

Universidad Central “Marta Abreu” de Las Villas



Facultad Matemática – Física – Computación

Centro de Estudios Informáticos

Laboratorio de Bio-Informática

2014 – 2015

Algoritmo Recocido Simulado para predecir la estructura terciaria de proteínas

Autor:

Fernando Wong Delgado (fwong@uclv.edu.cu)

Tutores:

Lic. Yasser B. Ruiz Blanco (yasserrb@uclv.edu.cu)

Lic. Elizabeth Martínez Pérez (elizabethmp@uclv.cu)

Dr. Gladys Casas Cardoso (gcasas@uclv.edu.cu)

ANEXO ACTA DE CONFORMIDAD PARA ESTUDIANTES DE PREGRADO



Universidad Central "Marta Abreu" de Las Villas

ACTA DE CONFORMIDAD

Por una parte:

Fernando Wong Delgado estudiante de la carrera de: Ingeniería Informática en la facultad de Matemática-Física-Computación, en lo adelante **EL ESTUDIANTE**. Con número de identidad permanente: 91081430506.

Y por otra parte:

Gheisa Ferreira Lorenzo Jefe del Departamento Docente de: Ingeniería Informática en la ya mencionada facultad, en lo adelante **EL JEFE DE DEPARTAMENTO**, y Yasser B. Ruiz Blanco, Elizabeth Martínez Pérez y Gladys Casas Cardoso profesores encargados de tutorar el Trabajo de Diploma **DEL ESTUDIANTE**, en lo adelante **EL TUTOR**.

Reconocen que:

- I. A **EL ESTUDIANTE** se le ha aprobado como tema de investigación para su Trabajo de Diploma el titulado: Algoritmo Recocido Simulado para predecir la estructura terciaria de proteínas.
- II. **EL ESTUDIANTE** no divulgará información concerniente a la investigación, tanto durante el desarrollo como tras la culminación de esta sin la debida autorización **DEL TUTOR o EL JEFE DE DEPARTAMENTO**.

- III. Que el Trabajo de Diploma fruto de la labor investigativa **DEL ESTUDIANTE** y la asesoría **DEL TUTOR**, resulta de **TITULARIDAD EXCLUSIVA** de la Universidad Central “Marta Abreu” de las Villas.
- IV. **EI ESTUDIANTE** una vez aprobada su tesis para la defensa, depositará una copia electrónica de la misma en el Repositorio Digital Institucional de la Universidad Central “Marta Abreu” de las Villas.
- V. A partir de la defensa y aprobación del Trabajo de Diploma, la publicación total, parcial o la elaboración de cualquier obra que se derive de esta investigación por parte **DEL ESTUDIANTE**, contará con la coautoría **DEL TUTOR** y viceversa, resultando de referencia obligada esta obra en cualquier otra que se elabore. El incumplimiento de esta cláusula, puede llevar consigo el inicio de procesos de plagio. Todo lo anterior de acuerdo a la normativa de Derecho de Autor vigente en Cuba.

Y para que así conste se firma la presente en la Universidad Central “Marta Abreu” de Las Villas, a los ____ días del mes de _____ del año 20____.

Fernando Wong Delgado

Estudiante

Gheisa Ferreira Lorenzo

J’ de Departamento

Lic. Yasser B. Ruiz Blanco

Tutor

Lic. Elizabeth Martínez Pérez

Tutor

Dr. Gladys Casas Cardoso

Tutor

La investigación que se propone tiene como objetivo predecir la estructura terciaria de las proteínas a partir de la secuencia de aminoácidos. Ello, a su vez, podría potenciar el desarrollo de nuevos medicamentos y servir de base para el desarrollo en nuestro país de una de las áreas más novedosas de la Biotecnología. Los estudios relacionados a este problema han tratado de simplificar algunos aspectos como la función de energía y el espacio de posibles conformaciones, aplicando técnicas de optimización heurísticas. Se implementó la metaheurística Recocido Simulado con la función objetivo propuesta por Ruiz-Blanco (2014) y cuatro métodos de enfriamiento (Boltzmann, Cauchy, Geométrico, Lundi-Mess). Para la selección de las soluciones se obtuvieron 30 réplicas de cada proteína por configuración seleccionando de ellas cinco con un análisis de clúster para realizar las comparaciones. Se compararon los resultados entre cuatro métodos de enfriamiento, obteniendo que los resultados favorables en cuanto a calidad los tiene el método de Lundi-Mess. La comparación de los resultados contra las soluciones de VMO, arrojó como resultados la superación de la metaheurística SA a VMO en cuatro de las cinco métricas medidas. De la comparación que se realizó con soluciones de predictores internacionales podemos destacar que el promedio de nuestras soluciones se presentan en el segundo cuartil, además la metaheurística implementada proporciona un mejor alineamiento global (GDT) que local (LGAS).

Palabras Claves: meta-heurística, recocido simulado, estructura terciaria de proteínas, servidor LGA.

Abstract

The objective of the proposed research is to predict the protein structure from a sequence of aminoacids. This could help to develop new medications and serve as base for the development of our country in one of the most novel areas in Biotechnology. The studies related to this problem have been about simplifying some aspects like the energy function and the space of possible conformations, applying techniques of heuristic optimization. The Simulated Annealing metaheuristic was implemented, using the objective function proposed by Ruiz-Blanco (2014) and four cooling methods (Boltzmann, Cauchy, Geometric, Lundy-Mess). For the selection of the solutions we obtained 30 replicas of each protein by configuration, selecting five of them with a cluster analysis to carry out the comparisons. The results of the four cooling methods were compared, obtaining favorable results in terms of quality when using the Lundy-Mess method. The comparison of the results against the VMO solutions shows that the SA metaheuristic is able to outperform VMO in four of the five measured metrics. From the comparison using international predictors we can notice that the average of our solutions is located in the second quartile. Besides, the implemented metaheuristic provides a better global alignment (GDT) than local (LGAS).

Key Words: met-heuristic, simulating annealing, folding proteins, LGA server.

Dedicatoria:

“Educar es depositar en cada hombre toda la obra humana que le ha precedido... es hacer de cada hombre resumen del mundo viviente; es preparar al hombre para la vida”

José Martí

“Cien veces al día me recuerdo a mí mismo que mi vida interior y exterior se alimentan de los esfuerzos de otros hombres, vivos y muertos, y que debo aferrarme por dar en la misma medida en que he recibido”

Albert Einstein

Dedico mi trabajo:

A mis Padres que con su amor y entrega infinita contribuyeron a que me convirtieran en quien soy hoy. Por su apoyo incondicional y su entrega sin límites, por ser quienes son y estar ahí siempre que los necesite con toda su sabiduría y cariño. A mi tía Beba y mi prima Silvita, quienes a pesar de la distancia con su amor incondicional siempre han tenido fe en mí y creyeron que con mi sacrificio y esfuerzos cualquier proyecto que me propusiese podía ser posible y tener éxito, facilitaron mi vida y colaboraron en todo para que este día fuera posible. A mis abuelos Celina y Jorge que me han acompañado siempre a lo largo de todos estos años con su amor y cariño infinito, brindándome toda su experiencia, siempre con el consejo oportuno para convertirme en el hombre de bien que soy hoy. A mi abuela Ita que aunque hoy ya no está con nosotros, donde quiera que este cuida y vela por mi salud, abuela te extraño muchísimo. A todos ellos mil gracias por existir en mi vida.

Agradecimientos

Agradecimientos:

Agradezco a mis tutores Elizabeth, Gladita y Yasser que me ayudaron y estuvieron a mi disposición cuando los necesite, que en todo momento con su sabiduría e inteligencia me asesoraron para que esta investigación pudiera llevarse con éxito, porque siempre confiaron en mi para realizarlo, agradezco profundamente el tiempo que he pasado con ustedes y todo el conocimiento que me han transmitido y que repercutirá satisfactoriamente en mi futura vida profesional. Mejor que ustedes nunca los hubiera encontrado, gracias.

Agradezco profundamente a mi hermanita Claudia por haberte conocido, por estar ahí siempre para mí cuando lo necesite, por ser incondicional y comprensiva, por hacer más llevadera mi vida con su cariño sin límites, Felicia la vida nos depara muchas cosas buenas.

A mi tía Mercedita por su amor y cuidado incondicional hacia mí, por estar siempre que la necesite, por sus sabios consejos para hacer de mí un hombre de bien y por la confianza que siempre me has dado.

A Edel que desde mis primeros años de vida siempre me guió como un verdadero padre, apoyándome, cuidándome, confiando en mí, demostrándome su amor y gracias por ayudarme a levantar cuando he estado caído, gracias papá.

A mis hermanos de siempre Diana, Mario Ángel, Ernesto, Jorge Félix y Yordano, mil gracias por ser quienes son, por todas las experiencias que me han brindado, por no juzgarme, por darme apoyo siempre que lo he necesitado, ustedes son mi familia.

A mis padres Rosmery y Mayito y mi hermano Luis Enrique, gracias por tener fe en mí, apoyarme y guiarme en los momentos más difíciles de mi vida, demostrándome que con sacrificio y tenacidad siempre se logra lo que se quiere.

A Yailen mi profesora y amiga que siempre me ayudó cuando lo necesite y nunca puso objeciones a pesar de sus múltiples ocupaciones y de la hora que la requiriera.

Agradecimientos

A Carmen y Cusi por su cariño y apoyo en todo como si fuera su propio hijo, por sus sabios consejos que me hicieron mucho bien.

A mi compañero y amigo en todos estos años de carrera universitaria Alejandro Roche (Pelusi) por ayudarme cada vez que lo necesite, por compartir conmigo como un hermano, usted es mi hermano.

A mis compañeros de aula y a mis profesores por todos los recuerdos que me han brindado y que voy atesorar toda la vida, por las enseñanzas que me han transmitido, muchas gracias.

A las personas que me apoyaron en mis años de universidad gracias por lo que me enseñaron de la vida y que me forjaron como persona.

A mi novia por su apoyo, comprensión y cariño en este tiempo de mucho estrés y conflictos, eres muy linda lo sabes, te quiero mucho.

Contenidos

Introducción.....	1
I. Marco Teórico	6
I.1. Estructura de la Proteína.....	7
I.1.1 Jerarquía de estructuras.....	8
I.2. Definición del problema de plegamiento de proteínas.....	9
I.3. Métodos computacionales para el plegamiento de proteínas	11
I.3.1 Recocido Simulado.....	12
I.3.2 Seudocódigo general del Recocido Simulado	14
I.3.3 Optimización basada en Mallas Variables	15
I.3.4 Comparación entre VMO y Recocido Simulado.....	16
I.4. Competencia CASP-ROLL.....	17
I.5. Conclusiones parciales del capítulo.	19
II. Diseño e Implementación.....	20
II.1. Elementos y diseño del problema	21
II.1.1 Proteína Computacional	21
II.2. Representación espacial.....	22
II.2.1 Coordenadas internas y formato de la Z-Matriz.....	22
II.2.2 Método de conversión de coordenadas internas a cartesianas.....	24
II.3. Función de Amplitud.....	24
II.3.1 Restricciones del problema	28
II.3.2 Espacio de búsqueda del problema.....	29
II.4. Recocido Simulado aplicado al problema del plegamiento de proteínas	30
II.4.1 Diagrama de Actividad de la metaheurística RS.....	31

Contenidos

II.4.2	Diagrama de Actividad del RS-Paralelo.....	33
II.4.3	Diagrama de Actividad de cómo encontrar el Sucesor	34
II.4.4	Métodos de enfriamientos	35
II.5.	Implementación.....	36
II.5.1	Atributos utilizados para el algoritmo	36
II.6.	Conclusiones parciales del capítulo.	37
III.	Análisis de los Resultados.....	38
III.1.	Procedimiento para comparar la calidad de las soluciones	39
III.2.	Alineamiento Local – Global (LGA) de estructuras de proteínas	41
III.2.1	Métricas calculadas por el servidor LGA	41
III.2.2	Parámetros asociados al servidor LGA	43
III.3.	Comparación de los resultados.....	43
III.3.1	Comparación entre los cuatro métodos de enfriamiento	44
III.3.2	Comparación entre el método SA-enfriamiento y VMO	45
III.3.3	Comparación entre el método SA-LundiMess y VMO contra predicciones internacionales	46
III.4.	Conclusiones del capítulo	47
	Conclusiones.....	49
	Recomendaciones.....	51
	Referencias Bibliográficas	53
	Anexos	57
I.	Datos Generales del proceso de selección de la proteína R0007	58
II.	Datos Generales del proceso de selección de la proteína R0015.....	59

Contenidos

III.	Datos Generales del proceso de selección de la proteína R0023.....	60
IV.	Datos Generales del proceso de selección de la proteína R0033.....	61
V.	Comparación entre los métodos de enfriamiento en las proteínas	62
VI.	Comparación entre las metaheurísticas SA y VMO en las proteínas	63
VII.	Comparación entre SA y VMO contra predicciones internacionales.	64

Índice de Tablas

Tabla I-1: Los 20 aminoácidos naturales que conforman proteínas y su clasificación según su afinidad al agua. El * indica que no todos los científicos coinciden con la clasificación que se muestra.	8
Tabla III-1: Nombre de las proteínas y sus respectivas longitudes	39
Tabla III-2: Atributos generales del SA y sus valores	40
Tabla III-3: Atributos específicos por configuración.....	40
Tabla III-4: Generalización de las métricas por método de enfriamiento.....	44
Tabla III-5: Generalización de las métricas por metaheurísticas	45
Tabla III-6: Cantidad de soluciones por competidores	46

Índice de Figuras

Figura II-1: Diagrama de Actividad de la metaheurística RS	31
Figura II-2: Diagrama de Actividad del RS-Paralelo	33
Figura II-3: Diagrama de Actividad de cómo encontrar el Sucesor	34
Figura III-1: Comparación entre los cuatro métodos de enfriamiento	45
Figura III-2: Comparación entre SA y VMO	46
Figura III-3: Comparación entre SA, VMO y competidores internacionales	47

Introducción

La Bioinformática y la Biología Computacional son áreas que se ubican en la frontera entre la Biología y las Ciencias de la Computación. El principal objetivo de estas disciplinas es el desarrollo y uso de técnicas matemáticas y computacionales para ayudar en la resolución de problemas de la biología, en particular de la biología molecular.

En lo que se refiere al impacto dentro de las ciencias computacionales, la bioinformática incluye problemas de gran complejidad. Uno de estos problemas de alta complejidad es la predicción de la estructura terciaria de proteínas que constituye una de las áreas más enigmáticas de la biología molecular, pues, integra aspectos relacionados no solo con la predicción, sino además con la codificación y diseño de su estructura tridimensional. Durante más de 30 años, investigadores de las áreas de Biología, Química, Física, Matemática y Computación han abordado el asunto de manera independiente o en equipos multidisciplinarios, obteniendo hasta ahora resultados limitados.

