Docking Molecular Directo Entre Proteína y Ligando Flexible Utilizando Técnicas de Inteligencia Artificial

Nydia Paola Rondón Villarreal
Ingeniera de Sistemas

Universidad Industrial de Santander
Facultad Ingenierías Fisicomecánicas
Escuela de Ingeniería de Sistemas e Informática
Programa de Maestría en Ingeniería de Sistemas e Informática
Bucaramanga, 2011

Docking Molecular Directo Entre Proteína y Ligando Flexible Utilizando Técnicas de Inteligencia Artificial

Nydia Paola Rondón Villarreal

Trabajo de grado presentado para optar el título de Magister en Ingeniería de Sistemas e Informática

Directores del Proyecto:
Henry Arguello Fuentes, *Mpe.*
Rodrigo Gonzalo Torres Sáez, *PhD*

Universidad Industrial de Santander
Facultad Ingenierías Fisicomecánicas
Escuela de Ingeniería de Sistemas e Informática
Programa de Maestría en Ingeniería de Sistemas e Informática
Bucaramanga, 2011

a Dios

a mis Padres

a mi Hermana

a Ti...

Agradecimientos

Al profesor Rodrigo Torres por sus invaluables aportes, apoyo, confianza y excelente orientación en el desarrollo del presente proyecto. Profe mil gracias.

A Darío, María, Erick, Iván, Johann, Oveimar, mis compañeros del GIIB y del GIBIM, por su amistad, sus conocimientos, consejos, por todo el apoyo brindado y por su compañía en todo el tiempo de la maestría. Amigos mil gracias.

Al profesor Alfonso Mendoza por todo el apoyo y confianza brindados, por sus consejos y por buscar siempre lo mejor para todos los estudiantes del grupo.

Al profesor Henry Arguello por su dirección y orientación.

y a todas aquellas personas que directa o indirectamente me ayudaron en este proceso de formación.

Contenido

Índice de figuras

Lista de tablas

Resumen

TITULO: Docking Molecular Directo Entre Proteína y Ligando Flexible Utilizando Técnicas de Inteligencia Artificial [1]

AUTOR: Nydia Paola Rondón Villarreal. [2]

PALABRAS CLAVE: Docking molecular, complejo proteína-ligando, RMSD, Algoritmo Genético

El diseño de fármacos es un proceso ineficiente y costoso, que necesita ser mejorado para poder brindar a la humanidad medicinas seguras, efectivas y asequibles. Las herramientas computacionales han permitido mejorar este proceso mediante la simulación de las posibles interacciones existentes entre las proteínas y los posibles medicamentos. Sin embargo, dichas herramientas no suministran resultados fiables para todos los tipos de complejos proteína-ligando.

Una de estas herramientas computacionales es el docking molecular, que consiste en predecir el complejo más estable en términos de energía entre un receptor o proteína y un ligando o compuesto con las propiedades químicas necesarias para poder considerarse como posible fármaco. En los últimos años, el proceso de docking molecular se ha venido trabajando con algoritmos genéticos, búsqueda tabú, recocido simulado, entre otras técnicas y los resultados obtenidos han sido satisfactorios.

En el presente trabajo de investigación se desarrolló un proceso computacional de docking molecular utilizando una variante de un algoritmo genético de estado estable, en donde la función de evaluación de energía es aquella suministrada por el campo de fuerza AMBER. En la primera parte del documento se ilustra brevemente el proceso bioquímico de la unión de una proteína con un ligando, la selección de la técnica de inteligencia artificial utilizada y la selección de la función de energía. Posteriormente, se describen los pasos realizados para el desarrollo del proceso de docking molecular y finalmente, se presentan las simulaciones realizadas y los resultados obtenidos.

[1] Trabajo de grado

[2] Facultad de Ingenierías Físico Mecánicas. Escuela de Ingeniería de Sistemas e Informática. Director: Henry Arguello. Codirector: Rodrigo Torres.

Abstract

TITLE: Direct Molecular Docking Between Protein and Flexible Ligand Using Artificial Intelligence Techniques [3]

AUTHOR: Nydia Paola Rondón Villarreal [4]

KEY WORDS: Molecular Docking, protein-ligand complex, RMSD, Genetic Algorithms

Drug design is an inefficient and expensive process that needs to be improved to provide safe, effective and affordable medicines to the humanity.Computational tools have improved this process by simulating the possible interactions between proteins and ligands. However, these tools do not provide reliable results for all types of protein-ligand complexes.

One of these computational tools is molecular docking, which allows to predict the most stable complex in terms of energy between a protein or receptor and a ligand or compound with the chemical properties needed to be considered as a potential drug. In recent years, the molecular docking process has been working with genetic algorithms, tabu search, simulated annealing, among other techniques and results have been satisfactory.

In this research was developed a computational process of molecular docking using a variant of a steady-state genetic algorithm, where the energy evaluation function is that supplied by the AMBER force field. The first part briefly illustrates the biochemical process of the binding of a protein with a ligand, the selection of the artificial intelligence technique and the choice of the energy function. Then, the steps taken for the development of molecular docking process are described and finally the simulations and results are presented.

[3] Research work

[4] Faculty of Physical-Mechanical Engineerings. Systems engineering and informatics department. Advisor: Henry Arguello. Co-advisor: Rodrigo Torres

1. Docking Molecular: Modelando un Proceso Bioquímico

1.1 Organismos Vivos

Todo organismo vivo en la tierra está compuesto por células que controlan su estructura y funcionamiento. Estas células requieren de la presencia de cuatro grandes grupos de biomoléculas (carbohidratos, lípidos, proteínas y ácidos nucléicos), para que la vida exista.

Todos los organismos, incluidas las células, utilizan los carbohidratos como fuente de energía. Los lípidos además de cumplir funciones estructurales y de regulación del organismo, se convierten en una reserva concentrada de energía. Las proteínas son biomoléculas de gran tamaño que desempeñan un papel fundamental para la vida, están compuestas por aminoácidos y varían tanto en estructura como en función. Existen diversos criterios para clasificar las proteínas entre los cuales se encuentran la forma, la solubilidad, la composición química y su función. Según la forma las proteínas se clasifican en globulares y fibrosas, según la solubilidad se clasifican en insolubles, solubles y poco solubles, de acuerdo a la composición química se clasifican en simples y conjugadas, dentro de estas últimas se encuentran las lipoproteínas, glicoproteínas, fosfoproteínas, hemoproteínas, flavoproteínas y metaloproteínas [1]. Finalmente, según la función biológica que realizan, las proteínas se clasifican en enzimas, de transporte, de reserva, contráctiles, estructurales, de defensa y reguladoras. Por otra parte, los ácidos nucléicos son macromoléculas formadas por la repetición de monómeros llamados nucleótidos, que se encuentran al interior de las células. Existen dos tipos de ácidos nucléicos, el ácido desoxirribonucléico (ADN), el cual contiene la información genética del organismo y permite controlar el crecimiento, función y reproducción celular, y el ácido ribonucléico (ARN) que permite transferir información vital

durante la síntesis de proteínas (producción de las proteínas que necesita la célula para sus actividades y su desarrollo) [2].

1.2 Aminoácidos y Proteínas

Las proteínas son polímeros de aminoácidos unidos por medio de enlaces peptídicos. Este tipo de enlace es covalente y se forma al unir el grupo α-carboxilo del primer aminoácido y el grupo α-amino del segundo. A la unión de estos dos aminoácidos se le conoce como dipéptido. De manera similar, tres aminoácidos se pueden unir mediante dos enlaces peptídicos para formar un tripéptido, cuatro aminoácidos forman un tetrapéptido y así sucesivamente. A la unión de pocos aminoácidos, generalmente menos de 10, se le conoce con el nombre de oligopéptido, mientras que a la unión de múltiples aminoácidos se le conoce como polipéptido.

