estructuras generadas, lo que puede obedecer a la posición de los átomos de las cadenas laterales (las cuales no son optimizadas) y esto se refleja en las medidas *Clashscore* y PROCHECK (*all*) de la validación de la estructura.

Para la proteína 1CRN se obtuvo una conformación con un valor RMSD global de 4.22 Å y un valor RMSD local de 1.16 Å, que indican la similitud con el modelo experimental reportado, como se muestra en la Tabla 4-3 y en la Tabla 4-4. Se observa que las hélices alfa se ubicaron en los residuos correctos como se muestra en la Figura 4-2 y en la Tabla 4-4 en los segmentos de residuos 7-17 y 23-30. Sin embargo, no hubo presencia de láminas beta a pesar de que los ángulos de torsión figuran en el rango favorecido. Así mismo, en los residuos en los cuales se presentan estructuras secundarias como hélices y láminas el RMSD local es menor a 1 Å, lo cual indica que los átomos del *backbone* se encuentran espacialmente con una configuración similar a los átomos del modelo reportado. Con respecto a la validación de la estructura, al comparar los valores obtenidos en el modelo reportado y en el modelo generado, se establece de acuerdo con la Tabla 4-5 y la Tabla 4-6 que la conformación obtenida tiene puntajes similares en algunas de las medidas, no obstante la diferencia en el valor del *Clashscore* es alta, lo cual significa que en la conformación generada se presentan superposiciones estéricas desfavorables. De acuerdo con las medidas de PROCHECK sobre los ángulos phi y psi los puntajes no son inusuales, mientras que para todos los ángulos de torsión sí reporta valores altamente inusuales.

Para la proteína 2KBQ el modelo generado es similar al reportado experimentalmente al observar su geometría; de acuerdo con los resultados presentados en la Tabla 4-7 y en

la Tabla 4-8 se establece que el valor RMSD global es de 4.42 Å y el valor RMSD local es de 4.16 Å, lo cual indica que en los segmentos en los cuales la estructura secundaria es una hélice al superponer las estructuras la distancia máxima fue de 1.69 Å entre los residuos 5-16, debido a que la hélice inicial se formó entre los residuos 2-12 y 14-17, dejando los ángulos del residuo 13 en el rango de los *loop*. Por tanto, las cinco hélices de la estructura 2KBQ se formaron, aunque es necesario minimizar más la energía de la estructura con el objetivo de ajustar los ángulos de algunos residuos que no se encuentran en el valor indicado, a pesar de que el 100% de estos se encuentran en la región favorecida según el gráfico de Ramachandran de la Figura 4-6. Con respecto a la validación de la estructura, al comparar los valores obtenidos en el modelo reportado y en el modelo generado, se establece de acuerdo con la Tabla 4-9 y la Tabla 4-10 que la conformación obtenida tiene puntajes (Z-score) similares en algunas de las medidas, siendo destacable la diferencia entre el valor del *Clashscore* que es alta, lo cual significa que hay superposiciones estéricas desfavorables en la estructura generada. De acuerdo con las medidas de PROCHECK para los ángulos phi y psi los puntajes son números positivos, lo cual indica un buen valor, en contraste con el puntaje de todos los ángulos de torsión cuyos valores son altamente inusuales.

Para la proteína 1ROP el modelo generado es similar geométricamente al modelo reportado experimentalmente, como se puede establecer al comparar las estructuras presentadas en la Tabla 4-11 y en la Tabla 4-12, en las cuales se observa que el RMSD global es de 3.65 Å y el RMSD local es de 2.41 Å, lo cual indica que los modelos son similares ya que se presentan dos hélices que conforman esta proteína, no obstante como se muestra en la Figura 4-8 la hélice formada entre los residuos 32-55 está interrumpida en el segmento de residuos 42-44 en los cuales se presenta un *loop,* a pesar de encontrarse en el intervalo favorecido de ángulos de torsión de acuerdo con el gráfico de Ramachandran de la Figura 4-9; así mismo, se evidencia que en el giro presente entre las dos hélices (residuos 29-31) se observa que hay un ángulo de abertura mucho mayor. Por tanto se establece que es necesario minimizar más la estructura, con el propósito de que las hélices alfa sean formadas en su totalidad y, de esta manera, los ángulos de torsión de los residuos que se encuentran en la forma indicada sean ajustados. Con respecto a la validación de la estructura, al comparar los valores obtenidos en el modelo reportado y en el modelo generado, se establece de acuerdo con la Tabla 4-13 y la Tabla 4-14 que los puntajes Z-score de la conformación

generada difieren de los obtenidos en el modelo reportado, en especial, en la medida de todos los ángulos de torsión de PROCHECK en la que se presentan valores altamente inusuales y superposiciones estéricas desfavorables en la estructura generada según el *Clashscore.* Sin embargo, los valores de PROCHECK para los ángulos phi y psi presentaron números positivos en ambos casos, lo cual indica un buen Z-score.