En este ámbito de investigación la predicción de la estructura terciaria de la proteína obliga preguntarse, dada la secuencia lineal de aminoácidos que conforman una proteína, ¿cuál es la estructura tridimensional correspondiente?, cuestionamiento que constituye el problema científico de la presente investigación.

La investigación que se propone puede lograr predecir la estructura terciaria de las proteínas a partir de la secuencia de aminoácidos. Ello, a su vez, podría potenciar el desarrollo de nuevos medicamentos y servir de base para el desarrollo en nuestro país de una de las áreas más novedosas de la Biotecnología.

Actualmente existen aproximadamente unas 68 048 estructuras proteicas identificadas (H. Berman, 2000, Mikael Oliveberg, 1995, Berman, 2008), (Mikael Oliveberg, 1995), cada una conformada en promedio por 50 a 1500 aminoácidos. A pesar de que la secuencia de los aminoácidos de una proteína se obtiene fácilmente, usando técnicas y reacciones casi completamente automatizadas, el proceso por el que estas asumen su estructura nativa aún no se conoce bien.

Es principalmente el conocimiento de la estructura terciaria, y no la secuencia primaria, lo que permite a los bioquímicos predecir el comportamiento de una proteína. Por lo anterior existe una gran necesidad de obtener la estructura terciaria de las proteínas que se hallen sometidas a investigación.

Los estudios relacionados a este problema han tratado de simplificar algunos aspectos como la función de energía y el espacio de posibles conformaciones, aplicando técnicas de optimización heurísticas como algoritmos de Ascenso de Colina (Hill Climbing), Recocido Simulado (B Suman) (E Aarts, 1990, Berman, 2008), Monte Carlo (2001, Berman, 2008) y Algoritmos Evolutivos (Escuela, 2006, Berman, 2008).

A pesar de este esfuerzo, el asunto sigue representando un reto cuando se trabaja con proteínas reales cuya longitud siempre incluye varias decenas de aminoácidos. Según los resultados internacionales, este tipo de modelos parecen ser prometedores en el estudio del proceso de plegamiento de las proteínas, aunque han sido poco aplicados en este contexto.

Desde esta contradicción planteamos el problema científico que dio origen al presente trabajo. Se reitera en este punto que como los métodos heurísticos generales no siempre pueden explicar procesos singulares David Wolpert (1997), como el del plegamiento de las proteínas, el modelo que aquí se propone puede acercarse a la solución de dicha problemática con las consiguientes ventajas que ello reportaría a las ciencias biológicas y médicas de la nación. (D.H Wolpert, 1997, Berman, 2008)

Enmarcado en el contexto de meta-heurísticas para la solución de problemas complejos, en el grupo de Investigación **Unit of Computer-Aided Molecular “Biosilico” Discovery and Bioinformatics (CAMD-BIR)** de la facultad de Química-Farmacia necesita diseñar e implementar un método para el pliegue de proteínas utilizando la metaheurística Recocido Simulado. Hill Climbing es un algoritmo incompleto porque puede estancarse en óptimos locales y no alcanzar el óptimo global pero eficiente, dado el tamaño del espacio de soluciones. Por el contrario, un algoritmo que se moviese hacia un sucesor elegido de forma uniformemente aleatoria de entre un

conjunto de sucesores sería completo pero muy ineficiente. El algoritmo Recocido Simulado (RS) está a medio camino entre estos extremos combina el Hill Climbing con el seguimiento de un camino aleatorio de modo que se pueda conseguir tanto eficiencia como completitud.

De aquí la posibilidad de probar el comportamiento de este algoritmo frente a un problema tan complejo y de una naturaleza real dando origen a las siguientes preguntas de investigación:

- ✓ ¿Será posible adaptar la metaheurística de Recocido Simulado para obtener mejores resultados que los encontrados actualmente?
- ✓ ¿Se encontrarán estructuras tridimensionales que satisfagan las necesidades de los especialistas?

Teniendo en cuenta lo antes mencionado nos hemos trazado como **objetivo general**: Implementar la metaheurística Recocido Simulado para obtener un modelo tridimensional óptimo de proteínas, conociendo únicamente su secuencia de aminoácidos.

Para lograr un mejor desarrollo del objetivo general se ha subdividido en los siguientes **objetivos específicos**:

- ✓ Diseñar la metaheurística Recocido Simulado para la predicción de la estructura terciaria de proteínas.
- ✓ Implementar y evaluar la metaheurística Recocido Simulado con cuatro métodos de enfriamiento.
- ✓ Realizar la comparación estadística de los resultados entre esta metaheurística y la de Optimización basada en Mallas Variables (VMO) en cuanto a calidad de las soluciones.

En función de la problemática antes descrita se formuló la siguiente **hipótesis de investigación**: Emplear la metaheurística Recocido Simulado para la predicción de estructura terciaria de proteínas permitirá obtener resultados mejores o comparables a la metaheurística Optimización basada en Mallas Variables.

Aporte

El aporte de la tesis es que se logra comparar los resultados de la metaheurística a implementar con la que ya se implementó brindando información muy importante para la investigación y los especialistas ya que se busca una solución que sea factible en cuanto a calidad y parámetros de rendimiento computacional.

Justificación

Esta investigación se ve justificada porque se propone que puede lograr predecir la estructura terciaria de las proteínas a partir de la secuencia de aminoácidos. Ello a su vez podría potenciar el desarrollo de nuevos medicamentos y servir de base para el desarrollo en nuestro país de una de las áreas más novedosas de la Biotecnología.

Este trabajo consta de tres capítulos y está estructurado de la forma siguiente

En el Capítulo 1, “Marco Teórico”, se describe el plegamiento de proteínas y se realiza una revisión general sobre modelos meta heurísticos y de otra índole que aborda el plegamiento proteico y finalmente se muestra un resumen de los mejores métodos y estrategias aplicados en esta área de la Biología computacional.

En el Capítulo 2 “Diseño e Implementación” se describe la metodología desarrollada para afrontar el problema del plegamiento de proteínas, así como algunas implementaciones y consideraciones tomadas durante el diseño. Por último se describe brevemente la meta heurística Recocido Simulado para problemas continuos.

En el Capítulo 3 “Análisis de los resultados” se detallan los experimentos realizados con el fin de evaluar el método propuesto, y se presentan los resultados obtenidos con los resultados que ya se obtuvieron del VMO.

I. Marco Teórico

En este capítulo se presentan los aspectos biológicos relacionados con el trabajo, entre los que se destaca la descripción de la estructura jerárquica de las proteínas. Por otra parte, se hace referencia acerca del estado del arte de las metodologías computacionales para resolver el problema de predicción de la estructura terciaria, se expone una breve teoría sobre la hipótesis termodinámica. Por último, se detalla las estrategias del modelado para el problema, particularmente se referencia el modelado ab initio.

I.1. Estructura de la Proteína

Las proteínas son las macromoléculas más abundantes y funcionalmente más versátiles, que forman parte de las células, constituyendo en definitiva el producto final de los genes.

Las proteínas están conformadas por una variedad de 20 aminoácidos que se encadenan entre sí mediante enlaces peptídicos, combinándose en distinto orden y número (de 50 a 1500 aproximadamente). Cada una de estas secuencias o cadenas, adopta espacialmente una estructura tridimensional que determina la funcionalidad biológica de la proteína. Esta estructura contiene regiones que permiten el acoplamiento de otras proteínas para formar estructuras más complejas, o para bloquear el funcionamiento de otras.

Los enlaces peptídicos se forman por la interacción química entre dos aminoácidos, liberando una molécula de agua. Una de las características fisicoquímicas más importantes que diferencian los aminoácidos entre sí, es su afinidad o repulsión al agua, lo que permite clasificar a estos monómeros en polares e hidrofóbicos, respectivamente (ver Tabla I-1).

De acuerdo a su morfología y solubilidad, las proteínas se clasifican en fibrosas y globulares. Las proteínas fibrosas son generalmente estructurales o de soporte, como la elastina y el colágeno. Por otra parte, las proteínas globulares tienden hacer más solubles en agua y presentan una forma compacta y casi esférica.

Capítulo I: Marco Teórico

Las mayorías de las proteínas conocidas son globulares, como las enzimas, las proteínas del plasma y las presentes en las membranas celulares.

Tabla I-1: Los 20 aminoácidos naturales que conforman proteínas y su clasificación según su afinidad al agua. El * indica que no todos los científicos coinciden con la clasificación que se muestra.

Aminoácido	Código	Letra	Clasificación
Ácido aspártico	Asp	D	Polar
Ácido glutámico	Glu	E	Polar
Alanina	Ala	A	Hidrofóbico*
Arginina	Arg	R	Polar
Asparagina	Asn	N	Polar
Cisteina	Cys	C	Hidrofóbico*
Fenilalanina	Phe	F	Hidrofóbico
Glicina	Gly	G	Polar*
Glutamina	Gln	Q	Polar
Histidina	His	H	Polar*
Isoleucina	Ile	I	Hidrofóbico
Leucina	Leu	L	Hidrofóbico
Lisina	Lys	K	Polar*
Metionina	Met	M	Hidrofóbico
Prolina	Pro	P	Hidrofóbico*
Serina	Ser	S	Polar
Tirosina	Tyr	Y	Hidrofóbico*
Triptófano	Trp	W	Hidrofóbico
Treonina	Thr	T	Polar
Valina	Val	V	Hidrofóbico

I.1.1 Jerarquía de estructuras

El nivel más básico de la estructura proteica, llamado estructura primaria, es la secuencia lineal de aminoácidos que está determinada a su vez, por el orden de los nucleótidos en el ácido desoxirribonucleico (ADN) o en el ácido ribonucleico (Jarosław Pillardy et al.) . Considerando el número posible de aminoácidos, la conformación o secuencia primaria de una proteína puede ser abstraída como una cadena de caracteres sobre un alfabeto $\mathcal{S}=\{A;C;B;E;F;G;H;I;K;L;M;N;P;Q;R;S;T;V;W;Y\}$ de tamaño

20, donde cada símbolo representa la letra que corresponde al aminoácido, según la tabla 1.1 (Quintero, 2011).

Fuerza como los enlaces de hidrógenos, los puentes disulfuro la atracción entre cargas positivas y negativas y los enlaces hidrofóbicos (repelentes al agua) e hidrofílicos (afines al agua) hacen que la molécula se repliegue y adopte una estructura secundaria. Las estructuras secundarias más comunes se denominan α -hélices, hojas- β y giros.

Aparte los patrones recurrentes de estructuras secundarias se denominan estructuras supersecundarias, como hélice –giro-hélice, cremallera de leucinas, horquilla α -hélices. Cuando las fuerzas provocan que la molécula se vuelva todavía más compacta, como ocurre en las proteínas globulares, se constituye una estructura terciaria donde la secuencia de aminoácidos adquiere una conformación tridimensional, denominado estado nativo. Se dice que la molécula tiene estructura cuaternaria cuando está formada por más de una cadena poli peptídica como ocurre en la hemoglobina y en algunas enzimas (Quintero, 2011).

Los dominios son las unidades funcionales y tridimensionales de un poli péptido el núcleo de un dominio está construido por combinaciones de estructuras supersecundarias denominadas también motivos.

Algunas proteínas tienen la propiedad de desempeñar funciones múltiples, adicionales o secundarias a su función principal, mediante la utilización de un mismo dominio o de dominios distintos. Por otra parte existen proteínas denominadas mosaicos, que consisten en varias copias repetidas de uno o más dominios, todos dentro de una cadena poli peptídicas muchas proteínas celulares presentan esta característica. Los dominios en cuestión son llamados módulos y en ocasiones son relativamente cortos (Quintero, 2011).

I.2. Definición del problema de plegamiento de proteínas

“El Plegamiento de Proteínas” es el término utilizado para describir el proceso complejo en el que las cadenas poli peptídicas adoptan su conformación tridimensional “nativa”.

Para llevar a cabo sus funciones deben plegarse con rapidez y lo suficientemente estable, razón por la que deben cumplir el siguiente requerimiento cinético; “que el plegamiento se complete en un tiempo razonable” y el siguiente requisito termodinámico: “que la conformación plegada sea estable bajo condiciones fisiológicas” (Levinthal, 1969).

Si una proteína está compuesta por 100 aminoácidos, hay cerca de 10^{49} posibles conformaciones de la cadena poli peptídica completa.

Una característica importante del plegamiento de proteínas es que la secuencia de aminoácido de la proteína determina su estructura única general (Anfinsen, 1973), que es una combinación de estructuras secundarias (las regiones de α -hélices, hojas- β) y la estructura terciaria (patrón de plegamiento general).

Las diferencias en las secuencias dan lugar a diferencias en la estructura secundaria y terciaria. Hasta ahora, las estructuras tridimensionales de aproximadamente 6000 proteínas han sido determinadas por cristalografía de rayos X y espectroscopia de RMN (Resonancia Magnética Nuclear). Los dominios de estas proteínas pueden ser agrupados en aproximadamente 350 familias de pliegues, que consisten en secuencias con estructuras similares (Chothia, 1992).

Es decir, la evolución ha seleccionado las secuencias con un mínimo de energía profundo para el estado nativo, lo que elimina las estructuras mal plegadas o parcialmente desplegadas a temperaturas fisiológicas. El proceso de plegamiento de proteínas es uno de los procesos biofísicos fundamentales. Dicho proceso resulta de interés debido al importante papel que desempeña en los mecanismos y controles de una amplia variedad de procesos celulares. Estos incluyen la regulación de eventos complejos durante el ciclo celular y la translocación de proteínas a través de las membranas de los órganos (ML Riley, 1997).

Por otra parte, la falta de proteínas a veces se asocia con el mal funcionamiento de los sistemas biológicos, dando lugar a una amplia gama de enfermedades. Algunas de estas, como el Alzheimer y las enfermedades de Creutzfeldt-Jakob, están asociadas a la conversión de proteínas solubles en insolubles en placas amiloides y agregados de

fibrillas (Prusiner, 1998). Otras como la fibrosis quística, por ejemplo, son el resultado de mutaciones que dificultan el plegamiento normal y la secreción de proteínas específicas (SC Hyde, 1990).

Como cualquier otra reacción química, la comprensión del plegamiento de proteínas requiere del conocimiento de las interacciones que lo dominan, así como una visión de la cinética y la dinámica del proceso.

Como parte del interés mundial para solucionar esa problemática, pueden mencionarse dos hechos recientes de gran trascendencia. Por un lado, la creación de un grupo de evaluadores críticos de predicción de la estructura terciaria CASP (Nicholas Hamilton, 2004, Mikael Oliveberg, 1995, Jarosław Pillardy et al., 2001, Richard Bonneau and Baker, 2001),(Mikael Oliveberg, 1995) (Critical assessment of structure prediction) que se reúnen bianualmente desde el año 1994.