Las proteínas se caracterizan por tener miles de aminoácidos en un orden específico. A esta secuencia de aminoácidos se le conoce como estructura primaria de una proteína y determina el plegamiento de la misma, generando una única estructura tridimensional y en consecuencia, determinando la función de dicha proteína. Las funciones de las proteínas dependen casi siempre de interacciones con otras moléculas y éstas se pueden ver significativamente afectadas por cambios, sútiles o grandes, en la conformación de la proteína. Miles de enfermedades genéticas humanas se deben a la producción de proteínas defectuosas [2].

Las funciones de varias proteínas involucran la unión reversible con otras moléculas. Una molécula que se une a una proteína de forma reversible es conocida como ligando y las interacciones de éste con la proteína son las que le permiten a un organismo responder de forma rápida y reversible a los cambios del entorno y del metabolismo. Este tipo de interacciones son específicas, es decir, la proteína puede discriminar entre miles de moléculas diferentes en su ambiente y seleccionar una o pocas moléculas a las cuales unirse. Esta unión se realiza en un sitio específico de la proteína llamado sitio de enlace, o sitio activo para el caso de las enzimas, el cual es complementario al ligando tanto en tamaño, forma, carga y carácter hidrofóbico o hidrofílico [2]. La unión de

una proteína con un ligando rara vez se realiza mediante enlaces covalentes debido a su naturaleza reversible. Por lo anterior, la mayoría de las interacciones proteína-ligando se realizan a través de fuerzas intermoleculares, no covalentes, entre las que se destacan las interacciones electrostáticas, de van der Waals e hidrofóbicas y los puentes de hidrógeno. En las primeras, una carga positiva y una carga negativa se atraen mutuamente y las interacciones pueden ser de tres tipos: iónicas, dipolo-ión y dipolo-dipolo. Las interacciones de van der Waals se forman entre grupos moleculares neutros cercanos entre sí. Las interacciones hidrofóbicas dependen del elevado grado de desorden (entropía) del agua y son características de aquellas moléculas que no pueden interaccionar fácilmente con el agua. Finalmente, los puentes de hidrógeno, se establecen entre moléculas capaces de generar cargas parciales y se forman por átomos de hidrógeno localizados entre átomos electronegativos.

1.3 Docking Molecular

En el diseño de medicamentos asistido por computador (*in silico*), el principal objetivo es encontrar el complejo más estable (aquel con la energía más baja), entre un receptor (proteína) y un ligando. En este caso el término ligando hace referencia a una molécula, generalmente pequeña, que cumple ciertas propiedades químicas para poder ser considerada como un posible fármaco. Lipinski en el año 1997 planteó una serie de reglas, conocidas como "Rule–of–Five", que permiten realizar un filtrado de aquellos ligandos que no sirven como medicamentos en las personas. Una explicación más detallada de éstas y otras reglas desarrolladas, se puede encontrar en [3–5].

El proceso computacional que permite encontrar complejos energéticamente favorables, entre dos tipos de moléculas, se conoce como docking molecular. Dependiendo de estas moléculas de entrada, las herramientas de docking se pueden clasificar en dos tipos. El primero hace referencia al docking macromolecular, en el cual dos macromoléculas como las proteínas o el ADN se unen. El segundo tipo es el docking de pequeñas moléculas, en el cual una macromolécula se une a una pequeña molécula. Para el caso del diseño de medicamentos este proceso es conocido como docking entre proteína y ligando [6].

El docking molecular tiene tres enfoques: rígido, semiflexible y flexible [7]. En el primer caso, tanto el receptor (en este caso una proteína) como el ligando son considerados como estructuras químicas rígidas, limitando el problema a la búsqueda de las posiciones de cada uno de ellos en el espacio [8]. En el segundo, se considera al receptor rígido y al ligando flexible (su estructura puede cambiar de acuerdo a los enlaces rotables que posea), este enfoque tiene un espacio de búsqueda más amplio, puesto que no sólo se debe encontrar las posiciones del receptor y el ligando sino que se debe buscar una estructura geométrica óptima para el ligando [9].

El tercer enfoque, es el más complejo y costoso computacionalmente, puesto que tanto el receptor como el ligando son considerados flexibles, ocasionando que el espacio de búsqueda para el complejo con la estructura más estable sea de un tamaño de varios ordenes de magnitud mayor que en los casos anteriores [10]. Sin embargo, existe una reducción considerable del espacio de búsqueda si el sitio activo de la proteína es conocido. En el caso en el que el sitio activo no es conocido, se habla de docking ciego o docking global, mientras que el docking directo hace referencia al caso en el que se conoce el sitio activo de la proteína [7].

1.4 Función de Evaluación de Energía

Un aspecto importante en el proceso de docking molecular es la función de energía que permite calcular la energía libre del complejo proteína-ligando, la cual es un indicador de la estabilidad del mismo. Estas funciones se pueden agrupar en empíricas, basadas en campos de fuerza y basadas en el conocimiento [7].

Las funciones empíricas se basan en las propiedades de la estructura del complejo receptor-ligando. La primera desventaja de este tipo de funciones, consiste en que es difícil determinar con exactitud lo que cada término de la función aporta y es difícil evaluar de donde provienen los errores. Una segunda desventaja consiste en que sólo se obtienen predicciones exitosas si las moléculas hacen interacciones similares a aquellas que realizaron los comple-

jos en el conjunto de entrenamiento [11].

Las funciones basadas en el campo de fuerza son las más utilizadas en el docking molecular, y permiten calcular la energía de un complejo en función de sólo las posiciones de los núcleos. Este tipo de funciones permite obtener buenos resultados en un corto tiempo computacional [12].

Finalmente, en las funciones basadas en modelos de conocimiento, para cada tipo de par de átomos, se cuenta la frecuencia de aparición dependiendo de la distancia. Los histogramas resultantes se convierten en una función de energía y el valor para un complejo dado se calcula mediante la suma de los valores de la función de energía para todos los pares del complejo [13].

En general, las funciones de evaluación para el caso de docking rígido sólo contemplan la energía de interacción entre el receptor y el ligando. Para los casos del docking semiflexible y flexible, se deben contemplar tanto la energía de interacción entre las moléculas, como la energía interna del ligando para el primer caso, y la energía interna de ambas moléculas para el segundo. La función de energía general, se puede escribir mediante la siguiente ecuación [9]:

$$E_{total} = E_{inter} + E_{intra} \quad [kcal/mol] \tag{1}$$

donde E_{total} representa la energía total del complejo, E_{inter} la energía intermolecular y E_{intra} la energía intramolecular.

Teniendo en cuenta lo anterior, el proceso de docking molecular se puede ver como un problema de optimización con un espacio de búsqueda de tamaño considerable, que debe recorrerse para localizar el complejo más estable, aquel con la menor energía libre de enlace, que corresponde al mejor mínimo local.

Las técnicas de inteligencia computacional, se han convertido en herramientas útiles en la solución de este tipo de problemas y permiten obtener mejores resultados que los obtenidos con los métodos de búsqueda y optimización tradicionales [14,15]. Algunas de las técnicas de inteligencia computacional para problemas de optimización son la computación evolutiva, algoritmos genéticos, enjambre de partículas y colonia de hormigas [16].

En los últimos años, el problema del docking molecular se ha venido trabajando con algoritmos genéticos, búsqueda tabú, recocido simulado, entre otras técnicas y los resultados obtenidos han sido satisfactorios [17–20].