El modelo generado para la proteína 2MQL tiene un valor RMSD global de 20.3 Å y un valor RMSD local de 7.17 Å, de acuerdo con la Tabla 4-15 y con la Tabla 4-16. El primero indica que espacialmente el modelo generado es distante geométricamente del modelo reportado experimentalmente, posiblemente debido a la cantidad de residuos en *loop*, en los cuales los ángulos φ y ψ se mueven libremente en el rango [-180º,180º], mientras que el valor RMSD local muestra que algunos ángulos de torsión de los residuos involucrados en esos segmentos de la estructura se encuentran bien ubicados.

Sin embargo, como se muestra en la Figura 4-11 se formaron sólo dos hélices alfa de tres que se presentan en la estructura reportada, y cuatro láminas beta de cinco posibles que tiene el modelo del PDB. Al observar los residuos en los que se formaron estas estructuras secundarias se observa que la hélice presente entre los residuos 59-69 abarca las dos hélices que se presentan en los segmentos 59-63 y 66-69, convirtiendo en hélice el segmento 64-65 que es un *loop,* esto debido a que posiblemente los ángulos de torsión de ese segmento coincidieron con los de una hélice.

Así mismo, las láminas presentes en los residuos 15-19 y 74-75 no se formaron debido, posiblemente, a la ausencia de más puentes de hidrógeno entre estos residuos; mientras que la lámina que se formó en los residuos 89-91 no corresponde con los motivos estructurales de la proteína reportada, ya que ese segmento debería ser un *loop;* esto se pudo presentar por las combinaciones de ángulos de torsión que quedaron en el rango de las láminas beta, así como muestra el gráfico de Ramachandran de la Figura 4-12. Con respecto a la validación de la estructura, los valores de las medidas entre el modelo generado y el modelo reportado experimentalmente presentadas en la Tabla 4-17 y la Tabla 4-18 son similares, con excepción de los obtenidos en PROSAII y el Clashscore de Molprobity, lo cual indica que la estructura generada presenta superposiciones estéricas desfavorables y errores en los pliegues de la misma.

Igualmente, al comparar los modelos generados con los modelos de los servidores de proteínas I-TASSER y QUARK se observa que las estructuras generadas por el método propuesto tienen un buen desempeño, ya que el RMSD en todos los modelos fue similar al presentado por estos, es decir, que la distancia no fue superior a 3 Å, con excepción del modelo de la proteína 2MQL en la cual la diferencia es más alta posiblemente por la gran cantidad de residuos en *loop,* como se explicó anteriormente. Así mismo, se compararon tres métodos con cuales se puede predecir la estructura terciaria de las proteínas, como son: *ab initio*, reconocimiento del plegamiento (*threading)* y el modelo propuesto en esta investigación que usa ensamblaje de fragmentos, obteniendo buenos resultados mediante los mismos. También se puede afirmar que el comportamiento de I-TASSER en proteínas de mayor tamaño es mejor, debido al uso de plantillas de proteínas homólogas.

5. Conclusiones y recomendaciones

5.1 Conclusiones

En la presente investigación se desarrolló un método híbrido para la predicción de la estructura tridimensional de las proteínas a partir de su secuencia de aminoácidos, que consistió en el uso del ensamblaje de fragmentos como insumo, con el fin de reducir el espacio conformacional y, así mismo, obtener una estructura geométricamente similar a la reportada para luego refinarla con el algoritmo de búsqueda dispersa AbYSS.

Los resultados obtenidos satisfacen los objetivos propuestos que consistieron en seleccionar un modelo de representación computacional de las proteínas, establecer un método de búsqueda de conformaciones, utilizar una función de puntuación que guiara el proceso de búsqueda e integrar todo lo anterior en una herramienta computacional; posteriormente, se realizó la validación y el análisis del desempeño del modelo propuesto con respecto a otros resultados reportados en la literatura.

Las representaciones computacionales de las proteínas que se utilizaron fueron adecuadas, ya que facilitaron la manipulación de las conformaciones tanto a nivel de coordenadas cartesianas como de coordenadas trigonométricas.