Por el otro, la incursión de IBM en el estudio del plegado de las proteínas dentro de un proyecto de supercomputación, cuyo principal actor, el Blue Gene/L, sucesor del famoso jugador de ajedrez Big Blue, está completamente instalado desde el 2005 con una velocidad de 360 teraflops y con un sistema formado por unos 130.000 procesadores.

I.3. Métodos computacionales para el plegamiento de proteínas

Los algoritmos metaheurísticos son aquellos algoritmos que emplean una heurística para resolver una clase muy amplia de problemas mediante la combinación de distintos procedimientos. La mayoría de los problemas resueltos por los algoritmos metaheurísticos son de optimización combinatoria o en su defecto deben ser adaptados en esta forma. Estos algoritmos metaheurísticos suelen ser utilizados en problemas donde no existe un algoritmo para hallar soluciones en tiempo polinomial. (Aponte, 2006)

El poder ser aplicado a una clase muy general de problemas tiene muchos beneficios en problemas NP, ya que a ninguno de éstos se le conoce una solución específica. Sin

embargo, para otro tipo de problemas, esta característica puede no ser tan beneficiosa. El uso indiscriminado de metaheurísticas puede llevar ignorar algún algoritmo específico, las metaheurísticas tienen herramientas para muy diversos usos, pero a veces una herramienta específica se presta mejor a una situación. Aún no se conoce ninguna heurística efectiva y especializada a ser aplicada al problema de predicción de estructuras de las proteínas (Aponte, 2006).

A medida que ha ido avanzando las investigaciones relacionadas con la predicción de la estructura terciaria de las proteínas se ha hecho necesario encontrar métodos de búsqueda de soluciones aplicables para este tipo de problema. Se ha experimentado con algunas metaheurísticas (VMO, Recocido Simulado, etc.) obteniendo resultados satisfactorios. En este acápite se expondrán los aspectos esenciales de las metaheurísticas mencionadas anteriormente ya que son parte esencial del objetivo de investigación de esta tesis de grado.

I.3.1 Recocido Simulado

El Recocido Simulado (RS), también denominado Simulated Annealing (SA), Enfriamiento Simulado, Método Descendiente Probabilístico o Relajación Estocástica, es un algoritmo de búsqueda metaheurística de optimización global. En el marco de la metalurgia esta técnica consiste en calentar el material para dotar a los átomos de energía y que así puedan desplazarse. Posteriormente, se deja enfriar el material lentamente para hacer que recristalice en configuraciones con menor energía que la inicial. Cuanto más lento sea el procedimiento de enfriamiento del material más probabilidad tendremos de conseguir un cristal perfectamente estructurado con baja energía interna. Las ideas en la que se basa el algoritmo fueron introducidas en Metrópolis et al (1953), en un algoritmo para simular el proceso de enfriamiento antes descrito. Posteriormente, fue desarrollado independientemente por Scott Kirkpatrick, C. Daniel Gelatt y Mario P. Vecchi en 1983, y por Vlado Černý en 1985 que propusieron que este algoritmo podía ser empleado para resolver problemas de optimización (Gallardo, 2014).

Simulated Annealing es una técnica que utiliza enfriamiento controlado para reducir los defectos en los metales. En el proceso de recocido el calor causa que los átomos se vean liberados de sus posiciones iniciales (mínimo local) y en consecuencia se paseen aleatoriamente por estados de mayor energía.

Por analogía el algoritmo SA busca en el espacio de soluciones mediante un iterador estocástico. En cada iteración el algoritmo escoge un vecino de la solución y se calcula la energía (función a minimizar) del vecino. Con este valor se calcula la probabilidad de transición. El valor de esta función (que toma en cuenta además de la energía, la temperatura) determina si la nueva solución es tomada en cuenta o no. Esta función permite que vecinos con peores soluciones sean tomados en cuenta cuando se comienza a iterar y disminuye luego la probabilidad de que peores soluciones sean escogidas cuando ya la solución va más avanzada.

El parámetro más importante del algoritmo SA es la generación de vecinos. La selección del vecino puede ser aleatoria pero dirigida en cierta forma (Aponte, 2006).

Como el nombre del algoritmo sugiere, la variación de la temperatura durante la computación es una característica distintiva de este método y su principal ventaja. El algoritmo comienza con un valor de T muy alto, que va decreciendo en cada iteración siguiendo un cierto *protocolo de recocido*, que puede ser diferente para cada problema, pero que siempre debe terminar con $T=0$. Así el sistema será libre inicialmente de explorar una gran porción del espacio de búsqueda, ignorando pequeñas variaciones de la energía entre los estados vecinos evaluados, para más tarde centrarse en regiones con estados de baja energía y, al final, cambiar solo a estados con energía menor que la inicial, hasta alcanzar un mínimo. La probabilidad de que el algoritmo acabe encontrando el mínimo global para un problema dado se aproxima a 1 a medida que el protocolo de recocido se extiende.

Hay autores que han trabajado con Recocido Simulado para la predicción de la estructura terciaria de las proteínas. En el año 2002 Yoshida y coautores realizaron experimentos utilizando un híbrido de SA para paralelo con operadores de cruce

genéticos. Se han publicado muchos artículos e investigaciones del Recocido Simulado para la predicción de la estructura terciaria de las proteínas como (Aponte, 2006) que realizó una investigación para utilizar Recocido Simulado y Búsqueda Tabú para el problema de predicción de la estructura terciaria de proteínas publicado en el 2006 en la Universidad Simón Bolívar.

I.3.2 Seudocódigo general del Recocido Simulado

```
▶ Ciclo externo
▶     Seleccionar estado = So;
▶     Seleccionar it = 0;
▶     Ciclo interno
▶         Seleccionar siguiente= sucesor_aleatorio (estado);
▶         Función objetivo DE = valor (siguiente) – valor (estado);
▶         Si DE < 0
▶             Entonces estado = siguiente;
▶         Sino
▶             Seleccionar q = min (1,  $e^{-DE/T}$ );
▶             Si aleatorio (0,1) < q
▶                 Entonces estado = siguiente;
▶         it = it + 1;
▶     Mientras que it = tope;
▶     So = estado;
▶ To = enfriar (To);
▶     t = t + 1;
▶     Mientras que To = ~ 0
▶     Retornar So;
```

A continuación se presenta la descripción del pseudocódigo general de la metaheurística RS el que parte desde el ciclo externo; dentro de este ciclo se declaran la variable de estado a la cual se le asigna la solución inicial (So) y un contador de las iteraciones (ML Riley) que es igual a 0.

Después se ingresa en un ciclo interno; dentro del cual se realiza la búsqueda aleatoria de un sucesor el que se asigna a la variable siguiente, se encuentra el valor de la función objetivo (DE) que se calcula: la función de calidad del sucesor menos la función de calidad del estado inicial.

Luego se evalúa la condición de $DE < 0$ si se cumple se actualiza el valor de la variable estado que tomaría el valor del siguiente esto quiere decir que el sucesor es mejor que el estado actual.

Si no se cumple la condición se crea una variable q que obtendrá el mínimo entre 1 y una función que depende de la temperatura y se evalúa otra condición en la cual se genera un numero aleatorio entre 0 y 1 y este número tiene que ser menor que el valor de q para que se acepte el sucesor como el estado actual esto quiere decir que si el sucesor es peor que el estado actual todavía tiene la posibilidad de tomarse si cumple con los anteriores requisitos y esto le permite a la metaheurística poder aceptar soluciones peores para evitar estancamientos en óptimos locales.

Más adelante se actualiza el contador de la iteración y el ciclo interno se realiza hasta que la iteración llegue al máximo de iteraciones a ejecutar; ya fuera del ciclo interno dentro del externo se van a actualizar las variables declaradas fuera de los ciclos como son la solución inicial que va a recibir el valor del estado actual, se actualiza la temperatura inicial que tendrá el resultado del método de enfriamiento de la temperatura, se actualiza además un contador de temperatura (t) y el ciclo externo se va a realizar hasta que la temperatura inicial sea menor o aproximado a 0.

Por último después de que se ejecuten los ciclos se retorna la So que recibe el mejor nodo encontrado hasta el momento.

I.3.3 Optimización basada en Mallas Variables

La esencia del método Optimización basada en Mallas Variables (VMO por sus siglas en inglés) es crear una malla de puntos en el espacio m dimensional, donde se realiza el proceso de optimización de una función $FO(x_1, x_2, \dots, x_m)$; la cual se mueve mediante un proceso de expansión hacia otras regiones del espacio de búsqueda.

El proceso de generación de nodos está comprendido por varios pasos como: generación de la malla inicial la cual consta de N_i nodos, los cuales en la primera iteración son generados de forma aleatoria o por otro método que garantice obtener soluciones diversas. La generación de nodos en dirección a los extremos locales constituye otro tipo de exploración que se realiza en VMO se lleva a cabo en las vecindades de cada uno de los nodos de la malla inicial.

Al igual que en el paso anterior, la generación de nodos en dirección hacia el extremo global tiene como propósito realizar una exploración global hacia el nodo que mejor calidad ha tenido hasta el momento (extremo global).

La generación de nodos a partir de los nodos más externos de la malla es un proceso más de expansión de nuevos nodos, que tiene lugar con el objetivo de explorar el espacio de búsqueda en dirección a las fronteras de cada dimensión.

La cantidad de nodos de la malla inicial (N_i) es muy importante al definir el tamaño de las poblaciones para obtener un buen desempeño de estos métodos. Para todos los casos, el tamaño total de la expansión es $N = 3N_i + N_i/2$, con el objetivo de generar con el operador de las fronteras una cantidad de nodos, al menos, igual a la mitad de la malla inicial. Además, el tamaño de la vecindad (k) y la condición de parada (M) son parámetros que utiliza el modelo y juegan un rol importante en el desempeño del mismo (Quintero, 2011).

El autor (Quintero, 2011) que ha implementado la metaheurística VMO para la predicción de la estructura terciaria de las proteínas; realizó una investigación en la Universidad Central “Marta Abreu” de Las Villas presentada como tesis de grado.

I.3.4 Comparación entre VMO y Recocido Simulado

La metaheurística VMO se caracteriza por ser un algoritmo poblacional que genera una malla inicial con una cantidad de nodos equivalente a: 3 veces la longitud de la proteína lo que permite realizar una mayor exploración del espacio de búsqueda. VMO también es lento computacionalmente debido a su característica de expansión-contracción.

A diferencia de la metaheurística Recocido Simulado que es un algoritmo que no es poblacional tiene una menor exploración del espacio de búsqueda; además Recocido Simulado ya ha sido utilizado con anterioridad para el problema de plegado de proteínas (Aponte, 2006), existen estudios realizados de como descender la temperatura y su principal ventaja es que garantiza poder salir de óptimos locales (Pardo, 2012-2013).

Una característica que tienen en común estas dos metaheurísticas es que utilizan las probabilidades para recorrer el espacio de búsqueda. Debido a las ventajas que presenta Recocido Simulado hemos decidido implementar este para comparar sus salidas contra VMO.

I.4. Competencia CASP-ROLL

El evento CASP-ROLL (Rolling Critical Assessment of Structure Predictions) es una competencia continua de predicción de estructuras de proteínas, derivada del evento más general CASP cuya frecuencia es bienal. El CASP surgió en el año 1994 en respuesta al marcado desbalance que se produjo entre el número de secuencias conocidas de proteínas y el número de estructuras elucidadas. Debido a las dificultades experimentales y el tiempo requerido para determinar la estructura de una proteínas es que ya en aquella fecha se apostaba por el desarrollo de métodos computacionales de predicción de estructura de proteínas y la competencia brindo el marco apropiado para que los principales investigadores de la rama confrontasen sus métodos en una prueba a ciegas y se pudiese determinar que algoritmos lograban los mejores resultados en este problema. Dando lugar a un espacio para el intercambio entre los investigadores y la publicación de los mejores métodos. En la actualidad la competencia sigue vigente y cuenta con una gran reputación científica y diversidad de predictores participantes, siendo evidencia de las dificultades del problema del plegamiento de proteínas y los avances que aún son necesarios hacer en este campo.

La competencia se organiza de modo que los grupos que en distintos laboratorios en el mundo se dedican a la elucidación experimental de estructuras de proteínas envían a

los organizadores la secuencia de las proteínas que están próximos a publicar. Los organizadores dan entonces un plazo limitado para que los distintos predictores computacionales envíen un máximo de cinco modelos estructurales para cada proteína ofertada. Una vez los experimentalistas hacen llegar las estructura reales, los organizadores evalúan las predicciones y publican los resultados junto con la estructura de las proteínas y los distintos modelos propuestos por los diferentes predictores.

El CASP_ROLL surgió hace dos años como un espacio donde se agrupan aquellas proteínas más complejas de modelar por poseer una secuencia con baja homología con secuencias existentes y donde por tanto los métodos basados en conocimiento no brindan los mejores resultados. De aquí que en este evento participan mayormente métodos *ab initio*, los cuales básicamente realizan una optimización de la estructura partiendo de candidatos aleatorios y empleando una determinada función de ajuste y un método de exploración apropiado.

Los métodos *ab initio* constan fundamentalmente de tres etapas en su algoritmo:

- i) Exploración, donde es usual restringir el espacio de búsqueda empleando por ejemplo bibliotecas de fragmentos prediseñados y conformando las estructuras a modo de puzle.
- ii) Selección, se clusterizan decenas o centenares de soluciones obtenidas de múltiples optimizaciones en la primera etapa, seleccionando un subconjunto reducido de candidatos.
- iii) Refinamiento, esta etapa comprende la aplicación de métodos comúnmente muy costosos computacionalmente como las Dinámicas Moleculares para tunear localmente las conformaciones de las cadenas de los candidatos seleccionados anteriormente.

Los experimentos presentados en la tesis difieren de los predictores *ab initio* contemporáneos en:

- i) Realizar una exploración no restringida del espacio de búsqueda.
- ii) La selección se realiza en una población total de solo 30 soluciones.

- iii) No se realizan etapas de refinamiento.

I.5. Conclusiones parciales del capítulo.

Al mostrar la descripción de la estructura jerárquica de las proteínas y definir el problema del plegamiento de proteínas podemos tener una mejor comprensión de los factores biológicos que originan enfermedades como “Alzheimer”, “Creutzfeldt-Jakob”, entre otras, que afectan a la población mundial y para las que se realizan estudios en búsqueda de una cura. Debido a la complejidad (NP) del problema de plegamiento de proteínas es necesario la utilización de metaheurísticas para encontrar una buena solución para la predicción de la estructura terciaria de la proteína y para resolver este problema se seleccionó la metaheurística SA debido a sus ventajas y características que garantiza que la metaheurística pueda salir de los óptimos locales favoreciendo una mejor búsqueda del óptimo global.