1.5 Características del Proceso de Docking Molecular Semiflexible

En el presente trabajo de investigación se realiza un proceso de docking molecular directo semiflexible entre proteína y ligando con las siguientes características: Las proteínas se limitan a enzimas para las cuales el sitio activo se encuentre registrado en la base de datos Catalytic Site Atlas [21]. En el proceso de docking molecular se utiliza la estructura tridimensional de la proteína presente en el complejo del conjunto de prueba, esto es, la estructura que adopta la proteína cuando ésta se une al ligando. La estructura de la proteína se mantiene rígida durante toda la simulación y los aminoácidos pertenecientes al sitio activo son delimitados por una malla tridimensional con espacios de 0.2 $\mathring{A}$.

Las moléculas de agua son removidas para disminuir la complejidad computacional de las simulaciones. La flexibilidad del ligando se trabajó diviendo dicha molécula en partes rígidas que giran de acuerdo con los enlaces rotables que posea. La función de evaluación de la energía utilizada en este estudio fue la del campo de fuerza AMBER [22]. Para el proceso de optimización se utilizó un algoritmo genético y para la validación de los resultados se realizó el cálculo de la distancia RMSD (Root Mean Square Deviation) entre la estructura obtenida y la estructura cristalina. Los átomos de hidrógeno se ignoraron en la evaluación de esta medida, debido a que los complejos suministrados por el Protein Data Bank no incluyen dichos átomos.

Finalmente, entre las principales características de los complejos proteína-ligando seleccionados para el conjunto de prueba, se encuentra que las estructuras fueron determinadas mediante difracción de rayos X de sus formas cristalinas, los complejos no poseen enlaces covalentes entre la proteína y el ligando, la molécula del ligando no contiene elementos químicos pocos comunes como Be, B, Si y átomos metálicos y los complejos no tienen múltiples

ligandos en un sitio de enlace común [23].

2. Métodos de Optimización utilizados en Docking Molecular

Teniendo en cuenta que el proceso de docking molecular directo y semiflexible se puede ver como un problema de optimización con múltiples mínimos locales, la utilización de métodos tradicionales de búsqueda como por ejemplo el método de Monte Carlo, Gradiente descendiente, entre otros, no es la opción más recomendada, puesto que dichos métodos pueden converger a mínimos locales y el proceso de realizar múltiples simulaciones con condiciones iniciales diferentes, aparte de incrementar el tiempo computacional, resulta ineficiente cuando la cantidad de mínimos locales incrementa de forma exponencial con el tamaño del problema [24]. El uso de técnicas de inteligencia artificial en el proceso de docking molecular ayuda a obtener un mejor rendimiento, debido a que estos métodos de optimización permiten buscar nuevas soluciones que compitan con el actual mínimo local, evitando así, que el resultado dependa de la configuración inicial del algoritmo y evitando converger a un mínimo local [24]. El uso de estas técnicas en el campo de docking molecular se ve respaldado con los múltiples trabajos de investigación que las utilizan. Algunos ejemplos de estos trabajos se listan a continuación.

En el año 1995, Jones et al. [25] desarrollaron uno de los primeros prototipos de docking molecular semiflexible, con un algoritmo genético de estado estable. En este prototipo se tuvieron en cuenta los posibles enlaces de hidrógeno que se pueden formar entre el ligando y el receptor, con el objetivo de mejorar el rendimiento de dicho algoritmo. Las simulaciones se realizaron 50 veces para cada complejo y los resultados obtenidos fueron favorables. Sin embargo, Jones et al. resaltan que múltiples mejoras pueden ser realizadas al prototipo diseñado en [25]. Dos años más tarde, Jones y su equipo de trabajo presentan a la comunidad científica, un nuevo prototipo conocido como GOLD (Genetic Optimisation for Ligand Docking) [26], el cual presenta

ciertas mejoras con respecto al anterior prototipo. En este estudio, el algoritmo genético desarrollado fue probado con 100 complejos proteína-ligando y el porcentaje de acierto fue de 71 % teniendo como umbral un valor de RMSD entre 2 y 3 $\mathring{A}$. En ese mismo año, Westhead et al. [15] realizaron un estudio comparativo entre 4 métodos heurísticos (algoritmo genético, programación evolutiva, recocido simulado y búsqueda tabú) aplicados al caso del docking molecular. Los resultados indicaron que todos los métodos analizados son efectivos y tienen un rendimiento satisfactorio. Sin embargo, como casos particulares, se destacan el algoritmo genético que arroja la menor energía promedio en los complejos y el algoritmo de búsqueda tabú que permite encontrar con alta probabilidad un mínimo local cercano al mínimo global.

Para el año 2003 Bursulaya et al. [23] llevaron a cabo un estudio comparativo de 5 programas para docking molecular flexible: DOCK 4.0, FlexX 1.8, AutoDock 3.0, GOLD 1.2 e ICM 2.8, en el cual las tasas de acierto para un RMSD menor a 2 $\mathring{A}$ fueron 46 %, 30 %, 35 %, 46 % y 76 % respectivamente. Los resultados sugieren que los programas evaluados realizan un buen trabajo de docking molecular y entre los más destacados se encuentran el software ICM que utiliza el método de Monte Carlo, el software GOLD que utiliza un algoritmo genético y el software DOCK que utiliza un algoritmo de reconstrucción incremental. Un año después Magalhães et al. [27] desarrollaron un algoritmo genético de estado estable para simular un proceso de docking molecular semiflexible. En las simulaciones utilizaron una población de 1000 individuos y 1.000.000 de evaluaciones de energía, utilizando para ello, la función de evaluación del campo de fuerza GROMOS [28]. La tasa de aciertos de este algoritmo varía entre el 10 % y el 60 % dependiendo del ligando trabajado.

Por otro lado, Liu et al., en el año 2005 [29] desarrollaron un nuevo método, llamado SODOCK, para resolver problemas de docking molecular utilizando la técnica de Enjambre de Partículas. Los resultados obtenidos en este trabajo fueron comparados con los resultados de varios software comerciales como GOLD 1.2, AutoDock 3.05, DOCK 4.0 y FlexX 1.8 y tomando sólo el mejor resultado de 30 simulaciones para cada uno de los 16 complejos analizados en cada herramienta, SODOCK obtiene el menor RMSD en 11 complejos. Posteriormente, en el año 2006, Oduguwa et al. [30] desarrollaron un algoritmo de docking molecular entre proteína y ligando, basado en optimización

multiobjetivo con algoritmos evolutivos. En estas simulaciones, el tamaño de la población fue de 100 individuos y el número de generaciones fue de 500, lo cual se traduce en 50.000 evaluaciones de energía. En este estudio los valores del RMSD promedio varían entre 15 $\mathring{A}$ y 21 $\mathring{A}$, indicando que los resultados obtenidos varían en gran medida con respecto a otros estudios realizados. En este mismo año, Chen et al. [31] desarrollaron un algoritmo de docking molecular basado en la técnica de Enjambre de partículas, el cual titularon Tribe-PSO. En este estudio los valores de energía obtenidos para 100 casos de prueba fueron comparados con los valores de energía arrojados por el software AutoDock. Los resultados muestran que Tribe-PSO permite obtener valores de energía más bajos que el software AutoDock.

Entre otros estudios que utilizan algoritmos genéticos se encuentra el trabajo de Wang et al., 2008 [32] en donde se desarrolla un proceso de docking molecular utilizando la paralelización de un algoritmo genético Lamarckian, el trabajo realizado por Kang et al., en el año 2009, [33] que propone un algoritmo genético mejorado que utiliza técnicas avanzadas como múltiples poblaciones, la auto-adaptación y el castigo cuasi-exacto. El porcentaje de acierto para este trabajo es del 66.2 %, teniendo como umbral un valor de RMSD menor o igual a 2 $\mathring{A}$. Finalmente, Thiriot y Monard en 2009 [34] desarrollaron un método de docking molecular que calcula la energía mediante un enfoque lineal cuántico semi-empírico y en el proceso de optimización utiliza un algoritmo genético.