El método de búsqueda AbYSS proporcionó características de exploración y explotación del espacio de conformaciones, así como las cualidades de una optimización multiobjetivo en la cual el frente de Pareto se caracterizó por tener los individuos mejor evaluados y más diversos que permitió evitar posibles mínimos locales y realizar una minimización de la función de puntuación satisfaciendo los diferentes objetivos. Igualmente, el enfoque utilizado para determinar las variables de decisión fue apropiado, teniendo en cuenta que no se modificaron las regiones que tenían una geometría correcta, es decir, segmentos de la proteína que presentaron hélices alfa o láminas beta de acuerdo con las estructuras secundarias proporcionadas por PSIPRED y DSSP, que permitieron la reducción de las variables de decisión y del espacio de búsqueda.

Con respecto a la función de puntuación Talaris2013, esta direccionó el algoritmo de búsqueda de manera correcta, lo cual se refleja en la similitud estructural de los resultados obtenidos con respecto a los modelos reportados experimentalmente.

Finalmente, se compararon las estructuras generadas con el método propuesto contra las reportadas en el PDB mediante la superposición estructural, obteniendo resultados con un valor RMSD global menor a 5 Å en la mayoría de los modelos; así mismo los gráficos de Ramachandran mostraron que los ángulos de torsión de los residuos se movieron en el intervalo favorecido para hélices y láminas. Del mismo modo, se validaron las estructuras generadas mediante la comparación de los datos obtenidos con las estructuras reportadas mediante el servidor PSVS.

5.2 Recomendaciones

De acuerdo con los resultados obtenidos en este trabajo, y con el propósito de mejorar la calidad de los mismos, se recomienda:

- Paralelizar la evaluación de la función de puntuación y el cálculo de la estructura secundaria de las conformaciones, ya que estos son procesos de un muy alto costo computacional.
- Realizar experimentación adicional con otros valores de los parámetros del algoritmo, así como ejecutar pruebas con nuevas proteínas de diversos tamaños, con el propósito de complementar la validación del modelo utilizado.
- Adicionar información sobre las cadenas laterales, para que con la ayuda de librerías de rotámeros se puedan modelar y, de este modo, construir un modelo más preciso para evitar colisiones estéricas entre los átomos de las mismas.
- Buscar una estrategia que permita un mejor modelamiento de los residuos en *loop.*

A. Anexo: Términos de la función de energía Talaris2013

Los términos que componen la función de puntuación Talaris2013 son [45] [46]:

- fa_atr: Lennard-Jones atractivo entre átomos de diferentes residuos.
- fa_rep: Lennard-Jones repulsivo entre átomos de diferentes residuos.
- fa_sol: energía de solvatación de Lazaridis-Karplus.
- fa_intra_rep: Lennard-Jones repulsivo entre átomos en el mismo residuo.
- fa_elec: potencial electrostático de Coulomb con una distancia-dependiente dieléctrica.
- pro_close: energía del anillo de prolina y la energía del ángulo Ψ del residuo precedente.
- hbond_sr_bb: puentes de hidrógeno en el *backbone* cercanos en secuencia.
- hbond_lr_bb: puentes de hidrógeno en el *backbone* lejanos en secuencia.
- hbond_bb_sc: enlaces de hidrógeno entre el *backbone* y las cadenas laterales.
- hbond_sc: enlaces de hidrógeno entre cadenas laterales.
- dslf_fa13: enlaces de disulfuro.
- rama: preferencias de Ramachandran.
- omega: preferencias del ángulo omega en el *backbone.*
- fa_dun: energía interna de los rotámeros de las cadenas laterales que se deriva de las estadísticas de Dunbrack (biblioteca de rotámeros 2010).

- p_aa_pp: probabilidad de observar un aminoácido particular dados los ángulos Φ y Ψ.

- ref: energía de referencia para cada aminoácido. Equilibra la energía interna de términos de aminoácidos.