II. Diseño e Implementación

En este capítulo se hace una descripción a cada uno de los aspectos computacionales que fueron aplicados en esta investigación durante el proceso de modelación del plegamiento de proteínas. Se muestra una revisión detallada de la función objetivo y las restricciones, finalmente se presenta una descripción sobre el algoritmo de optimización basado en Recocido Simulado, la representación general de la metaheurística y como se aplica al problema de plegamiento de proteínas.

En los siguientes epígrafes que tratan los temas relacionados con los conceptos químicos y estructuras moleculares que son parte de la solución del problema de investigación, se ha obtenido la información para esta tesis, de la tesis de pregrado (Quintero, 2011) en el que se argumentan con mayor profundidad.

II.1. Elementos y diseño del problema

El problema fundamental que se propone solucionar en esta tesis trata de encontrar la estructura terciaria de una proteína específica teniendo en cuenta un parámetro esencial que es su energía mínima; utilizando una función de evaluación de la calidad se definirá si la estructura terciaria encontrada es lo suficientemente buena.

La solución planteada para el problema es un algoritmo Recocido Simulado el cual se va a ejecutar en paralelo para realizar una mayor exploración del espacio de búsqueda. Primero se genera una población atendiendo a la cantidad de hilos o cantidad de veces que se ejecutara la metaheurística RS de forma paralela esto quiere decir que se utiliza la metaheurística RS para cada elemento de la población. Dependiendo del porcentaje que se le haya asignado en el fichero de los datos de entrada a las variables que representan a las hélices alfas y las hojas betas depende la estructura terciaria de las proteínas que conforman la población.

II.1.1 Proteína Computacional

La estructura de las proteínas reúne las propiedades de disposición en el espacio de las moléculas de proteína que provienen de su secuencia de aminoácidos, las características físicas de su entorno y la presencia de compuestos, simples o complejos

que las estabilicen y/o conduzcan a un plegamiento específico, distinto del espontáneo. Por ello, deriva de sus componentes, es decir de la propia estructura de los aminoácidos, de cómo interaccionan químicamente éstos, de forma jerarquizada y específica, y evidentemente está en relación con la función a acometer en el destino celular (Quintero, 2011). Para lograr todo un funcionamiento de la misma se implementó un paquete denominado Protein, compuesto por un conjunto de clases que definen la estructura de la misma, dicho paquete se encuentra dentro del paquete auxiliar de VMO implementado por (Quintero, 2011) en su tesis de grado y es reutilizado en este trabajo.

II.2. Representación espacial

Casi desde los primeros días de la teoría atómica de Dalton se utilizaron representaciones espaciales de los objetos y los procesos químicos, se trató de representar la complejidad de los sistemas químicos mediante modelos. Sin embargo, el trabajo de los químicos computacionales de hoy sobrepasa a simple vista por el uso de herramientas para construir “geometrías” moleculares, programas para dibujar complejos armazones de esferas y varillas y sobre todo, de nuevos algoritmos para el cálculo cuántico de propiedades que requieren como datos de entrada “las posiciones de los átomos en el espacio” o las coordenadas de las moléculas. Todo esto es básicamente incorrecto como “visión” de los átomos, las moléculas o las reacciones químicas; pero se usa como una aproximación a los resultados cuánticos (Quintero, 2011).

Para quien aborde el estudio de la química computacional, debe quedar claro que el uso de todas estas herramientas geométricas, se trata de una estrategia que nos facilita hacer los cálculos en forma aproximada además de una representación muy útil, pero no constituye un marco ampliado del mundo químico molecular.

II.2.1 Coordenadas internas y formato de la Z-Matriz

En química computacional los sistemas de N partículas ligadas (como es el caso de una molécula) pueden especificarse completamente con menos de $3N$ coordenadas. La

“armazón” molecular se puede mover como un todo, o puede rotar sin que con ello cambien las distancias internas y los ángulos entre los átomos que lo conforman. Por eso podemos ahorrar la especificación de algunos valores y aun así dar toda la información para construir el armazón si fijamos el origen en el centro de masas de la molécula (con lo cual descartamos tres coordenadas) y además, si impedimos que el sistema de ejes de referencia rote libremente, fijándolo con lo cual podemos descartar otras tres coordenadas si la molécula es no lineal, o dos si es lineal. Al conjunto resultante se denomina matriz de coordenadas internas (Quintero, 2011).

La Z-Matriz es un formato muy común coordenadas internas moleculares en el cual la geometría molecular se especifica a partir de $3N - 5$ o $3N - 6$ valores de distancias ángulos “de enlace”. Una geometría molecular se puede especificar mediante vías de una Z-Matriz.

Para lograr tal estructura se utilizó una clase a la cual se le llamó ZMatriz modelada e implementada en la tesis (Quintero, 2011), para lograr una representación de los átomos en forma de distancia, ángulos y ángulos de torsión. Computacionalmente la construcción de la Z-Matriz depende en gran medida de la secuencia por la dependencia que genera esta representación.

Los elementos empleados para la generación de una Z-Matriz pueden o no reflejar la conectividad de la molécula. La gran mayoría de las ocasiones el uso de una Z-Matriz permite realizar optimizaciones geométricas con mayor velocidad de computo respecto del uso de coordenadas cartesianas.

Para la implementación de la Z-Matriz en el modelo se utiliza una interfaz denominada ZAMin modelada e implementada en la tesis (Quintero, 2011), con el objetivo de agrupar los valores de distancias, ángulos y ángulos diedros entre los átomos de cada aminoácido, dejando como variable los ángulos diedros que conecta un aminoácido con otro. También se utiliza otra interfaz para definir la particularidad del primer aminoácido de contener un Angulo de torsión, la misma se denominó Cab_Amin modelada e

implementada en la tesis (Quintero, 2011). Las clases que implementan dichas interfaces están contenidas en dos paquetes denominado Cabeceras y Residuos.

II.2.2 Método de conversión de coordenadas internas a cartesianas.

Los algoritmos de modelación de proteínas que utiliza la técnica *ab initio* realizan la búsqueda de muchas conformaciones atómicas sobre una superficie potencial. Algunos métodos como Rosetta y Dyana, generan perturbaciones o prueban moverse en el espacio angular y entonces convierten las coordenadas angulares en cartesianas las cuales emplean para evaluar las funciones potenciales o campos de fuerza.

Esta conversión, puede denominar el tiempo necesario para evaluar una prueba de movimiento. Por lo tanto, una conversión eficiente de los ángulos de torsión a una representación cartesiana es de vital, interés tema profundizado en la tesis (Quintero, 2011).

II.3. Función de Amplitud

(Ruiz-Blanco, 2014) Plantea la formulación general del **modelo de estabilidad**, el cual emplearemos como **función objetivo**, el modelo se define como sigue:

$$\Delta G_f = \omega_H (\Delta G_{el} + \Delta G_{VdW} + \Delta G_{tor} + \Delta G_{HBd}) + \omega_S (\Delta G_w + \Delta G_{conf})$$

Ecuación II-1: Función Objetivo

Los valores de los coeficientes son: $\omega_H = 0.01024$ y $\omega_S = -0.08016$.

El **cambio en la energía libre conformacional** viene dado por:

$$\Delta G_{conf} = 2.08 RT N_{eff}$$

Ecuación II-2: Energía libre conformacional

Donde R y T son la constante de los gases (8.314×10^{-3} kJ/mol) y la temperatura (298 K) respectivamente y N_{eff} representa la *longitud efectiva* de la cadena correspondiente a la suma de contribuciones de cada residuo.

$$N_{eff} = \sum_{i=1}^N N_{eff}^i$$

$$N_{eff}^i = \frac{N-i}{N-1} * \frac{\sum_{j=i+2}^N \delta_{ij} |j-i|}{\sum_{j=i+2}^N |j-i|} + \frac{i-1}{N-1} * \frac{\sum_{j=1}^{i-2} \delta_{ij} |j-i|}{\sum_{j=1}^{i-2} |j-i|}$$

$$\delta_{ij} = \begin{cases} 1 & \text{si } \frac{|j-i|}{d_{ij}} \geq \frac{4}{6.4} \\ 0 & \text{otro caso} \end{cases}$$

Ecuación II-3: Longitud Efectiva

Donde d_{ij} representa la distancia espacial entre los residuos i y j .

El modelo del **efecto hidrofóbico** responde a la siguiente ecuación:

$$\Delta G_w = k_{np}^w \Delta A_{np} - RT k_p^w \sum_{i=1}^N \delta_i^S \gamma_i$$

Ecuación II-4: Efecto Hidrofóbico

Los valores de los coeficientes son: $k_{np}^w = 0.0452$ and $k_p^w = 0.2819$. El parámetro δ_i^S es una variable unitaria que toma valor 1 si el residuo es superficial o 0 en caso contrario.

El segundo término simula el incremento en los grados de libertad (γ_i) de las moléculas de agua al aumentar la densidad de residuos hidrofílicos en la superficie de una proteína.

$$\gamma_i = \ln \left[\frac{\left(\frac{g_i^F}{N_i^w} \right)}{\left(\frac{g_i^U}{N_i^w} \right)} \right] \quad g_i = N_i^W + \sum_{j=1}^N \delta_{ij}^{nb} \delta_j^S N_j^w$$

Ecuación II-5: Grados de Libertad

Donde δ_{ij}^{nb} es una variable binaria que toma valor 1 si los residuos i y j son vecinos o 0 en otro caso. N_i^w representa el número de moléculas de agua asociadas a la cadena lateral de cada tipo de residuo. (Jiang et al., 2005)

Para describir las **interacciones electrostáticas** se representó a cada residuo como un punto situado en las coordenadas del C α correspondiente. Se diseñó un potencial tipo Coulomb, modificado por la acción de un *factor de ponderación radial* (C_{ij}).

$$\Delta G_{el} = -k_{el} \sum_{i=1}^{N-2} \sum_{j>i+1}^N C_{ij} \frac{q_i q_j}{r_{ij}} \quad C_{ij} = \frac{r_i r_j}{r^2}$$

Ecuación II-6: Interacciones electrostáticas

Donde q representa el ECI de cada residuo (Collantes and Dunn-III, 1995), r_{ij} la distancia entre los residuos i y j medida desde sus $C\alpha$, r_i representa la distancia entre el residuo i y el centro de la proteína y r^2 representa el cuadrado del radio medio de la estructura. El valor del coeficiente es: $k_{el} = 7.608$.

Las **interacciones de Van der Waals** emplean un potencial de Lennard-Jones considerando como centros de interacción las coordenadas de su $C\alpha$. La distancia mínima de interacción es fijada en 3.96.

$$\Delta G_{LJ} = k_{LJ} \sum_{i=1}^{N-2} \sum_{j>i+1}^N \left[\left(\frac{3.96}{r_{ij}} \right)^{12} - \left(\frac{3.96}{r_{ij}} \right)^6 \right]$$

Ecuación II-7: Interacciones de Van der Waals

Donde r_{ij} representa la distancia entre los $C\alpha$ de cada par de residuos, N la longitud de la proteína y el coeficiente es: $k_{LJ} = 63.98$.

El modelo de **penalización por déficit de puentes de hidrógeno** se formula en la siguiente forma:

$$\begin{aligned}\Delta G_{Hbd} &= \Delta G_{Hbd}^N + \Delta G_{Hbd}^O \\ \Delta G_{Hbd}^N &= \sum_{i=1}^N c_N \delta_i^N \\ \Delta G_{Hbd}^O &= \sum_{i=1}^N c_O \delta_i^O\end{aligned}$$

Ecuación II-8: Penalización por Déficit de Puentes de Hidrógeno

Donde δ_i^N toma valor 1 si el átomo N del residuo i es interno y no esta enlazado por H-Bond con un átomo O de otro residuo, o 0 en caso contrario. En forma similar δ_i^O toma valor 1 si el átomo O del residuo i es interno y no está enlazado por H-Bond con un átomo N de otro residuo, o 0 en caso contrario. Las penalizaciones correspondientes a las valencias no ocupadas fueron fijadas en: $c_N = 2.931$ kJ/mol y $c_O = 11.724$ kJ/mol, de acuerdo al análisis de las energías de unión de dímeros amida - amida, amida - agua y agua - agua calculadas por (Dixon et al., 1994)

El **potencial de torsión de la cadena central** se formula de la siguiente forma:

$$\Delta G_{tor} = k_{tor} \sum_{i=1}^N [(\cos^2 2\phi_i - 1) + 0.256(\cos^2 2\psi_i - 1)]$$

Ecuación II-9: Potencial de Torsión de la Cadena Central

El coeficiente es $k_{tor} = 1.219$. Donde ψ_i y ϕ_i representa los ángulos diedros de la cadena central asociados al residuo i .

II.3.1 Restricciones del problema

Existen dos restricciones fundamentales a lo largo de la optimización, las cuales son aplicadas una en cada sistema de coordenadas. La primera de ellas, realizada a nivel de la Matriz-Z, realiza un conteo de la cantidad de ángulos ψ y ϕ cuyos valores están ubicados en rangos poco favorables (irregulares). Las soluciones cuyo conteo supere el máximo permitido por la restricción son descartadas. Esta restricción se evalúa antes de seleccionar la mejor solución y al generar las soluciones iniciales de la optimización. La segunda restricción se aplica a nivel de la matriz de coordenadas cartesianas y busca

identificar estructuras con átomos superpuestos, para lo cual fija el valor mínimo entre un par de átomos. Esta restricción es evaluada para hallar los nodos sucesores que se utilizan en la metaheurística.

II.3.2 Espacio de búsqueda del problema

La metaheurística persigue optimizar la columna asociada a los ángulos diedros de la Matriz-Z de la proteína. Esta matriz es una representación del espacio de coordenadas internas de la estructura. En sus tres columnas de coordenadas reúne los valores de distancias de enlace, ángulos de enlace y ángulos diedros necesarios para ubicar a un determinado átomo en un punto del espacio cartesiano común conociendo las posiciones de los átomos que le preceden.

La optimización asume que la columna de las distancias y ángulos de enlace son fijos y por tanto no constituyen variables para el proceso de exploración. De los ángulos diedros que conforman la tercera columna de coordenadas internas se hace distinción de tres clases:

- i) los ángulos omega (ω) de la cadena central son igualmente considerados fijos con valor igual a 180° , este ángulo está asociado al enlace peptídico que es un elemento estructural plano y considerablemente rígido a por lo que su variación es mínima entre los distintos residuos de la cadena. Esto constituye una aproximación usual en los métodos de predicción estructural de proteínas.
- ii) Los demás ángulos diedros de la cadena central (ψ_i y ϕ_i) cuyos valores determinan la forma tridimensional de la estructura y constituyen por tanto las variables principales de la metaheurística.
- iii) El resto de los ángulos diedros de la estructura, correspondientes a las cadenas laterales de cada residuo. Este grupo no determina la forma global de la estructura sino que sus valores dependen considerablemente de la forma adoptada por la proteína.