2.1 Métodos de Optimización en softwares comerciales

En la actualidad existen múltiples software comerciales para realizar procesos de docking molecular, entre los que se destacan AutoDock [35], DockVision [36], GOLD [37], DOCK [38] y FlexX [39], entre otros. Al revisar los métodos de optimización utilizados por estos software altamente reconocidos, se encuentra que la mayoría de ellos permite trabajar con algoritmos genéticos. La tabla 1 muestra los métodos de optimización presentes en dichas herramientas computacionales.

Software	**Método de Optimización utilizado**
AutoDock	Algoritmo Genético Lamarckiano
DockVision	Monte Carlo, Algoritmo Genético
GOLD	Algoritmo Genético
DOCK	Gradiente Conjugado
FlexX	Método de Construcción Incremental

Tabla 1: Métodos de optimización de los software comerciales más utilizados

2.2 Selección de la Técnica de Inteligencia Artificial

Para el proceso de selección de la técnica de inteligencia artificial a utilizar se tuvieron en cuenta los siguientes aspectos:

- La técnica seleccionada debe haber sido utilizada en estudios previos de docking molecular, para garantizar el éxito en la implementación del primer prototipo de docking molecular al interior de la Universidad Industrial de Santander.
- Los resultados de los estudios consultados deben indicar que es posible obtener un valor de RMSD menor de 2 $\mathring{A}$ con la técnica de inteligencia artificial seleccionada.
- La técnica seleccionada debe ser utilizada en al menos un software comercial altamente utilizado. Lo anterior permite tener una certeza de la validez y fiabilidad de los resultados mostrados en los estudios presentes en la literatura.

El primer y segundo criterio, limitan la selección de la técnica de inteligencia artificial a algoritmos genéticos y enjambre de partículas, debido a que estos métodos de optimización han sido utilizados en estudios de docking molecular que obtienen valores de RMSD menores de 2 $\mathring{A}$. Finalmente, la selección se realiza en base al último criterio, puesto que entre los software comerciales más conocidos, no existe una implementación de la técnica de enjambre de partículas, en contraste con la técnica de algoritmos genéticos para la cual existen tres software renombrados que la utilizan. Por lo anterior, se seleccionan los algoritmos genéticos como la técnica de inteligencia artificial a utilizar en la implementación del proceso de docking molecular.

3. Selección de la Función de Evaluación de Energía Basada en el Campo de Fuerza

El campo de fuerza se conoce como la función que permite calcular la energía potencial molecular, junto con el conjunto de parámetros asociados a ésta. Los términos presentes en la función del campo de fuerza se pueden agrupar en aquellos relacionados con los átomos unidos por enlaces covalentes y aquellos que representan las interacciones electrostáticas y de van der Waals provenientes de los átomos que no están enlazados [7].

La forma general de la función del campo de fuerza es [7]:

$$E_{total} = E_{str} + E_{bend} + E_{tors} + E_{vdw} + E_{elec} \quad [kcal/mol] \tag{2}$$

en donde E_{total} es la energía total de la molécula, que corresponde a la suma de los términos E_{str}, E_{bend}, E_{tors}, E_{vdw} y E_{elec}, los cuales se ilustran en la figura 1.

Algunos software comerciales utilizan funciones de energía basadas en el campo de fuerza. Ejemplo de ello, son los software AutoDock cuya función de energía está basada en el campo de fuerza AMBER, el software SYBYL que posee dos funciones de energía, D-Score y G-Score, basadas en el campo de fuerza Tripos 5.2 [40] y el software GOLD que posee entre sus funciones de energía, la función GoldScore basada en el conjunto de parámetros GOLD [37].

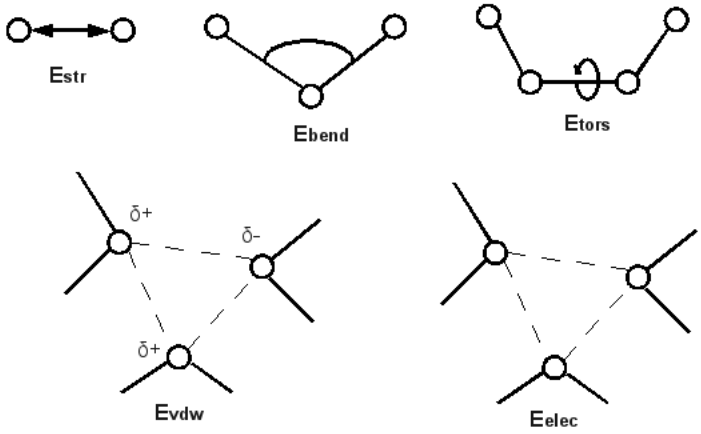


Figura 1: E_{str} describe el cambio de energía que ocurre cuando un enlace se extiende o contrae a partir de su longitud ideal, E_{bend} corresponde al cambio de energía que ocurre cuando se varía el ángulo de un enlace respecto de su ángulo ideal, E_{tors} describe el cambio de energía producido por la rotación de un enlace, E_{vdw} representa las interacciones entre los átomos que no están conectados directamente y finalmente, E_{elec} representa las fuerzas electrostáticas entre los átomos que no están enlazados.

3.1 Criterios de selección

La selección de la función de evaluación de energía, implementada en el presente trabajo de investigación, se basó en los siguientes criterios:

- La función de evaluación debe ser basada en un campo de fuerza.
- La función debe tener un buen rendimiento en los estudios comparativos existentes en la literatura, sin que esto implique que se debe seleccionar la de más alto rendimiento.
- Los estudios comparativos deben tener condiciones de simulación muy parecidas a las condiciones seleccionadas en el desarrollo del presente trabajo de investigación. De esta forma, la efectividad de la función seleccionada no debería verse afectada de forma significativa.
- El tiempo de cómputo requerido para llevar a cabo la evaluación de la energía de un complejo, debe ser el menor posible, debido a que en un proceso computacional de docking molecular es necesario realizar una gran cantidad de evaluaciones de energía.
- Facilitar la implementación y adaptación de la función en el proceso de docking molecular desarrollado.

3.2 Proceso de selección

Teniendo en cuenta el primer criterio de selección, las funciones de evaluación de energía opcionadas son las siguientes:

1. Autodock:: Amber [41]
2. SYBYL:: D-Score [42]
3. SYBYL:: G-Score [42]
4. GOLD:: GoldScore [42]

3.2.1 Estudios Comparativos y Selección

En la literatura se encuentran varios estudios comparativos entre las diferentes funciones de evaluación de energía utilizadas en procesos de docking molecular [42–45]. La mayoría de estos estudios realiza comparaciones entre funciones de evaluación empíricas, basadas en el campo de fuerza y basadas en el conocimiento. En el presente proyecto de investigación se optó por trabajar con funciones basadas en el campo de fuerza debido a que permiten obtener buenos resultados en un corto tiempo computacional y por lo tanto, sólo se tuvieron en cuenta los resultados obtenidos para las funciones previamente seleccionadas, en aquellos estudios que eliminan o ignoran las moléculas de agua que rodean a la proteína y al ligando y que contemplan al menos dos funciones de las seleccionadas previamente.

Renxiao Wang et al., 2003 [41], realizaron un estudio comparativo entre 11 diferentes funciones de evaluación de energía, de las cuales 3 eran basadas en el campo de fuerza (AutoDock, SYBYL::G-Score y SYBYL::D-score). En este estudio, las funciones de evaluación fueron probadas con 100 complejos proteina-ligando, se eliminaron las moléculas de agua y la tasa de aciertos estaba determinada por un valor de RMSD menor de 2 $\mathring{A}$. De las tres funciones de interés, la función del software AutoDock fue la que obtuvo una mayor tasa de aciertos, seguida por la función SYBYL::G-Score y finalmente la función SYBYL::D-Score. En la tabla 2 se puede apreciar los resultados obtenidos en este estudio, para las funciones antes mencionadas.