A continuación se exponen las ecuaciones de los términos de energía nombrados anteriormente:

Preferencias de ángulos de torsión de Ramachandran

$$rama = \sum_{i} -\ln[P(\phi_i \psi_i | aa_i, ss_i)]$$

donde:

i: Índice del residuo

ϕ, ψ: Ángulos de torsión del *backbone*

aa: Tipo de aminoácido

ss: Tipo de estructura secundaria asignada por DSSP

Interacciones de Lennard-Jones

$$LJ = \sum_{i} \sum_{j>i} \begin{cases} \left[\left(\frac{r_{ij}}{d_{ij}}\right)^{12} - 2\left(\frac{r_{ij}}{d_{ij}}\right)^{6}\right] e_{ij}, & si \ \frac{d_{ij}}{r_{ij}} > 0.6 \\ \left[-8759.2\left(\frac{d_{ij}}{r_{ij}}\right) + 5672.0\right] e_{ij,} & otro\ caso \end{cases}$$

donde:

i, j: Índices del residuo

d: Distancia interatómica

e: Media geométrica del átomo más inferior del pozo potencial

r: Sumatoria de radios de van der Waals

Nota 1: esta función no es evaluada para pares de átomos en los cuales la distancia interatómica depende de los ángulos de torsión de un solo residuo.

Nota 2: los radios son determinados mediante el ajuste de las distancias de los átomos de estructuras determinadas por rayos X con los potenciales 6 y 12 de Lennard Jones usando CHARMm19.

Puentes de hidrógeno

$$hb = \sum_i \sum_j \left(-\ln\left[P\left(d_{ij}|h_j ss_{ij}\right)\right]\right) - \ln\left[P\left(\cos\theta_{ij}\,|d_{ij}h_j ss_{ij}\right)\right] - \ln\left[P\left(\cos\psi_{ij}\,|d_{ij}h_j ss_{ij}\right)\right]$$

donde:

i: Índice del residuo donante

j: Índice del residuo aceptor

d: Distancia interatómica del aceptor-protón

h: Hibridación

ss: Tipo de estructura secundaria, las cuales son asignadas como helicoidales (j – i = 4, cadena principal); láminas (|j - i| > 4, cadena principal), u otra.

θ: Base del ángulo del enlace protón-aceptor-aceptor

ψ: Ángulo del enlace donante-protón-aceptor

Nota: evaluada solo para pares de donantes y aceptores en los cuales d se encuentra en el rango $1{,}4 \leq d \leq 3{,}0$ y $90° \leq \psi, \theta \leq 180°$. Puentes de hidrógeno de la cadena lateral involucrados con átomos de la cadena principal no son tenidos en cuenta.

Solvatación

$$hb^f = \sum_i \left[\Delta G_i^{ref} - \sum_j \left(\frac{2\Delta G_j^{free}}{4\pi^{3/2}\lambda_i r_{ij}^2} e^{d_{ij}^2} V_j + \frac{2\Delta G_j^{free}}{4\pi^{3/2}\lambda_j r_{ij}^2} e^{-d_{ij}^2} V_i\right)\right]$$

donde:

i, j: Índices del átomo

d: Distancia entre átomos

r: Sumatoria de los radios de van der Waals

λ: Longitud de la correlación

V: Volumen atómico

$\Delta G^{ref}, \Delta G^{free}$: Energía de un átomo solvatado completamente

Nota: los valores son tomados de Lazaridis y Karplus

Interacciones entre un par de residuos (electrostáticas y disulfidos)

$$pair = \sum_{i} \sum_{j>i} -\ln\left[\frac{P(aa_i, aa_j|d_{ij})}{P(aa_i|d_{ij})P(aa_j|d_{ij})}\right]$$

donde:

i, j: Índices del residuo

aa: Tipo de aminoácido

d: Distancia entre residuos

Energía de los rotámeros

$$dun = \sum_{i} -\ln\left[\frac{P(rot_i|\emptyset_i, \psi_i)P(aa_i|\emptyset_i, \psi_i)}{P(aa_i)}\right]$$

donde:

i, j: Índices del residuo

rot: Rotámero de Dunbrack *backbone*-dependiente

aa: Tipo de aminoácido

Φ, ψ: Ángulos de torsión del *backbone*

Energía de referencia del estado desplegado

$$ref = \sum_{aa} n_{aa}$$

donde:

aa: Tipo de aminoácido

n: Número de residuos

Bibliografía

[1] L. S. Jeremy M. Berg, John L. Tymoczko, *Biochemistry*, Fifth Edit. W.H.Freeman & Co Ltd, 2002.

[2] J. R. Gunn, "Protein Structure Prediction: Methods and Protocols Edited by David M. Webster (Southern Cross Molecular, Bath, UK). Methods in Molecular Biology 143. Humana Press: Totowa, NJ. 2000. x + 422 pp. $89.50. ISBN 0-89603-637-5.," *J. Am. Chem. Soc.*, vol. 123, p. 1796, 2001.

[3] J. Gu and P. Bourne, *Structural bioinformatics*. 2009.