No obstante al determinar la posición de los átomos de las cadenas laterales, estos valores afectan la función objetivo. En la metaheurística estos valores se exploran

aleatoriamente cada vez que se encuentra una mejor solución es encontrada a lo largo de la optimización, manteniendo aquella configuración de valores que minimice la función objetivo de la mejor solución encontrada.

En correspondencia a lo anterior la implementación de la optimización debe continuamente convertir la Matriz-Z de la proteína, con los valores de los diedros de una determinada solución, en una matriz de coordenadas cartesianas; que es empleada para evaluar la función objetivo.

II.4. Recocido Simulado aplicado al problema del plegamiento de proteínas

En correspondencia a lo anterior la implementación de la optimización debe continuamente convertir la Matriz-Z de la proteína, con los valores de los diedros de una determinada solución, en una matriz de coordenadas cartesianas; que es empleada para evaluar la función objetivo.

II.4.1 Diagrama de Actividad de la metaheurística RS

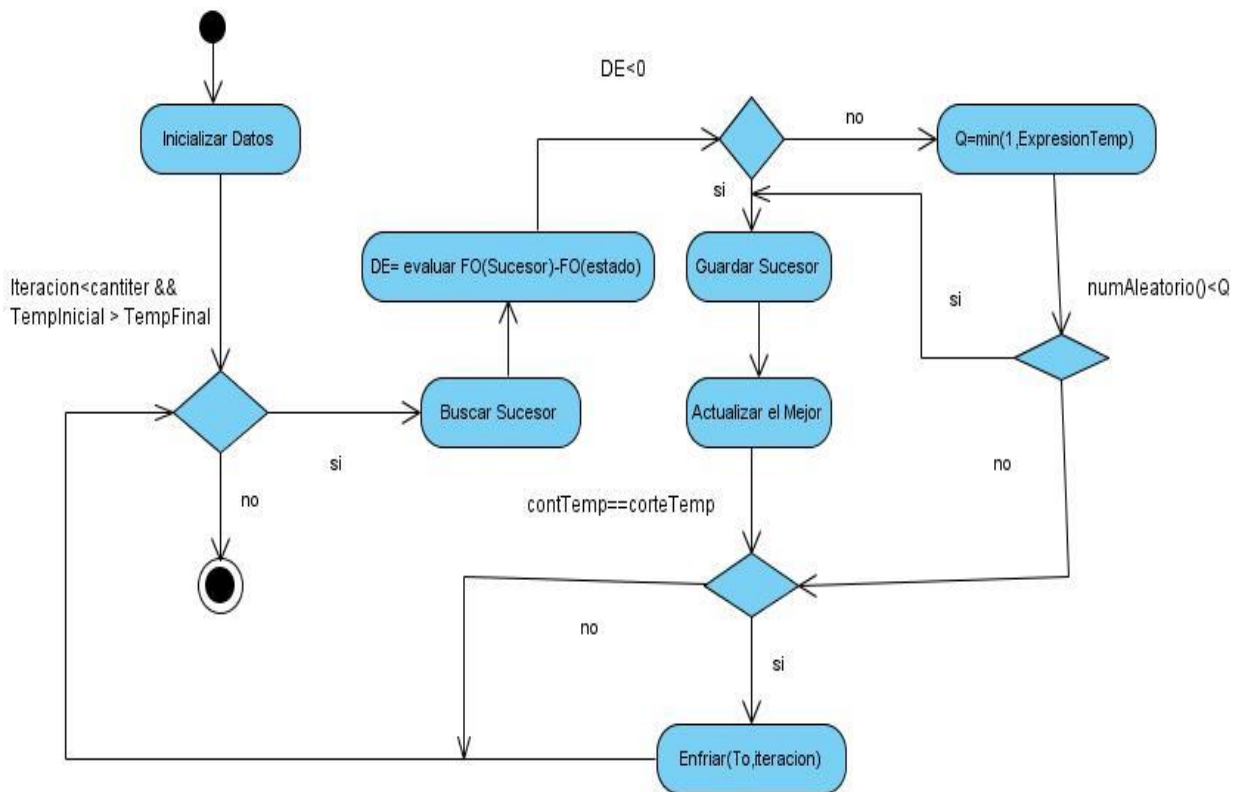


Figura II-1: Diagrama de Actividad de la metaheurística RS

En el diagrama de actividad se explica cómo se realizó la implementación de la metaheurística RS en esta tesis de grado.

Después de leer los datos de entrada se analiza dentro de un “for” (sentencia que representa un ciclo de java) si se cumple la condición de que la iteración en la que se encuentra el programa en ese momento es menor que el total de iteraciones (el valor se decide por el usuario en el fichero de entrada) y también verifica que la temperatura inicial sea mayor que la final; si se cumple se busca un sucesor de la solución inicial para esto se llama a la función **Mut2RegularZones()** a la cual se le pasan como parámetros la solución inicial, la cantidad de ángulos diedros que se van a mutar (este valor se decide por el usuario en el fichero de entrada) y la iteración actual en la que se encuentra el programa.

Luego se encuentra la función objetivo que se calcula función de calidad del sucesor menos la función de calidad de la solución inicial y se analiza en un “if” (sentencia que representa un ciclo de java) si se cumple la condición de si la función objetivo es menor que 0; si se cumple se guarda el sucesor y se actualiza el mejor llamando a la función **BestNodeSA ()** que se le pasa por parámetro la mejor solución encontrada.

Inmediatamente se comprueba la condición para poder enfriar la temperatura; si no se cumple se haya el mínimo entre 1 y una expresión matemática que depende de la temperatura para verificar dentro de otro “if” si el mínimo encontrado es mayor que un número generado aleatoriamente entre 0 y 1 y si la condición se cumple se acepta el sucesor aunque este sea peor que la solución inicial; si esta condición no se cumple se pasa directamente a examinar si se cumple la condición de que el contador de temperatura sea igual al corte de temperatura (este valor es decidido por el usuario en el fichero de entrada).

Esta condición decide cada cuantas iteraciones se va a realizar el enfriamiento de la temperatura, a continuación se hace el enfriamiento de la temperatura y dependiendo del método de enfriamiento que se escoja que se indagará más adelante depende la velocidad de enfriamiento de la temperatura. Posteriormente se analiza el ciclo for del principio y si no cumple los requisitos del ciclo se termina la ejecución de la metaheurística.

II.4.2 Diagrama de Actividad del RS-Paralelo

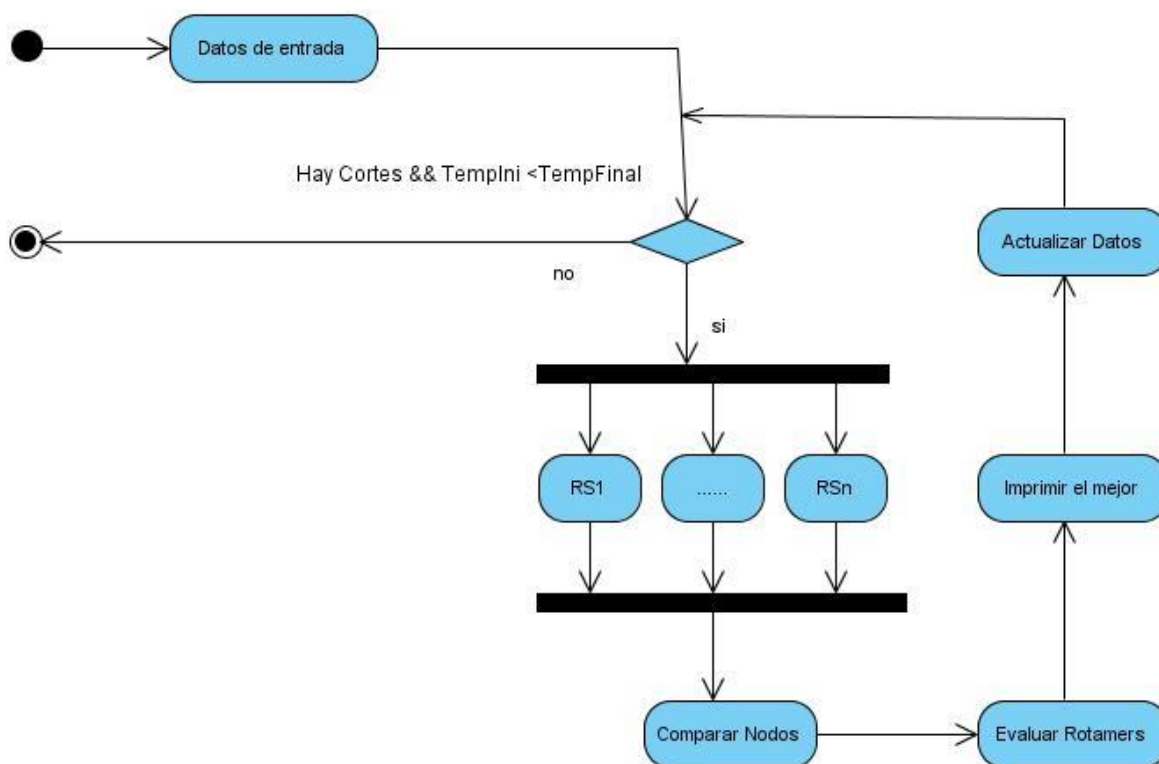


Figura II-2: Diagrama de Actividad del RS-Paralelo

En este diagrama de actividad se explica el funcionamiento en paralelo de la metaheurística RS.

Después de leer los datos de entrada se crea la población inicial de nodos que contienen las cadenas aminoácidos con los que se van a formar las proteínas, para crear esta población se llama al método **inicializate_Mesh_Helice_Beta ()** al ejecutar el programa, el tamaño de la población dependerá de la cantidad de hilos con los que se vaya a ejecutar RS (este valor es decidido por el usuario en el fichero de entrada), la mitad de la población se crea con hélices y la otra mitad con hojas betas.

A continuación se analiza si se cumple la condición de que si hay cortes en las iteraciones y si la temperatura inicial es menor que la temperatura final; si se cumple se ejecutan de manera paralela los hilos que contienen el RS y por cada ejecución de la metaheurística se asignan un individuo de la población y la temperatura inicial del RS.

Luego de terminado la ejecución de los hilos se realiza una comparación para sacar el mejor nodo encontrado con la función **find_MinGlobalBestNode_SA ()** que se le pasa un arreglo (estructura de dato en java) con el mejor nodo de cada hilo y al mejor nodo de todos se le evalúan los rotámetros con la función **Rotamer_SA ()**; cuando se termina de evaluar los rotámetros se imprime el mejor nodo, se actualizan los datos y se vuelve a examinar la condición del inicio y si no se cumple se termina la ejecución del programa.

II.4.3 Diagrama de Actividad de cómo encontrar el Sucesor

A continuación se mostrara y explicara el diagrama de actividad del método que encuentra los sucesores.

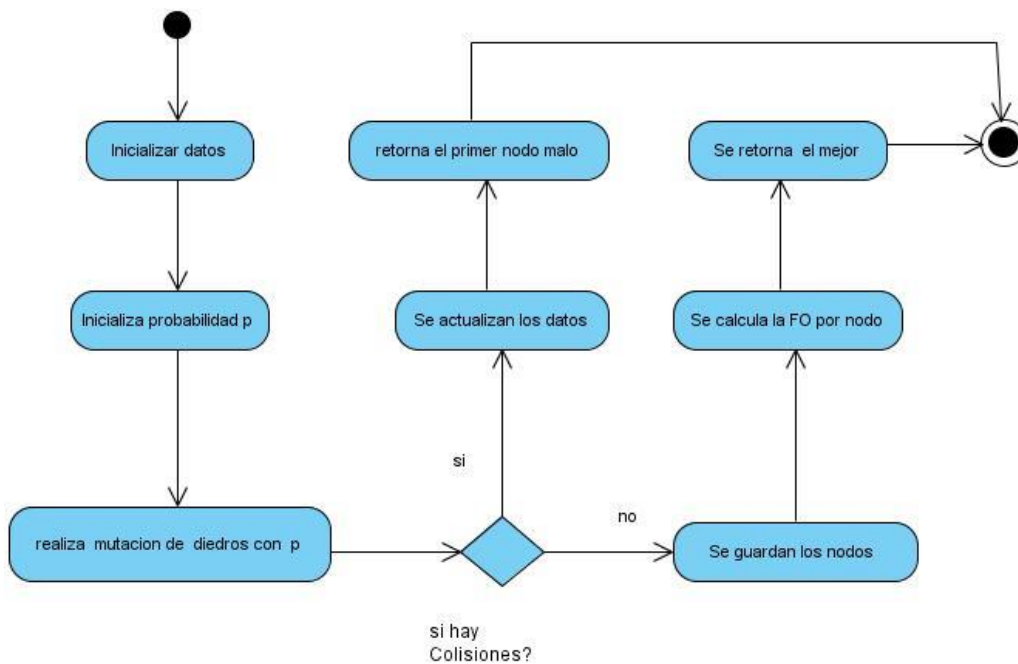


Figura II-3: Diagrama de Actividad de cómo encontrar el Sucesor

El método que encuentra los nodos sucesores se llama **Mut2RegularZones ()** al que se le pasan como parámetros el nodo actual, el número de ángulos diedros a mutar y la iteración en la que está este método crea un número de nodos igual a la cantidad de hilos que están corriendo con la metaheurística RS. Dentro del método se crea un

arreglo que va a guardar los aminoácidos a los que se le va a cambiar su posición en el espacio con el método **Auxiliar.randomArray_AA_Generate ()**, la cifra de aminoácidos que se cambia es igual al número de diedros que se pasó por parámetro.

Luego se inicializa una variable **p** que va a tener un número que se va a generar aleatoriamente del cual va a depender la mutación que se realiza después en un ciclo, en el que se va a llamar el método **Auxiliar.randomValue_Generate ()**, se utiliza el arreglo con los aminoácidos y la probabilidad, primero según la probabilidad que se genere se escoge un aminoácido del arreglo y con el método antes mencionado se cambia su posición o ángulo de rotación en cualquiera de los cuatro cuadrantes del sistema cartesiano.

Después se evalúa la condición de si el nuevo nodo que se genero tiene colisiones; si este no tiene colisiones se añade el nodo a una lista de nodos en el que se calcula la función objetivo por nodo y se retorna el mejor de la lista con el método **Auxiliar.find_MinGlobalBestNode ()**; si el nodo tiene colisiones se actualizan la configuración de datos de entrada, se guardan los nodos con errores y se retorna el primer nodo malo que encuentre.