Posteriormente, en el año 2009, Tiejun Cheng et al. [42] compararon 16 diferentes funciones de evaluación de energía, entre las cuales se encuentran tres funciones basadas en el campo de fuerza (GOLD::GoldScore, SYBYL::G-

Función de evaluación	**Tasa de aciertos RMSD < 2 $\mathring{A}$ %**
AutoDock:: Amber	62
SYBYL::G-Score	42
SYBYL::D-Score	26

Tabla 2: Tasa de aciertos para las funciones, basadas en el campo de fuerza, del estudio [41]

Score y SYBYL::D-Score). Para este estudio comparativo, los autores escogieron un conjunto de prueba de 195 complejos proteína-ligando, las moléculas de agua fueron eliminadas y la tasa de aciertos incluía complejos con un valor de RMSD menor de 2 $\mathring{A}$. Los resultados obtenidos, indican que de las tres funciones de interés la función GOLD::GoldScore fue la que obtuvo un mejor desempeño, seguida por SYBYL::G-Score y SYBYL::D-Score. La tabla 3 presenta las tasas de acierto para las funciones basadas en el campo de fuerza analizadas.

Función de evaluación	**Tasa de aciertos RMSD < 2 $\mathring{A}$ %**
GOLD::GoldScore	68
SYBYL::G-Score	42
SYBYL::D-Score	31

Tabla 3: Tasa de aciertos para las funciones, basadas en el campo de fuerza, del estudio [42]

Teniendo en cuenta que la función del software AutoDock basada en el campo de fuerza AMBER y la función GOLD::GoldScore fueron las que tuvieron un mejor desempeño en los estudios consultados, se procedió a escoger una de estas dos funciones, en base a los demás criterios de selección.

La ecuación que representa la función de evaluación GOLD::GoldScore es

$$GoldScore = E_{H-b,e} + E_{v,e} + E_{v,i} + E_{t,int} \tag{3}$$

en donde $E_{H-b,e}$ y $E_{v,e}$ corresponde a la energía proveniente de los enlaces de hidrógeno y la energía de van der Waals entre la proteína y el ligando, respectivamente. El término $E_{v,i}$ hace referencia a la energía de van der Waals al interior del ligando y el término $E_{t,i}$ corresponde a la energía de torsión al interior del mismo.

la ecuación que corresponde a la función de evaluación del campo de fuerza AMBER es

$$Amber = E_b + E_a + E_d + E_{e-v} \tag{4}$$

en donde E_b representa la energía proveniente de los enlaces covalentes, E_a corresponde a la energía proveniente de los ángulos entre los enlaces, E_d indica la energía proveniente de los ángulos de torsión y finalmente, E_{e-v} corresponde a la energía electrostrática y de van der Waals tanto del ligando como de la unión proteína-ligando.

Teniendo en cuenta que la precisión de la evaluación de la energía aumenta a medida que se incrementan los aspectos bioquímicos evaluados y que la función GOLD::GoldScore incluye un término específico para evaluar la energía proveniente de los enlaces de hidrógeno, en contraste con la función del campo de fuerza AMBER, se espera que la precisión de los resultados sea mayor en la primera función, pero a su vez, se espera un incremento del tiempo computacional, debido a que éste aumenta a medida que el número de términos a evaluar se hace mayor [46]. Considerando que en un proceso de docking molecular el cálculo de la energía del complejo debe realizarse varias miles de veces, se deben utilizar aquellas funciones que requieran cortos tiempos de cómputo [11, 13]. Por lo anterior, se escoge trabajar con la función de evaluación del campo de fuerza AMBER.

Finalmente, la energía total calculada para cada complejo en el proceso de docking molecular semiflexible desarrollado, fue calculada como la suma de la energía interna del ligando más la energía de interacción entre éste y la proteína.

4.Proceso Computacional de Docking Molecular

El desarrollo del proceso de docking molecular directo y semiflexible fue llevado a cabo mediante dos etapas tituladas: Selección del conjunto de prueba e implementación del algoritmo de optimización. La efectividad de este proceso computacional es analizada en una etapa posterior titulada Validación de resultados.

4.1 Selección del conjunto de prueba

Al realizar un proceso de docking molecular es necesario verificar los resultados obtenidos con el fin de determinar la tasa de aciertos de dicha herramienta computacional.

La prueba que generalmente se aplica a un algoritmo de docking molecular, consiste en realizar simulaciones con complejos proteína-ligando donde se conoce la estructura 3D de las moléculas. La precisión de la predicción de la estructura de los complejos proteína-ligando, se puede evaluar mediante la desviación cuadrática media (RMSD) existente entre la estructura predicha y la estructura cristalina [11], la cual está dada por:

$$RMSD = \sqrt{\frac{\sum_{i=1}^{N_{at}} d_i^2}{N_{at}}} \tag{5}$$

donde N_{at} es el número de átomos sobre los cuales se va a calcular el RMSD y d_i es la distancia existente entre las coordenadas del átomo i en las dos estructuras, cuando éstas se superponen.

Para realizar dicho proceso de verificación se hace necesario tener un conjunto de prueba de complejos proteína-ligando con una estructura de alta resolución y además es deseable que dicho conjunto tenga una buena variedad de estructuras posibles [42]. Por lo anterior, se seleccionaron en el presente trabajo, aquellos complejos presentes en el conjunto refinado de la base de datos PDBbind (versión 2009) [47], que están presentes en la base de datos Catalytic Site Atlas [48].

Para cada complejo seleccionado es necesario tener los archivos del receptor, del ligando y del complejo para el cual la estructura cristalina es conocida. Una representación gráfica de la obtención de los archivos para un complejo X, se puede apreciar en la figura 2.

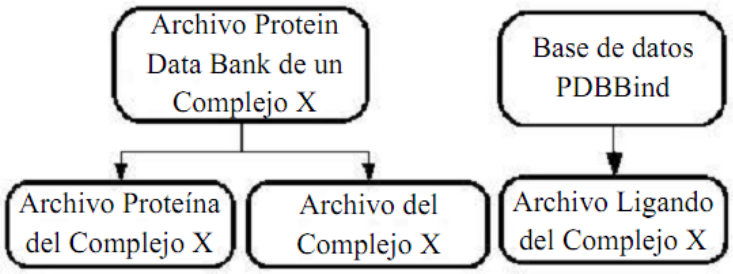


Figura 2: Para un complejo X, los archivos tando de la proteína como del mismo complejo, se obtienen a partir del archivo suministrado por el Protein Data Bank (PDB), mientras que el archivo del ligando se obtiene de la base de datos PDBBind.

4.1.1 Obtención y pre-procesamiento del archivo de la proteína (Receptor)

Para el caso de la proteína se hizo necesario modificar el archivo del complejo suministrado por el Protein Data Bank (PDB) [49]. La primera modificación consiste en agregar las cargas de Kollman utilizando para ello el software Autodock 4.2 [35]. Seguidamente se procede a generar un archivo de extensión .pdbq en el cual se eliminan los átomos del ligando y las moléculas de agua y se selecciona la opción que permite imprimir los campos *CONNECT*, aquellos campos del archivo PDB que indican las conexiones que tiene cada uno de los átomos presentes en la molécula, basados en las distancias entre los átomos. El archivo que se obtiene es con el cual se inicia el proceso de simulación.

El esquema para la creación del archivo de la proteína se ilustra en la figura 3.