[4] W. Widłak, *Molecular Biology Not Only for Bioinformaticians*. 2013.

[5] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, and J. Watson, *Biología molecular de la célula*, Tercera Ed. Ediciones Omega, S.A, 1996.

[6] H. Lodish, "Molecular cell biology," 2008.

[7] R. C. Bohinski, *Bioquímica*, Quinta Edi. Addison Wesley Longman de México, 1998.

[8] G. John, C. Rose, and S. Takeuchi, "Understanding Tools and Techniques in Protein Structure Prediction," *Syst. Comput. …*, 2011.

[9] "molecularsciences.org," 2007. [Online]. Available: http://www.molecularsciences.org/structural_bioinformatics/protein_structures. [Accessed: 02-Sep-2014].

[10] N. Heda, "B for Biology," 2013. [Online]. Available: http://namrataheda.blogspot.com/2013_03_01_archive.html. [Accessed: 04-Sep-2014].

[11] D. Voet and J. Voet, *Bioquímica*, Tercera Ed. Media Panamericana, 2006.

[12] Y. Xu, *Computational methods for protein structure prediction and modeling*. 2007.

[13] M. Hoque, M. Chetty, and A. Sattar, "Genetic Algorithm in Ab Initio Protein Structure Prediction Using Low Resolution Model: A Review," *Biomed. Data Appl.*, pp. 317–342, 2009.

[14] A. L. Jaimes and C. A. Coello, "An introduction to multi-objective evolutionary algorithms and some of their potential uses in biology," *Stud. Comput. Intell.*, vol. 122, no. 2008, pp. 79–102, 2008.

[15] S. Russell and P. Norvig, *Inteligencia Artificial. Un enfoque moderno. 2da Edición*. 2004.

[16] R. Martí and M. Laguna, “Búsqueda Dispersa,” *Business*, 1977.

[17] A. J. Nebro, F. Luna, and E. Alba, “Un algoritmo multiobjetivo basado en b ´ usqueda dispersa.”

[18] Y. Xu, *Computational methods for protein structure prediction and modeling*. 2007.

[19] C. Anfinsen, “Principles that govern the folding of protein chains,” *Science (80-.).*, vol. 181, no. 4096, pp. 223–230, 1973.

[20] R. Zwanzig, A. Szabo, and B. Bagchi, “Levinthal’s paradox,” *Proc. …*, vol. 89, no. January, pp. 20–22, 1992.

[21] S. D. J. Alas-guardado and A. Rojo, “La paradoja de Levinthal: cuando una contradicción se vuelve lógica / Levinthal paradox: When a contradiction turns logical,” vol. 22, no. 1, pp. 51–54, 2011.

[22] S. Seguí, “Plegamiento y funcionalidad biológica del inhibidor de metalocarboxipeptidasas LCI (Leech Carboxypeptidase Inhibitor),” 2004.

[23] A. a. Nickson and J. Clarke, “What lessons can be learned from studying the folding of homologous proteins?,” *Methods*, vol. 52, pp. 38–50, 2010.

[24] B. Honig, “Protein folding: from the levinthal paradox to structure prediction.,” *J. Mol. Biol.*, vol. 293, no. 2, pp. 283–93, Oct. 1999.

[25] S.-Q. Liu, X.-L. Ji, Y. Tao, D.-Y. Tan, K.-Q. Zhang, and Y.-X. Fu, “Protein folding, binding and energy landscape: A synthesis,” *Protein Eng.*, pp. 207–252, 2012.

[26] J. Lau, “Protein structure database for structural genomics group,” The State University of New Jersey, 2005.

[27] A. Murzin and S. Brenner, “SCOP: a structural classification of proteins database for the investigation of sequences and structures,” *J. Mol. …*, pp. 536–540, 1995.

[28] J. Xiong, *Essential bioinformatics*. 2006.

[29] Z. Xiang, “Advances in homology protein structure modeling,” *Curr. Protein Pept. Sci.*, vol. 7, no. 3, pp. 217–227, 2006.

[30] Z. Zhang, “An Overview of Protein Structure Prediction: From Homology to Ab Initio,” 2002.

[31] R. Bitello and H. S. Lopes, “A Differential Evolution Approach for Protein Folding,” *Computer (Long. Beach. Calif).*

[32] H. Guo, “Solving 2D HP Protein Folding Problem by Parallel Ant Colonies,” *Lect. Notes Comput. Sci.*, pp. 0–4, 2009.