II.4.4 Métodos de enfriamientos

Los métodos de enfriamientos es uno de los aspectos de mayor importancia en el RS además existe una amplia investigación realizada alrededor de este tema, porque los mismos, son los encargados de regular el decrecimiento de la temperatura cuyo valor determina en qué medida pueden ser aceptadas soluciones vecinas, mientras más lento sea el descenso de la temperatura hay mayor posibilidad de que converge al optimo global por eso es necesario realizar un estudio de los diferentes métodos de enfriamiento que existen antes de poder usar uno. A continuación se exhiben los métodos utilizados en la implementación. Aclaraciones T_i : Temperatura en la iteración actual y T_o : Temperatura inicial.

Descenso Geométrico: $T_{i+1} = \alpha * T_i$; $\alpha \in [0.8, 0.99]$

Criterio de Boltzmann: $T_i = T_o / (\log (i))$

Esquema de Cauchy: $T_i = T_0 / (1+i)$

Lundi-Mess: $T_{i+1} = T_i / (1+\beta \cdot T_i)$ con β muy pequeña

Para el estudio que se realizó de los métodos de enfriamientos se recopiló información de diferentes artículos, publicaciones y presentaciones extraídos de Internet a las que se le realiza mención ahora (Díaz, 1996),(K.A. Dowsland 2003),(B Suman OCT 2006),(Aponte, 2006) por si se desea seguir investigando este tema.

II.5. Implementación

Para la implementación de la metaheurística se decidió realizarlo en el NetBeans IDE 8.0, el NetBeans IDE es un entorno de desarrollo integrado libre principalmente para el lenguaje de programación Java, es un producto libre, gratuito y sin restricciones de uso.

También nos permite utilizar el SDK (kit de desarrollo de software) para los elementos químicos que son necesarios para el problema de plegado de proteínas. Se utilizó Java porque es un lenguaje de programación de propósito general, concurrente, orientado a objetos y basado en clases que fue diseñado específicamente para tener tan pocas dependencias de implementación como fuera posible (James Gosling, 2005). Su intención es permitir que los desarrolladores de aplicaciones escriban el programa una vez y lo ejecuten en cualquier máquina virtual Java (JVM) sin importar la arquitectura de la computadora. Se utiliza como proyecto base el de (Quintero, 2011) porque ya tiene implementado sobre la misma plataforma de programación los paquetes con las estructuras y jerarquías de los átomos, aminoácidos y demás conformaciones químicas que son necesarias para encontrar la estructura terciaria de proteínas.

II.5.1 Atributos utilizados para el algoritmo

Los parámetros de entrada utilizados en la ejecución del RS son: el número de hilos para ejecutar la metaheurística en paralelo, el porcentaje de hélices alfas y hojas betas constituyen características de la proteína, las réplicas son las veces que se puede ejecutar el algoritmo diferentes nodos aleatorios, los rotameros representan la optimización de las cadenas laterales de la proteína, el número de diedros son los

ángulos de rotación de los aminoácidos que conforman la proteína los cuales se van a mutar, la temperatura inicial con que comienza el algoritmo ,método de enfriamiento se representa con un número para escoger el procedimiento que se va a utilizar Boltzmann-1; Cauchy-2; Geométrico-3; Lundy-Mess -4; los parámetros para cortes intermedios de la metaheurística: tope interno y delta de temperatura; el contador de temperatura que es la frecuencia con la que se realizara el enfriamiento en el algoritmo y para perpetrar la parada del algoritmo esta se detiene cuando se alcance el total de generaciones o la temperatura actual menos la anterior sea menor o igual a 0.001.

II.6. Conclusiones parciales del capítulo.

Se muestra el diseño y se presenta como se realizó la implementación de la metaheurística Recocido Simulado en la solución del problema de predecir la estructura terciaria de proteínas. En la realización del RS se utilizaron coordenadas internas y formato de la Z-Matriz, el método de conversión de coordenadas internas a cartesianas, la función de amplitud y las restricciones del problema. Para la implementación de la metaheurística se decidió realizarlo en el NetBeans IDE 8.0, el mismo es un entorno de desarrollo integrado libre principalmente para el lenguaje de programación Java y además de ser un software libre también es gratuito y sin restricciones de uso. También se utilizó el SDK (Kit de Desarrollo de Software) para los elementos químicos que son necesarios para el problema de plegado de proteínas. Se usó como proyecto base el de (Quintero, 2011) porque ya tiene implementado sobre la misma plataforma de programación los paquetes con las estructuras y jerarquías de los átomos, aminoácidos y demás conformaciones químicas.

III. Análisis de los Resultados

En este capítulo se explica cómo es el proceso de selección de las mejores predicciones por métodos de enfriamientos para después realizar las comparaciones contra VMO y las proteínas internacionales calculadas en la competencia del CASP-ROLL. También se realiza una descripción de las métricas utilizadas para comparar las predicciones de proteínas contra la nativa en el servidor LGA y se mostraran los resultados generales que se obtuvieron de la comparación por métodos de enfriamiento, contra VMO y las predicciones internacionales.

III.1. Procedimiento para comparar la calidad de las soluciones

Para estudiar la calidad de las soluciones obtenidas con el SA implementado, se realizaron los siguientes pasos:

1. Seleccionar cuatro proteínas del CASP-ROLL.
2. A partir de la secuencia de aminoácidos de cada proteína y las cuatro configuraciones del SA se obtienen 30 réplicas de cada proteína por configuración.
3. Seleccionar 5 de las réplicas por proteína y compararlas con la estructura nativa correspondiente y con las soluciones de otros competidores del CASP-ROLL mediante las métricas del alineamiento estructural generado por el servidor LGA.

El primer paso consistió en seleccionar cuatro proteínas de la competencia CASP-ROLL. Para ello se seleccionaron las que tuvieran una longitud entre 100 y 300 aminoácidos (longitud de la proteína) para limitar el costo computacional de los experimentos, con estos requisitos los cumplían 21 proteínas. Se dividieron en distintos rangos (100-150; 150-200; 200-250; 250-300) y se seleccionó una proteína en cada rango las cuales se muestran en la Tabla III-1.

Tabla III-1: Nombre de las proteínas y sus respectivas longitudes

Nombre Proteína	R0007	R0015	R0023	R0033
Longitud	161	261	100	211

Para el segundo paso se realizaron cuatro configuraciones de parámetros al algoritmo, las configuraciones comunes son las presentadas en la Tabla III-2.

Tabla III-2: Atributos generales del SA y sus valores

Atributos	Valor
Nº de Hilos	5
Por ciento de Hojas Alfas	60
Por ciento de Hojas Betas	75
Total de generaciones	1000
Replicas	30
Rotameros	10
Numero de diedros	3
Temperatura Inicial	200
Tope Interno	10
Delta de Temperatura	5
Contador de Temperatura	1

Las cuatro configuraciones se diferenciaron en el parámetro tipo de enfriamiento que se muestra en la Tabla III-3 . Se ejecutó el algoritmo y se obtuvieron 30 réplicas de cada proteína que en el paso tres se analizan.

Tabla III-3: Atributos específicos por configuración.

Configuración	1	2	3	4
Enfriamiento	Boltzmann	Cauchy	Geométrico	Lundi-Mess

El tercer paso se realizó en 3 fases: Primero se seleccionaron las réplicas donde el valor de la función objetivo (Tanford) de la solución final estuviera en el siguiente rango $[FO_{min}, (FO_{min} + 0.7 * \Delta FO)]$, donde FO_{min} es el mínimo de los valores de la FO entre las soluciones finales de las 30 réplicas, $\Delta FO = FO_{min} - FO_{max}$ y FO_{max} es el máximo de los valores de la FO entre las soluciones finales de las 30 réplicas. Esta selección se realiza para eliminar las soluciones que convergen a mínimos locales intermedios cuyos valores de FO quedarían en el intervalo más próximo a FO_{max} . En la segunda fase, a las réplicas seleccionadas se les aplicó en el STATISTICA 7.0 (2004) un Análisis de Clúster de las K-medias con los parámetros siguientes: Variables: Función Objetivo y

Radio, Clusterizar por casos, Número de clúster a obtener: 5 y Cantidad de Iteraciones: 10. El número de clústeres se fijó en 5 debido a que en la competencia CASP-ROLL se compite con 5 soluciones por predictor y al final de este capítulo se realizará una comparación contra los resultados de sus competidores. La tercera fase consistió en seleccionar las soluciones que poseían menor distancia hacia el centro de cada clúster como las más representativas, obteniendo así los 5 modelos de cada proteína y por cada configuración de atributos (ver Anexos I, 0, 0, 0).

III.2. Alineamiento Local – Global (LGA) de estructuras de proteínas

LGA (Zemla, 2003) es un servidor web (<http://PredictionCenter.llnl.gov/local/lga>) diseñado para comparar estructuras de proteínas buscando la mejor superposición estructural de dos proteínas o de fragmentos estructurales. El alineamiento estructural se sostiene sobre el principio de que puede definirse una función objetivo con su valor óptimo correspondiente a la superposición estructural más significativa. LGA emplea en su función de ajuste dos parámetros, LCS (*Longest continuous segment*) y GDT (*Global Distance Total*), introducidos por Zemla, (Zemla, 2003) los cuales han mostrado gran éxito en la detección de similitudes estructurales locales y globales entre dos proteínas. (Zemla et al., 2001, Cristobal et al., 2001) El procedimiento LCS localiza el conjunto continuo más largo de residuos que puede ajustarse bajo un umbral de RMSD predeterminado. Por otro lado el algoritmo de GDT está diseñado para complementar las evaluaciones realizadas con LCS, en este caso se busca el mayor conjunto de residuos (no es necesario un conjunto continuo) que no se desvíen por más de una distancia de corte máxima seleccionada. (Zemla, 2003, Zemla et al., 2007)

III.2.1 Métricas calculadas por el servidor LGA

El alineamiento estructural, entre cada modelo creado y la estructura nativa de cada proteína, es generado por el método LGA. Este alineamiento es evaluado mediante cinco métricas, provistas por la propia aplicación, las cuales se describen a continuación:

N: Número máximo de pares de residuos *correspondientes* alineados satisfactoriamente de acuerdo a la distancia de corte seleccionada.

RMSD: Raíz de la media cuadrática de las desviaciones entre los N pares de residuos *correspondientes* satisfactoriamente alineados. Donde v_i^2 y v_i^1 representa las coordenadas de la estructura predicha y la estructura nativa respectivamente.

$$RMSD = \sqrt{\sum_{i=1}^N (v_i^2 - v_i^1)^2 / N}$$

Ecuación III-1: Métrica RMSD

LGA_S: Función de puntuación del ajuste, definida como:

$$LGA_S = w * GDT_S + (1 - w) * LCS_S$$

Ecuación III-2: Métrica LGA_S

Donde w varía entre 0.0 y 1.0, LCS_S es el porcentaje de residuos *correspondientes* que se ajusta por debajo de un valor de corte de RMSD denominado v_{LCS} , GDT_S se asocia al porcentaje de residuos *correspondientes* que se ajustan por debajo de un valor de corte de distancia v_{GDT} . Estos últimos parámetros, son estimados como F_S:

$$\begin{aligned} & \text{foreach } v_i (v_1, v_2, \dots, v_k) \{ \\ & \quad Y = (k - i + 1) / k; \\ & \quad X = X + Y(F_S_{v_i}); \\ & \} \\ & F_S = X / ((1 + k)k / 2); \end{aligned}$$

Donde para LCS, $v_i = 1, 2, \dots, v_{LCS}$ y para GDT, $v_i = 1, 2, \dots, v_{GDT}$. Los valores v_{LCS} y v_{GDT} corresponden a las distancias (en Å) empleados en el alineamiento. En los cálculos realizados en el presente trabajo $v_{LCS} = 5$ Å, $v_{GDT} = 10$ Å y $w = 0.5$.

GDT_TS = Distancia total global, se define como $(P_1 + P_2 + P_4 + P_8)/4$, donde P_d son los porcentajes de residuos que pueden ajustarse bajo los valores de distancias de corte $d = 1, 2, 4$ y 8 Å.

LGA_Q: Función de calidad del ajuste, definida en función de **N** y **RMSD** como:

$$Q = 0.1 * N / 0.1 + RMSD$$

Ecuación III-3: Métrica LGA_Q

III.2.2 Parámetros asociados al servidor LGA

En la configuración del método de alineamiento estructural entre los modelos tridimensionales generados y la estructura nativa de cada proteína, se empleó el comando “-4 -sda -d:10 -ch2:A” que instruye al programa para realizar el alineamiento buscando dentro del conjunto de residuos *correspondientes*, el mayor subconjunto de donde se logre un alineamiento óptimo de acuerdo al parámetro de distancia fijado (opción -4).

La *correspondencia* entre los pares de residuos del modelo y la estructura nativa se estableció considerando solamente los pares de residuos que coincidan en su numeración y tipo (opción -sda).

La distancia empleada como valor de corte del alineamiento y también para el cálculo de la componente GDT de las métricas fue fijada en 10 Å (opción -d:10).

Se empleó además la opción -ch2:A que fija el uso de la cadena identificada como ‘A’ del fichero correspondiente a la estructura nativa para asignar la correspondencia con los residuos de los modelos obtenidos, aun cuando la cadena correspondiente de los ficheros de los modelos no posea un identificador o este sea distinto de ‘A’. Esta es una opción necesaria cuando se usan datos de entrada de diferentes fuentes y se requiere que la *correspondencia* inter-residuos sea dependiente de la secuencia (-sda).

III.3. Comparación de los resultados

De las 5 predicciones de cada configuración se analizaron las métricas de la mejor solución por configuración. Los resultados de las métricas obtenidas se tabularon con la ecuación

$$Rank (P, T) = \frac{T - P}{T} * 100$$

Ecuación III-4: Ranqueo de Posiciones

Donde P es la posición obtenida al renquear por la métrica y T es el total de predicciones analizadas. El objetivo de dicha ecuación es expresar en porciento: cuantas soluciones tenemos por debajo de la nuestra.

III.3.1 Comparación entre los cuatro métodos de enfriamiento

Se realizó la comparación entre los métodos de enfriamiento utilizados por el SA en el servidor de LGA donde se realizan los cálculos a las proteínas usando las configuraciones **4 -sda -ch2: A -d: 10**, los resultados que se encuentran en la tabla de comparación de los cuatro métodos de enfriamiento para la proteínas se localizan en el Anexo V. En la Tabla III-4 se presentan la generalización de los métodos de enfriamiento.