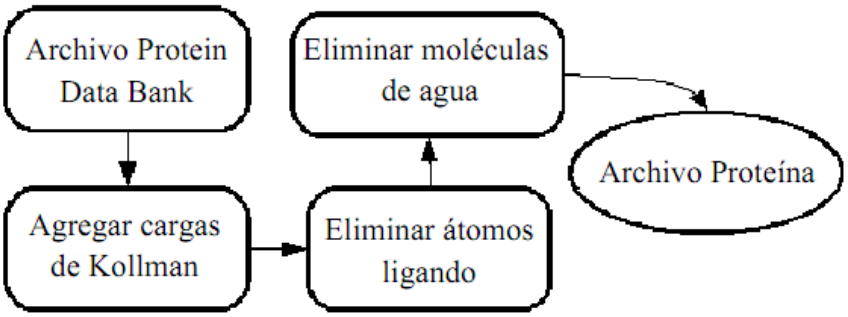


Figura 3: Creación del archivo de la proteína mediante el software Autodock 4.2

4.1.2 Obtención y pre-procesamiento del archivo del ligando

En el caso del ligando, el archivo utilizado es aquel suministrado por la base de datos PDBBind, el cual se convierte a un archivo de extensión .pdb utilizando el software de acceso libre Openbabel [50].

En el presente trabajo se contempla el ligando como flexible, por lo que se hace necesario conocer los enlaces rotables que éste posee y para ello se utiliza el software Autodock 4.2, que le permite al usuario seleccionar los enlaces rotables que desee y crear un archivo con extensión .pdbqt el cual será utilizado para generar el archivo final del ligando para el proceso de simulación.

Es importante destacar que el archivo generado por el software Autodock 4.2, ignora los átomos de hidrógeno, impidiendo utilizar todas las ventajas del campo de fuerza seleccionado. Por esta razón, se desarrolló un algoritmo que permitiera comparar el archivo de salida (archivo de extensión .pdbqt) con el de entrada (archivo de extensión .pdb) y de esta manera agregar los átomos de hidrógeno que han sido removidos. Para que el algoritmo implementado pueda ser utilizado exitosamente se debe garantizar que cada uno de los átomos posea un nombre diferente y exista su correspondiente campo *CONNECT*.

El siguiente paso consiste en traducir todos los átomos del ligando a la notación que utiliza el campo de fuerza seleccionado. En el presente trabajo se utilizó el campo de fuerza AMBER [22]. Un esquema gráfico de la obtención del archivo del ligando, se puede apreciar en la figura 4.

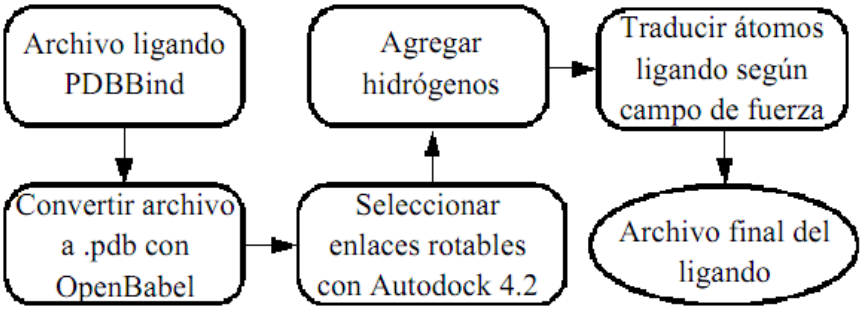


Figura 4: El archivo suministrado por la base de datos PDBBind se convierte a extensión .pdb, se seleccionan los enlaces rotables en el software Autodock 4.2, se agregan los hidrógenos con el algoritmo desarrollado y finalmente se traducen los átomos de acuerdo al campo de fuerza AMBER. El archivo resultante es con el cual se realiza la simulación de docking molecular.

4.1.3 Obtención y pre-procesamiento del archivo del complejo

Para el archivo de cada complejo perteneciente al conjunto de prueba, se hizo necesario eliminar las moléculas de agua existentes en el archivo suministrado por el PDB.

Finalmente, los tres archivos, se revisan y corrigen, en caso de ser necesario, obteniendo así, un conjunto de prueba que permite realizar las simulaciones de Docking Molecular.

4.2 Implementación del algoritmo de optimización

Para llevar a cabo el proceso de optimización, es necesario representar la información correspondiente a las estructuras tridimensionales, tanto de la proteína como del ligando, en estructuras de datos que puedan ser trabajadas en un algoritmo genético. En el presente trabajo de investigación se

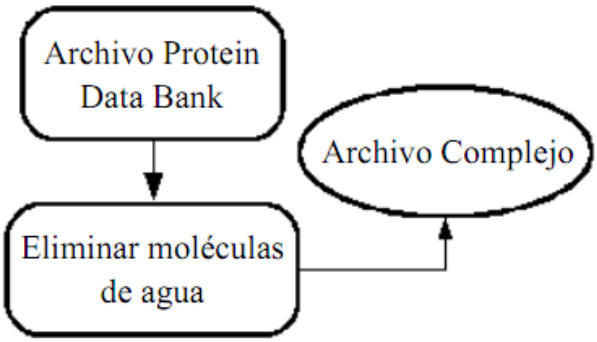


Figura 5: Al archivo del complejo suministrado por el PDB se le eliminan las moléculas de agua y el archivo resultante es el que se utiliza en el proceso de validación de resultados.

desarrolló una variante de un algoritmo genético de estado estable [24].

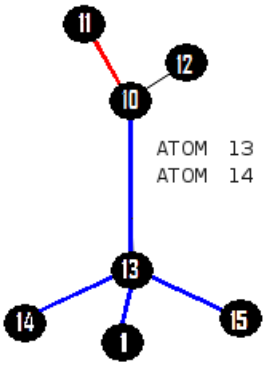
En términos generales, un algoritmo genético imita el proceso de evolución mediante la manipulación de una colección de estructuras de datos llamadas cromosomas. Cada uno de estos cromosomas codifica una posible solución al problema que se está trabajando. En este caso, cada cromosoma codifica la conformación de un posible complejo proteína-ligando.

El proceso de optimización desarrollado, inicia con la representación en texto plano de las estructuras tridimensionales de las moléculas (proteína y ligando), seguido por la generación de un complejo base, la creación de la población inicial, un proceso iterativo que incluye evaluación, cruce y mutación y finalmente, los resultados.

4.2.1 Representación en texto plano de la estructuras tridimensionales de la proteína y el ligando

La estructura tridimensional de una molécula puede ser representada mediante un archivo de texto plano que incluya la información de cada uno de los átomos presentes en la molécula, así como de las conexiones existentes entre ellos. Por ejemplo, los archivos provenientes del Protein Data Bank, permiten representar una estructura tridimensional a través de texto plano. En estos archivos, los renglones etiquetados como *ATOM* incluyen información de los átomos como por ejemplo, el símbolo, posición espacial, carga, entre otros. Las líneas del archivo etiquetadas como *CONNECT*, poseen una

lista de números, en la cual el primer número corresponde al átomo analizado y los siguientes corresponden a los átomos que forman un enlace covalente con este primero. En la figura 6 se puede apreciar un ejemplo de la representación de la información de un fragmento de molécula.




```
ATOM  13  C2  LOF A   1  -2.496  29.735  -8.525  1.00 -0.65   C
ATOM  14  O2  LOF A   1  -3.143  29.951  -7.521  1.00 -0.65   O

CONECT   11   10
CONECT   12   10
CONECT   13    1   10   14   15
CONECT   14    2   13
CONECT   15    3    4   13   16
CONECT   16   15   17   17   21
```

Figura 6: Se puede observar que para el átomo número 13, su correspondiente renglón *ATOM* indica que es un átomo de Carbono y su línea del archivo etiquetada como *CONNECT* indica que forma enlaces covalentes con los átomos 1, 10, 14 y 15.

Una molécula se puede representar computacionalmente mediante una matriz de coordenadas **C** que se conectan entre sí, de acuerdo con una matriz de conexiones **Mc**.