[33] M. T. Hoque, M. Chetty, and L. S. Dooley, “A Guided Genetic Algorithm for Protein Folding Prediction Using 3D Hydrophobic-Hydrophilic Model,” *2006 IEEE Int. Conf. Evol. Comput.*, pp. 2339–2346, 2006.

[34] R. Bonneau and D. Baker, “A B I NITIO P ROTEIN S TRUCTURE PREDICTION :,” 2001.

[35] J. C. Calvo, J. Ortega, and M. Anguita, “PITAGORAS-PSP: Including domain knowledge in a multi-objective approach for protein structure prediction,” *Neurocomputing*, vol. 74, no. 16, pp. 2675–2682, Sep. 2011.

[36] J. Zhang and S. Li, “A Simulating Algorithm for Protein Folding Process,” *2009 WRI Int. Conf. Commun. Mob. Comput.*, pp. 343–347, Jan. 2009.

[37] J. C. Calvo, J. Ortega, and M. Anguita, “Comparison of parallel multi-objective approaches to protein structure prediction,” *J. Supercomput.*, vol. 58, no. 2, pp. 253–260, Dec. 2009.

[38] S. Fidanova and I. Lirkov, “Ant colony system approach for protein folding,” *2008 Int. Multiconference Comput. Sci. Inf. Technol.*, pp. 887–891, Oct. 2008.

[39] D. Xu and Y. Zhang, “Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field.,” *Proteins*, vol. 80, no. 7, pp. 1715–35, Jul. 2012.

[40] C. Hardin, T. V Pogorelov, and Z. Luthey-schulten, “Ab initio protein structure prediction,” *Constraints*, pp. 176–181.

[41] R. Kowalski and M. Sergot, “Computer representation of,” *Computational Chemistry*, 1996. [Online]. Available: http://www.ccl.net/cca/documents/molecular-modeling/node4.html. [Accessed: 04-Sep-2015].

[42] A. J. Nebro, F. Luna, E. Alba, B. Dorronsoro, J. J. Durillo, and A. Beham, “AbYSS: Adapting scatter search to multiobjective optimization,” *IEEE Trans. Evol. Comput.*, vol. 12, no. 4, pp. 439–457, 2008.

[43] R. Becerra and D. Camilo, “A Multi-objective Ab-initio Model for Protein Folding Prediction at an Atomic Conformation Level,” no. February, pp. 1–11, 2010.

[44] A. Leaver-Fay, M. J. O’Meara, M. Tyka, R. Jacak, Y. Song, E. H. Kellogg, J. Thompson, I. W. Davis, R. a. Pache, S. Lyskov, J. J. Gray, T. Kortemme, J. S. Richardson, J. J. Havranek, J. Snoeyink, D. Baker, and B. Kuhlman, *Scientific Benchmarks for Guiding Macromolecular Energy Function Improvement*, vol. 523. 2013.

[45] R. Y. R. Wang and F. Dimaio, “Tutorial :* Rosetta * tools * for * structure * determination * in ** cryoEM * density !,” pp. 1–5, 2015.

[46] “Energy terms in Rosetta.” [Online]. Available: https://www.rosettacommons.org/docs/latest/rosetta_basics/scoring/score-types.

[47] T. H. Cormen, C. E. Leiserson, and R. Rivest, *Introduction to Algorithms*. 1990.

[48] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin, “UCSF Chimera—A visualization system for exploratory research and analysis,” *J. Comput. Chem.*, vol. 25, no. 13, pp. 1605–1612, 2004.

[49] R. Maiti, G. H. Van Domselaar, H. Zhang, and D. S. Wishart, “SuperPose: a simple server for sophisticated structural superposition.,” *Nucleic Acids Res.*, vol. 32, no. Web Server issue, pp. W590–4, 2004.

[50] A. Bhattacharya, R. Tejero, and G. T. Montelione, “Evaluating protein structures determined by structural genomics consortia,” *Proteins: Structure, Function, and Bioinformatics*, 2006. [Online]. Available: http://doi.wiley.com/10.1002/prot.21165.

[51] S. C. Lovell, I. W. Davis, W. B. Arendall, P. I. W. de Bakker, J. M. Word, M. G. Prisant, J. S. Richardson, and D. C. Richardson, “Structure validation by Calpha geometry: phi,psi and Cbeta deviation.,” *Proteins*, vol. 50, no. 3, pp. 437–450, 2003.

[52] PDBe, “The Protein Data Bank in Europe ,” 2013. [Online]. Available: http://www.ebi.ac.uk/pdbe/.