Tabla III-4: Generalización de las métricas por método de enfriamiento

	N	RMSD	GDT	LGAS	LGAQ	Promedio (%)
Boltzmann	81.25	75	78.75	78.75	81.25	79
Cauchy	82.5	80	71.25	82.5	82.5	79.75
Geométrico	81.25	86.25	81.25	78.75	82.5	82
Lundi-Mess	90	86.25	92.5	92.5	91.25	90.5

En la comparación general de los métodos de enfriamiento, podemos afirmar que el método de enfriamiento que arrojó los mejores resultados fue el de Lundi-Mess . Este método generó como promedio soluciones por encima del 90,5% de las soluciones de los demás métodos. Esta superioridad se ve presente en todas las métricas evaluadas.

Las métricas que obtuvieron los mejor resultado fueron **GDT** y **LGAS** con un 92.5% como se puede ver en la

Figura III-1.

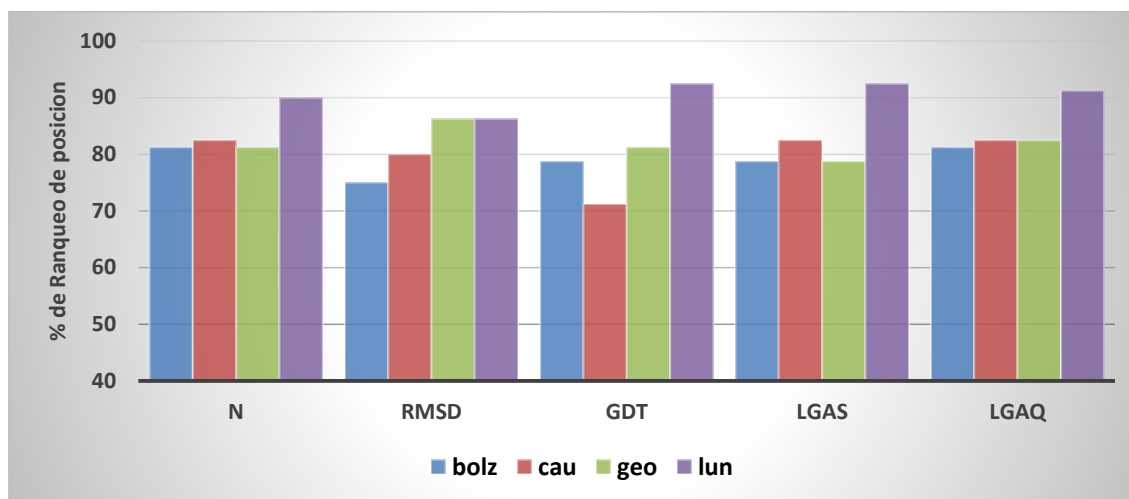


Figura III-1: Comparación entre los cuatro métodos de enfriamiento

III.3.2 Comparación entre el método SA-enfriamiento y VMO

Se realizó la comparación entre los algoritmos SA (del que se escogió las soluciones del método de enfriamiento Lundi-Mess) y VMO en el servidor de LGA donde se realizan los cálculos a las proteínas usando la configuración **-4 -sda -ch2:A -d:10**, los resultados que se encuentran en la tablas de comparación para las proteínas se localizan en el Anexo VI . La comparación general (ver Tabla III-5) de los resultados entre las metaheurísticas SA y VMO, de ella podemos afirmar que el algoritmo que arrojó el mejor resultado fue el de SA con un 92,25%.

Tabla III-5: Generalización de las métricas por metaheurísticas

	N	RMSD	GDT	LGAS	LGAQ	Promedio (%)
RS	90	93.75	92.5	92.5	91.25	92
VMO	93.75	90	91.25	90	91.25	91.25

En la Figura III-2 podemos apreciar el comportamiento de las dos metaheurística en cada métrica. El algoritmo SA supera en tres métricas al VMO y además se obtuvo los mismos resultados en la métrica **LGAQ**.

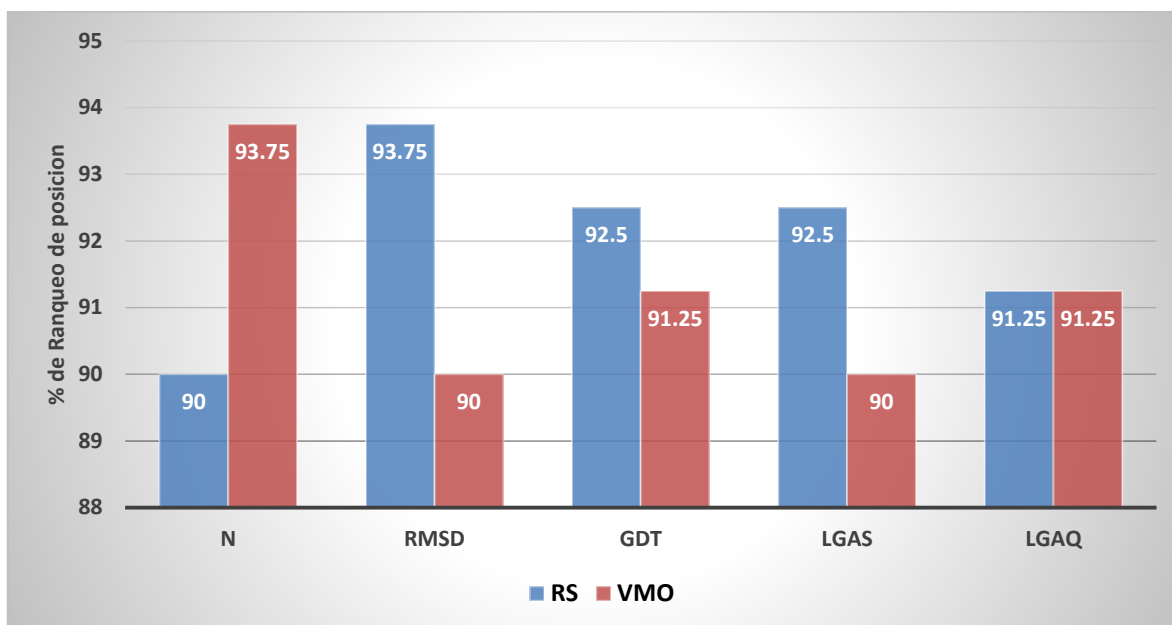


Figura III-2: Comparación entre SA y VMO

III.3.3 Comparación entre el método SA-LundiMess y VMO contra predicciones internacionales

Se realizó la comparación entre los algoritmos SA (del que se escogió las soluciones del método de enfriamiento Lundi-Mess) y VMO contra las predicciones internacionales del CASP-ROLL. En la Tabla III-6 se muestra la cantidad de soluciones por competidor; a estas soluciones se le calcularon las métricas en el servidor de LGA donde se realizan los cálculos a las proteínas usando la configuración **4 -sda -ch2: A -d: 10**.

Tabla III-6: Cantidad de soluciones por competidores

	R0007	R0015	R0023	R0033
RS	5	5	5	5
VMO	5	5	5	5
Internacionales	90	91	78	21
Total	100	101	88	31

Los resultados obtenidos se encuentran en el Anexo VI del cual podemos concluir que en la metaheurística SA la métrica **LGAQ** de la proteína R0033 obtuvo la quinta posición entre 31 soluciones, representado por 95.05% y siendo esta proteína la de

mejores resultados en todas las métricas. Además se afirma la superioridad del algoritmo SA sobre VMO en cuatro métricas.

En la Figura III-3 se presenta el promedio general por metaheurística en cada métrica. Se observa que el mayor promedio lo obtuvo la métrica **GDT** en el algoritmo SA con un valor de 69.56%. Atendiendo al significado de las métricas se puede concluir que la metaheurística implementada presenta mejores resultados en la métrica GDT (69.56%) contra LGAS (54.7%), lo que significa que tiene mejor alineamiento global que local. En todas las métricas las soluciones obtenidas por la metaheurística SA en promedio se encuentran en el segundo cuartil.

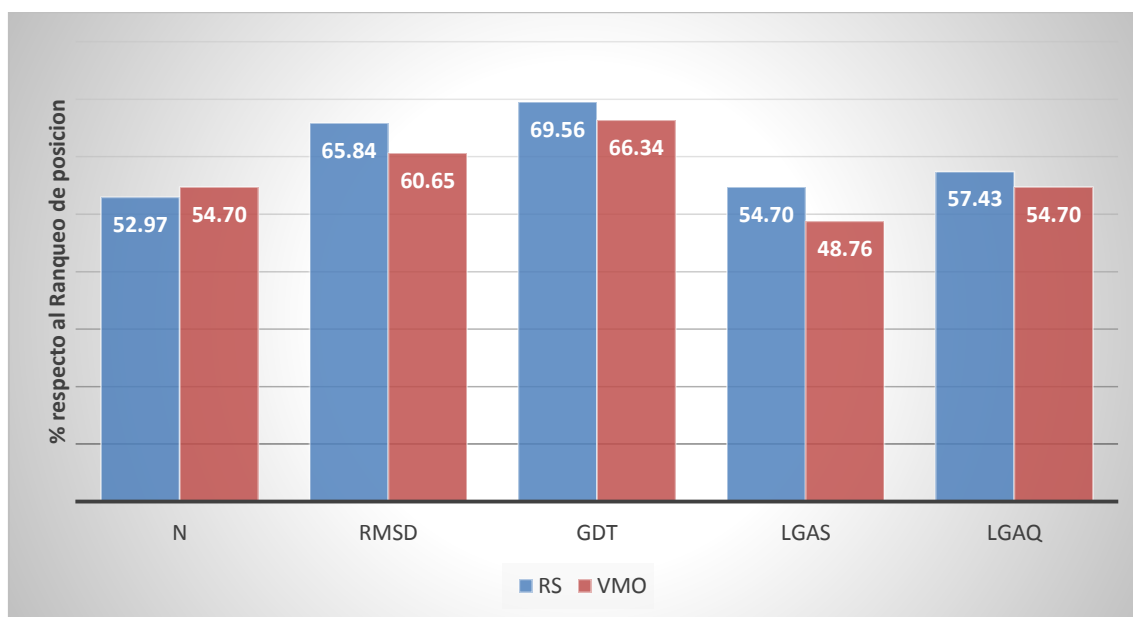


Figura III-3: Comparación entre SA, VMO y competidores internacionales

III.4. Conclusiones del capítulo

Para la selección de las soluciones se obtuvieron 30 réplicas de cada proteína por configuración seleccionando de ellas cinco con un análisis de clúster para realizar las comparaciones en el servidor LGA. Se compararon los resultados entre los cuatro métodos de enfriamiento, obteniendo que los resultados favorables en cuanto a calidad los tiene el método de Lundi-Mess. La comparación de los resultados contra las

soluciones de VMO, arrojó como conclusión la superación de la metaheurística RS a VMO en cuatro de las cinco métricas medidas. De la comparación que se realizó con soluciones de predictores internacionales podemos destacar que el promedio de nuestras soluciones se presentan en el segundo cuartil, además la metaheurística implementada proporciona un mejor alineamiento global (GDT) que local (LGAS).

Conclusiones

Conclusiones

En la investigación que se realizó del problema de predicción de la estructura terciaria de proteínas se presentaron los aspectos biológicos fundamentales al mostrar la descripción de la estructura de las proteínas, su jerarquía y la definición del problema de plegado de las mismas lo que posibilitó realizar el diseño de un algoritmo basado en Recocido Simulado ajustado a este problema. Esta investigación permitió una mejor comprensión de los factores biológicos que originan enfermedades que afectan a la población mundial.

Después del análisis perpetrado del algoritmo Recocido Simulado, se implementó el mismo con cuatro métodos de enfriamiento (Boltzmann, Cauchy, Geométrico, Lundi-Mess). En el estudio comparativo de las soluciones generadas por cada método para determinar la calidad de sus soluciones, se obtuvo que el método de enfriamiento Lundi-Mess fue el que mejor resultados ofreció.

En los estudios comparativos que se realizaron de las soluciones entre las metaheurísticas RS-Lundi-Mess y VMO, se concluye en el análisis de desempeño de las soluciones que el RS obtiene mejores resultados en tres de las cinco métricas obtenidas por el servidor LGA. Con respecto a la comparación de las soluciones de RS y predictores internacionales, el promedio por métricas de las soluciones generadas de RS se encuentran en el segundo cuartil; donde la métrica que alcanzó el mayor porcentaje fue la de alineamiento de la distancia global (GDT).

Después de realizadas las comparaciones entre RS- Lundi-Mess, VMO y las predicciones internacionales se ha llegado a la conclusión de que la metaheurística Recocido Simulado es viable para predecir la estructura terciaria de proteínas.

Recomendaciones

Recomendaciones

- ✓ Realizar un estudio exhaustivo sobre los parámetros del algoritmo de implementado.
- ✓ Utilizar un conjunto mayor de proteínas para el estudio de calidad.
- ✓ Realizar una cantidad superior a 100 réplicas de cada proteína para tener más soluciones que clusterizar y obtener una mejor representatividad de las soluciones.

Referencias Bibliográficas

2001. *Computational Biochemistry and Biophysics*, New York, Oren M. Becker, Alexander D. Mackerell, Jr., Bonoit Roux, Masakatsu Watanabe
2004. STATISTICA (data analysis software system). Tulsa: StatSoft, Inc. version 7. www.statsoft.com.
- ANFINSEN, C. B. 1973. Principles that govern the folding of protein chains. *Science*, 181, 223-230.
- APONTE, I. 2006. *RECOCIDO SIMULADO Y BÚSQUEDA TABÚ PARA EL PROBLEMA DE PREDICCIÓN DE LA ESTRUCTURA TERCIARIA DE PROTEÍNAS*. UNIVERSIDAD SIMÓN BOLÍVAR.
- B SUMAN , P. K. OCT 2006. *A survey of simulated annealing as a tool for single and multiobjective optimization*
- BERMAN, H. M. 2008. The Protein Data Bank: a historical perspective. *Acta Crystallographica Section A: Foundations of Crystallography*, A64, 88-95.
- COLLANTES, E. R. & DUNN-III, W. J. 1995. Amino acid side chain descriptors for quantitative structure-activity relationship studies of peptide analogues. *J. Med. Chem.*, 38, 2705-2713.
- CRISTOBAL, S., ZEMLA, A., FISCHER, D., RYCHLEWSKI, L. & ELOFSSON, A. 2001. A study of quality measures for protein threading models. *BMC Bioinformatics*, 2.
- CHOTHIA, C. 1992. *Nature*, 360, 543-544.
- D.H WOLPERT, W. G. M. 1997. No Free Lunch Theorems for Optimization. *IEEE Transactions of Evolutionary Computation*, 1, 67-82.
- DAVID WOLPERT, W. M. 1997. "No Free Lunch Theorems for Optimization". 1, 67.
- DÍAZ, A. 1996. *Algoritmos de Enfriamiento Simulado*
- DIXON, D. A., DOBBS, K. D. & VALENTINI, J. J. 1994. Amide- Water and Amide- Amide Hydrogen Bond Strengths. *The Journal of Physical Chemistry*, 98, 13435-13439.
- E AARTS, J. K. 1990. *Simulated Annealing and Boltzmann Machines.*, New York, Wiley.
- ESCUELA, G. 2006. *Algoritmos evolutivo con representación basada en sistemas-L para el problema del replegado de las proteínas*. Magíster en Ciencias de la Computación, UNIVERSIDAD SIMÓN BOLÍVAR.
- GALLARDO, M. C. 2014. *Heurística de Recocido Simulado para la resolución del problema del acarreo terrestre*. Escuela Técnica Superior de Ingeniería Universidad de Sevilla
- H. BERMAN, J. W., Z. FENG, G. GILLILAND, T. BHAT, H. WEISSIG, I. SHINDYALOV AND P. BOURNE. 2000. The protein data bank. . *Nucleic Acids Research*, 28, 235-242.