Teniendo en cuenta lo anterior, la estructura tridimensional del ligando se representa por medio de una matriz de coordenadas, $\mathbf{C_{lig}}$, y una matriz de conexiones $\mathbf{Mc_{lig}}$. A su vez, la matriz de coordenadas $\mathbf{C_{pr}}$ y la matriz de conexiones $\mathbf{Mc_{pr}}$ representan la estructura tridimensional de la proteína. Finalmente, el sitio activo de la proteína se representa mediante la matriz de coordenadas, $\mathbf{C_{pr-sa}}$, proveniente de $\mathbf{C_{pr}}$ y la matriz de conexiones, $\mathbf{Mc_{pr-sa}}$, proveniente de $\mathbf{Mc_{pr}}$. La matriz $\mathbf{C_{pr-sa}}$ está conformada por las coordenadas de los átomos pertenecientes a los aminoácidos del sitio activo (indicados por el Catalytic Site Atlas) y la matriz $\mathbf{Mc_{pr-sa}}$ está compuesta por las conexiones existentes entre los átomos de la matriz $\mathbf{C_{pr-sa}}$.

4.2.1.2 Flexibilidad en el ligando

Teniendo en cuenta que el algoritmo de optimización en un proceso de docking molecular semiflexible, debe encontrar el complejo más estable energéti-

camente, es decir, el complejo que posea la menor energía libre de enlace y una óptima estructura geométrica del ligando, es necesario que el cromosoma del algoritmo genético incluya en sus bits una representación de la flexibilidad de esta molécula. La estructura tridimensional del ligando varía de acuerdo a los enlaces rotables que éste posea. Cada enlace rotable permite dividir la estructura en dos partes rígidas, una de las cuales permanece estática en el espacio (aquella que posee la raíz), mientras que la otra gira θ grados alrededor del enlace. Por lo tanto, el cromosoma diseñado incluye un bit por cada enlace rotable que el ligando posea.

Un ejemplo de lo anterior se ilustra en la figura 7, en la cual se muestra la estructura del ligando correspondiente al caso de prueba 2CTC, en la cual los enlaces rotables se representan con una línea punteada y la raíz se representa por medio de líneas sombreadas. En esta ocasión se escogió realizar una rotación de 180° con respecto al enlace rotable número 1 y una rotación de 0° en los enlaces rotables restantes. Se puede observar que una nueva estructura tridimensional del ligando se puede obtener fácilmente, al generar un cromosoma de forma aleatoria.

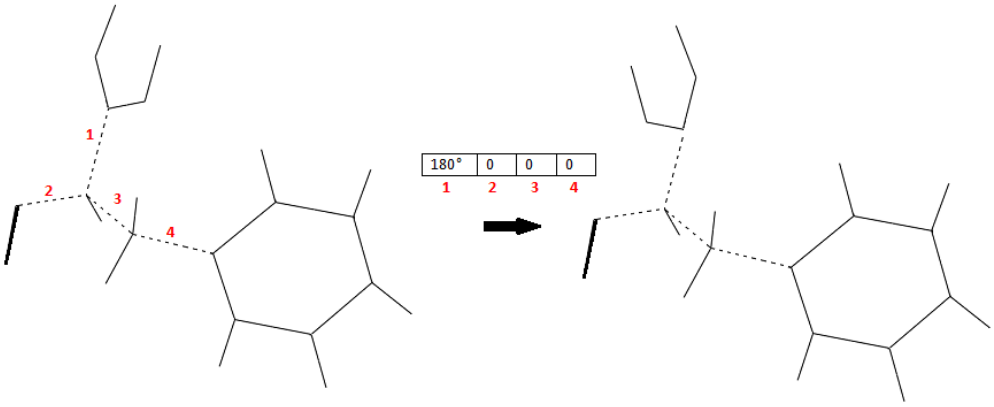


Figura 7: La molécula de la figura posee cuatro enlaces rotables, indicados por las líneas punteadas. Al aplicar las rotaciones indicadas por los bits del cromosoma a cada enlace rotable, se obtiene una nueva estructura tridimensional del ligando. En este caso, sólo se lleva a cabo una rotación de 180° en el enlace rotable número 1.

El valor entero que puede adoptar un bit que indica una rotación con respecto a un enlace rotable, debe pertenecer al conjunto {0,1,2,...,m}, donde m está dado por

$$m = \frac{360}{Mr_e} \tag{6}$$

en donde Mr_e representa la mínima rotación que se puede realizar en un enlace rotable. En este trabajo, Mr_e es conocida como la *resolución de rotación de enlace*.

4.2.1.3 Sitio activo y caja de enlace

En el presente trabajo de investigación, el docking directo se trabajó por medio de una malla tridimensional, llamada caja de enlace, que encierra todos los átomos de los aminoácidos pertenecientes al sitio activo de la proteína, $\mathbf{C_{pr-sa}}$. La distancia entre cada uno de los puntos de la malla tridimensional está dada por d_p y su tamaño final está dado por el número de puntos adicionales, n_{pa}. Estos puntos adicionales, hacen referencia al número de puntos (con espacio d_p entre sí) que se van a adicionar a la mínima caja de enlace posible (aquella que encierra los átomos del conjunto $\mathbf{C_{pr-sa}}$ y que posee un número de puntos n_{cx}, n_{cy} y n_{cz}, con distancia d_p entre sí, en los ejes x, y y z, respectivamente).

Una representación gráfica de la caja de enlace se puede observar en la figura 8.

Debido a que la posición de la estructura del ligando al interior del sitio activo de la proteína debe ser optimizada, se hace necesario incluir tres bits en el cromosoma del algoritmo genético, que indiquen el punto de la malla tridimensional con coordenadas (x,y,z) en el cual se ubicará el centroide del ligando. El valor entero que puede tomar el bit que representa la posición en x, y y z, debe pertenecer al conjunto X_p, Y_p y Z_p, respectivamente.

El conjunto X_p está conformado por los valores enteros {1,2,..., n_x} donde n_x corresponde al número de puntos que posee la caja de enlace en el eje x. De manera similar, los conjuntos Y_p y Z_p comprenden los valores enteros {1,2,...,n_y} y {1,2,...,n_z}, respectivamente.

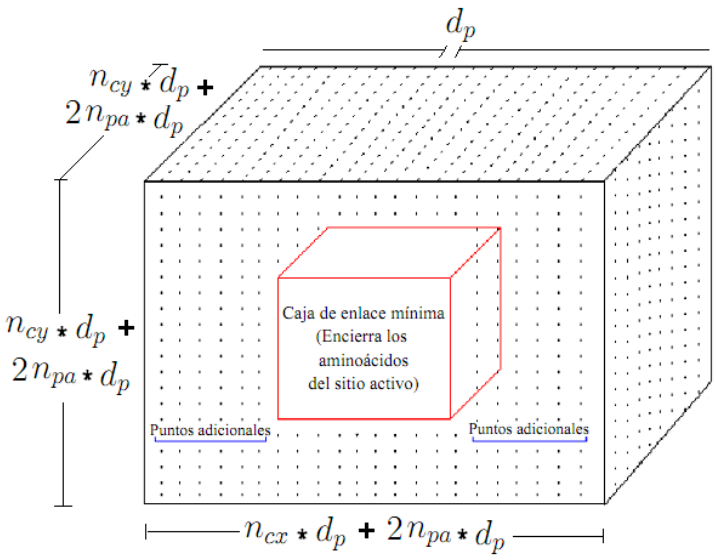


Figura 8: Esquema gráfico de la caja de enlace.