- JAMES GOSLING, B. J., GUY STEELE, Y GILAD BRACHA. 2005. The Java language specification.
- JAROSŁAW PILLARDY, C. C., ADAM LIWO, JOOYOUNG LEE, DANIEL R. RIPOLL, RAJMUND KAZ'MIERKIEWICZ,, STANISŁAW OŁDZIEJ, W. J. W., KENNETH D. GIBSON, YELENA A. ARNAUTOVA, JEFF SAUNDERS, YUAN-JIE YE, & SCHERAGA, A. H. A. 2001. Recent improvements in prediction of protein structure by global optimization of a potential energy function. *PNAS*, 98, 2329–2333.
- JIANG, L., KUHLMAN, B., KORTEMME, T. & BAKER, D. 2005. A “solvated rotamer” approach to modeling water-mediated hydrogen bonds at protein–protein interfaces. *PROTEINS: Structure, Function, and Bioinformatics*, 58, 893–904.
- K.A. DOWSLAND , A. D. 2003. *Diseño de heurísticas y fundamentos del recocido simulado*
- LEVINTHAL, C., J.T. 1969. *Mossbauer Spectroscopy in Biological Systems, in Proceedings of a meeting held at Allerton House,, Univ Illinois Press: Urbana. .*
- MIKAEL OLIVEBERG, Y.-J. T., AND AIAN R. FERSHT 1995. Negative activation enthalpies in the kinetics of protein folding. *Proc. Natl Acad. Sci.*, 92, 8926-8929.
- ML RILEY, B. W., SL FLITSCH, PJ BOOTH 1997. *Biochemistry,,* 36, 192-196.
- NICHOLAS HAMILTON, K. B., MARK A. RAGAN, AND THOMAS HUBER¹ 2004. Protein Contact Prediction Using Patterns of Correlation. *PROTEINS: Structure, Function, and Bioinformatics*, 56, 679–684.
- PARDO, E. G. 2012-2013. *Optimización de Sistemas de Comunicación Tema 3.3 Recocido Simulado.*
- PRUSINER, S. 1998. *Prion protein biology. Cell*, 93, 337-348.
- QUINTERO, I. M. 2011. *Optimizacion basada en Mallas Variables para Predecir el Plegamiento de Proteinas* Universidad "Martha Abreu" de Las Villas
- RICHARD BONNEAU, J. T., INGO RUCZINSKI, DYLAN CHIVIAN, CAROL ROHL, CHARLIE E. M. STRAUSS, AND & BAKER, D. 2001. Rosetta in CASP4: Progress in Ab Initio Protein Structure Prediction. *PROTEINS: Structure, Function, and Genetics* 5, 119–126.
- RUIZ-BLANCO, Y. B., MARRERO-PONCE, Y., GARCÍA, Y., PURIS, A., BELLO, R., GREEN, J., & SOTOMAYOR-TORRES. 2014. *A physics-based scoring function for protein structural decoys: Dynamic testing on targets of CASP-ROLL.* Universidad Central "Marta Abreu" de Las Villas.
- SC HYDE, P. E., MJ HARTSHORN, MM MINMACK, U GILEADI, SR PEARCE, MP GALAGHER, DR AND R.H GILL CF HIGGINS 1990. *Nature*, 346, 362-365.

Referencias Bibliográficas

- TANFORD, C. 1962. Contribution of Hydrophobic Interactions to the Stability of the Globular Conformation of Proteins. *J. AM. CHEM. SOC.*, 84.
- ZEMLA, A. 2003. LGA - A Method for Finding 3-D Similarities in Protein Structures. *Nucleic Acids Research*, 31, 3370-3374.
- ZEMLA, A., GEISBRECHT, B., SMITH, J., LAM, M., KIRKPATRICK, B., WAGNER, M., T. SLEZAK & ZHOU, C. E. 2007. STRALCP structure alignment-based clustering of proteins. *Nucleic Acids Research*, 35.
- ZEMLA, A., VENCLOVAS, C., MOULT, J. & FIDELIS, K. 2001. Processing and evaluation of predictions in CASP4. *PROTEINS: Structure, Function, and Genetics*, 45, 13-21.

Anexos

I. Datos Generales del proceso de selección de la proteína R0007

Config.	Clúster		Distancias al centro del Clúster			Solución Seleccionada	
	Nº	Cantidad	Mínima	Máxima	Promedio	FO	Radio
1	1	6	5.39	10.69	8.20	-736.40	20.12
	2	1	0.00	0.00	0.00	-783.82	25.47
	3	4	0.69	4.45	2.47	-694.53	23.83
	4	1	0.00	0.00	0.00	-683.27	32.23
	5	3	0.95	2.95	1.99	-677.70	25.38
	Total	15					
2	1	1	0.00	0.00	0.00	-797.206	19.99
	2	3	0.89	4.93	3.31	-730.78	25.24
	3	3	1.09	2.20	2	-706.039	23.29
	4	3	0.92	2.65	1.77	-717.177	23.62
	5	4	1.64	2.22	1.95	-698.881	21.18
	Total	17					
3	1	1	0.00	0.00	0.00	-676.978	34.63
	2	7	1.06	6.59	3	-684.029	22.64
	3	7	2.68	7.14	4.22	-699.218	28.6
	4	5	5.33	9.08	7	-735.809	30.54
	5	8	2.01	10.11	3.98	-641.231	26.43
	Total	28					
4	1	9	1.74	8.06	4.61	-715.11	22.55
	2	5	0.78	7.33	3.61	-738.09	19.81
	3	3	1.13	1.86	1.46	-686.59	21.87
	4	3	1.47	2.88	2.03	-683.40	26.87
	5	4	2.68	4.47	3.55	-667.50	27.45
	Total	24					

II. Datos Generales del proceso de selección de la proteína R0015

Config.	Clúster		Distancias al centro del Clúster			Solución Seleccionada	
	Nº	Cantidad	Mínima	Máxima	Promedio	FO	Radio
1	1	5	1.83	6.19	4	-1009.01	35.92
	2	2	5.48	5.48	5	-1029.994	40.58
	3	4	2.39	3.77	3.20	-988.234	46.76
	4	3	3.20	7.99	6	1050.475	28.49
	5	5	3.14	5.92	4.94	-970.853	29.21
	Total	19					
2	1	5	5.79	19.75	12	-1035.301	44.57
	2	8	4.04	9.03	5.45	-988.283	35.08
	3	2	1.97	1.97	1.97	-970.758	49.64
	4	2	3.24	3.24	3.24	-1114.245	41.28
	5	3	2.64	3.09	2.81	-951.238	33.2
	Total	20					
3	1	6	3.45	7.36	5.12	-991.356	33.17
	2	2	4.03	4.03	4.03	-965.903	41.82
	3	6	2.62	7.92	4.38	-1014.965	31.14
	4	7	3.52	25.78	10.07	-1044.269	29.84
	5	5	5.08	10.50	7.07	-945.624	37.28
	Total	26					
4	1	2	3.63	3.63	3.63	-1010.76	31.38
	2	2	4.78	4.78	4.78	-1028.144	38.04
	3	3	2.14	3.47	2.99	-1038.648	31.75
	4	1	0.00	0.00	0.00	-1061.329	27.09
	5	2	3.69	3.69	3.69	-972.002	30.24
	Total	10					

III. Datos Generales del proceso de selección de la proteína R0023

Config.	Clúster		Distancias al centro del Clúster			Solución Seleccionada	
	Nº	Cantidad	Mínima	Máxima	Promedio	FO	Radio
1	1	6	2.70	6.29	4.22	-400	18.65
	2	7	0.73	3.73	2.02	-417.55	21.82
	3	8	1.08	5.16	2.85	-384	20.55
	4	4	1.20	6.10	3.63	-433	19
	5	3	3.52	7.45	6.04	-465.96	22.5
	Total	28					
2	1	4	1.41	2.16	1.77	-429.714	21.49
	2	4	0.64	2.39	1.46	-430.609	14.4
	3	1	0.00	0.00	0.00	-476.11	13.24
	4	7	2.57	6.99	4.74	-446.842	22.05
	5	4	1.39	3.22	2.01	-418.832	19.05
	Total	20					
3	1	3	1.86	3.90	2.60	-422.531	15.15
	2	7	0.32	3.96	2	-411.597	20.78
	3	6	1.72	4.39	3.22	-442.953	16.9
	4	8	1.65	4.04	2.83	-401.087	22.48
	5	3	0.59	5.13	3.61	-464.691	15.93
	Total	27					
4	1	5	1.15	5.81	3.12	-455.705	15.41
	2	3	1.09	2.29	2	-440.459	18.83
	3	4	0.66	2.41	1.35	-436.976	14.79
	4	3	2.03	4.25	2.87	-475.374	15.04
	5	4	0.87	2.02	1	-423.919	18.32
	Total	19					

IV. Datos Generales del proceso de selección de la proteína R0033

Config.	Clúster		Distancias al centro del Clúster			Solución Seleccionada	
	Nº	Cantidad	Mínima	Máxima	Promedio	FO	Radio
1	1	3	2.77	6.55	4.97	-773.356	33.44
	2	5	2.80	6.94	4.83	-753.08	26.33
	3	4	3.89	8.88	6.42	-798.139	39.77
	4	8	2.02	12.93	5.54	-753.136	38
	5	1	0.00	0.00	0.00	-839.731	30.44
	Total	21					
2	1	2	5.39	5.39	5.39	-817.592	44.41
	2	2	3.53	3.53	3.53	-801.162	29.37
	3	8	1.08	4.10	3.00	-815.754	28.55
	4	2	7.14	7.14	7.14	-902.032	25.58
	5	4	2.55	7.88	5.35	-783.584	39.72
	Total	18					
3	1	4	2.36	7.79	4.98	-777.306	31.51
	2	9	2.21	8.30	4.16	-787.735	25.48
	3	5	2.64	11.13	5.73	-813.361	21.73
	4	6	0.64	4.65	3.11	-743.216	25.87
	5	2	3.25	3.25	3.25	-738.508	47.09
	Total	26					
4	1	3	2.47	5.55	4.01	-785.579	32.08
	2	5	1.74	6.61	3.67	-797.071	34.55
	3	6	1.55	5.37	3.84	-777.993	3078
	4	8	2.70	7.00	4.16	-742.79	38.27
	5	5	2.30	8.22	5.27	-826.783	28.17
	Total	24					

V. Comparación entre los métodos de enfriamiento en las proteínas

R0007	N	RMSD	GDT	LGAS	LGAQ	Promedio (%)
Boltzmann	80	95	70	80	85	82
Cauchy	90	70	80	95	90	85
Geométrico	70	75	65	70	80	72
Lundi-Mess	95	90	95	90	95	93
Promedio (%)	83.75	82.5	77.5	83.75	87.5	

R0015	N	RMSD	GDT	LGAS	LGAQ	Promedio (%)
Boltzmann	80	95	95	85	80	87
Cauchy	70	70	50	65	65	64
Geométrico	95	90	80	90	90	89
Lundi-Mess	90	85	85	95	95	90
Promedio (%)	83.75	85	77.5	83.75	82.5	

R0023	N	RMSD	GDT	LGAS	LGAQ	Promedio (%)
Boltzmann	75	45	80	65	70	67
Cauchy	95	85	75	90	95	88
Geométrico	65	90	90	60	65	74
Lundi-Mess	90	95	95	95	90	93
Promedio (%)	81.25	78.75	85	77.5	80	

R0033	N	RMSD	GDT	LGAS	LGAQ	Promedio (%)
Boltzmann	90	65	70	85	90	80
Cauchy	75	95	80	80	80	82
Geométrico	95	90	90	95	95	93
Lundi-Mess	85	75	95	90	85	86
Promedio (%)	86.25	81.25	83.75	87.5	87.5	

VI. Comparación entre las metaheurísticas SA y VMO en las proteínas

R0007	N	RMSD	GDT	LGAS	LGAQ	Promedio (%)
RS	98.02	99.01	99.01	99.01	99.01	98.81
VMO	99.01	97.03	98.02	97.03	98.02	97.82
Métrica más importante (%)	98.51	98.02	98.51	98.02	98.51	

R0015	N	RMSD	GDT	LGAS	LGAQ	Promedio (%)
RS	98.02	98.02	98.02	98.02	98.02	98.02
VMO	99.01	99.01	99.01	99.01	99.01	99.01
Métrica más importante (%)	98.51	98.51	98.51	98.51	98.51	

R0023	N	RMSD	GDT	LGAS	LGAQ	Promedio (%)
RS	99.01	99.01	99.01	99.01	99.01	99.01
VMO	98.02	98.02	97.03	97.03	97.03	97.43
Métrica más importante (%)	98.51	98.51	98.02	98.02	98.02	

R0033	N	RMSD	GDT	LGAS	LGAQ	Promedio (%)
RS	97.03	99.01	98.02	98.02	97.03	97.82
VMO	99.01	98.02	99.01	99.01	99.01	98.81
Métrica más importante (%)	98.02	98.51	98.51	98.51	98.02	

VII. Comparación entre SA y VMO contra predicciones internacionales.

	N		RMSD		GDT		LGAS		LGAQ	
	RS	VMO	RS	VMO	RS	VMO	RS	VMO	RS	VMO
R0007	35.64	41.58	52.48	29.7	61.39	51.49	23.76	17.82	38.61	32.67
R0015	23.76	27.72	45.54	54.46	57.43	64.36	36.63	41.58	29.7	38.61
R0023	65.35	50.5	76.24	73.27	68.32	57.43	70.3	43.56	66.34	59.41
R0033	87.13	99.01	89.11	85.15	91.09	92.08	88.12	92.08	95.05	88.12
Promedio	52.97	54.70	65.84	60.65	69.56	66.34	54.70	48.76	57.43	54.70