4.2.1.4 Orientación del ligando al interior del sitio activo

La orientación que posea la estructura tridimensional del ligando en un punto específico de la caja de enlace, constituye un nuevo aspecto que se debe tener en cuenta en el proceso de optimización. En este punto, la estructura tridimensional del ligando es considerada rígida y su orientación puede cambiar al rotar todos sus átomos con respecto a los tres ejes espaciales. Por esta razón, se deben incluir tres bits adicionales al cromosoma trabajado, que indiquen las rotaciones de la estructura rígida en el eje x, y y z.

Los valores enteros que pueden asumir estos tres bits, deben estar comprendidos en el conjunto {0,1,...,l}, donde l está dado por

$$l = \frac{360}{Mr_a} \tag{7}$$

en donde Mr_a representa la mínima rotación que se puede realizar en cada uno de los tres ejes de coordenadas.

4.2.1.5 Estructura del cromosoma

El cromosoma diseñado está dividido en tres partes. La primera está compuesta por un número de bits igual al número de enlaces rotables presentes en el ligando. Esta primera parte permite generar una nueva estructura tridimensional de esta molécula. La segunda parte del cromosoma, posee tres bits y hace referencia a la ubicación que va a tener el centroide de la estructura tridimensional del ligando al interior de la caja de enlace. Finalmente, la tercera parte indica la orientación que va a tener la estructura tridimensional del ligando, generada en la primera parte, cuando se ubica en un punto de la caja de enlace, indicado por la parte dos. Esta última parte del cromosoma está compuesta por tres bits. Una representación gráfica del cromosoma diseñado se puede apreciar en la figura 9.

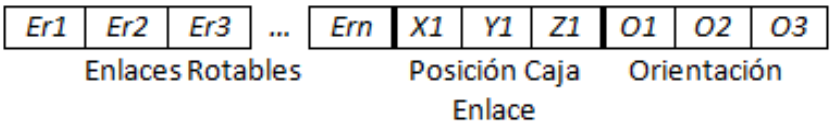


Figura 9: Esquema gráfico del cromosoma diseñado

4.2.2 Generación del Complejo Base

En el proceso computacional desarrollado, un complejo representa la unión de las matrices $\mathbf{C_{pr}}$ y $\mathbf{C_{lig}}$ y la unión de las matrices $\mathbf{Mc_{pr}}$ con $\mathbf{Mc_{lig}}$, lo cual se ve reflejado en la creación de una nueva molécula, en la cual el ligando está unido al sitio activo de la proteína, por medio de enlaces no covalentes.

El método para crear un nuevo complejo consiste en generar un nuevo cromosoma, de forma aleatoria, que indique una nueva configuración para la creación de la estructura tridimensional del complejo, partiendo de las estructuras tridimensionales de la proteína y del ligando, suministradas en el conjunto de prueba.

Teniendo en cuenta que la estructura del ligando suministrada se encuentra optimizada, y asumiendo que las coordenadas de los átomos del ligando indican la posición correcta al interior del sitio activo de la proteína, se hace

necesario crear un complejo base aleatorio que permita iniciar la simulación con una estructura del ligando no optimizada y una ubicación y orientación del ligando no óptima.

La generación del complejo base se realiza al inicio de la simulación y sobre dicha estructura tridimensional, se realiza el proceso de optimización.

4.2.3 Creación de la población inicial

El algoritmo genético de estado estable inicia con una población de tamaño n, definido por el usuario. Dicha población se crea a partir de la generación aleatoria de n cromosomas, en donde los números aleatorios siguen una distribución normal con un periodo de 2^{1492}.

4.2.4 Proceso iterativo

El proceso iterativo del algoritmo genético desarrollado, consta de tres etapas: Evaluación, cruce parcial y mutación. A continuación se describe cada una de ellas.

4.2.4.1 Proceso de evaluación

El proceso de evaluación de los individuos de una población, se realiza mediante el cálculo de la energía proveniente de la estructura interna del ligando y la energía de interacción entre éste y la proteína. La función de energía seleccionada permite evaluar y organizar los individuos de forma satisfactoria, con un tiempo de cómputo de aproximadamente 0.7 segundos por individuo, con ligandos que poseen entre 10 a 15 átomos, en un equipo con procesador AMD Athlon 64x2 5600 Dual Core de 2.9 GHz y memoria de 2GB.

En el proceso de docking molecular, los individuos más aptos son aquellos que poseen una menor energía.

4.2.4.2 Proceso de cruce parcial

El algoritmo genético desarrollado reemplaza un porcentaje de los individuos con el valor de desempeño más bajo (energía más alta), de la población

actual. El número de individuos reemplazados se fija al inicio de la simulación, el cual para este trabajo corresponde al 10 % de la población.

El operador de cruce diseñado permite generar dos nuevos individuos al intercambiar información entre los padres. Para llevar a cabo esta operación es necesario generar tres números aleatorios r_1, r_2 y r_3 que permiten realizar el intercambio de información en la primera, segunda y tercer zona del cromosoma, respectivamente.

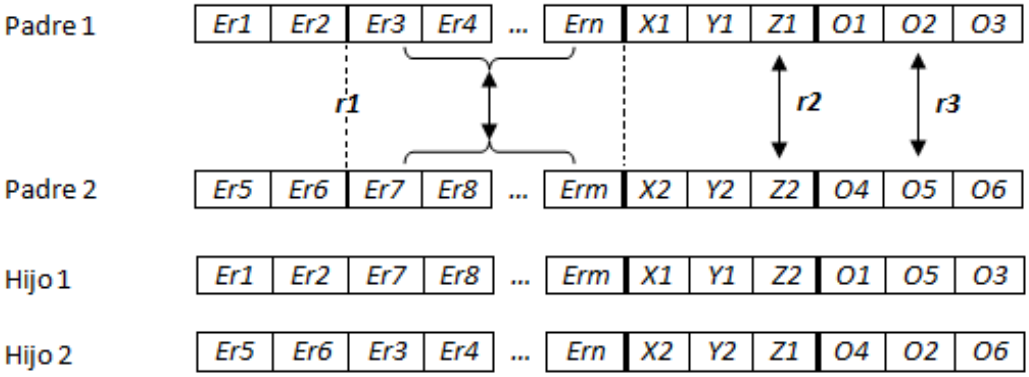


Figura 10: Ejemplo gráfico del operador de cruce: En el ejemplo de la figura, r_1, r_2 y r_3 toman los valores 2, 3 y 2 respectivamente. El valor de r_1 indica que a partir del siguiente bit y hasta el final de la zona 1, los bits deben intercambiarse entre los padres. Los números r_2 y r_3 indican el bit que debe intercambiarse en la zona 2 y 3, respectivamente.

En la primera zona, el número r_1 toma valores enteros que pertenezcan al conjunto {1,2,...,N_e-1} donde N_e es igual al número de enlaces rotables presentes en el ligando. El intercambio en esta parte, se realiza a partir de la posición r_1 hasta la posición N_e.

En la segunda parte del cromosoma, que corresponde a los tres bits de la ubicación del centroide del ligando al interior de la caja de enlace, el cruce se realiza por medio del intercambio del bit r_2 entre los padres. El número entero r_2 pertenece al conjunto {1,2,3}.

Finalmente, en la parte del cromosoma correspondiente a los tres bits que representan la orientación del ligando, el bit que se intercambia entre los padres está dado por r_3, en donde r_3 toma valores del conjunto {1,2,3}. Un

ejemplo gráfico del operador de cruce, se puede apreciar en la figura 10.

4.2.4.3 Proceso de mutación

El operador de mutación trabajado permite realizar pequeñas modificaciones a los nuevos individuos, por esto, una mutación corresponde a la alteración de un sólo bit. Para llevar a cabo este proceso, es necesario generar dos números aleatorios r_4 y r_5, en donde el primer número corresponde a la zona del cromosoma que va a cambiar y el segundo indica la posición del bit a mutar. El número r_4 debe pertenecer al conjunto {1,2,3} y el número r_5 debe pertenecer al conjunto $\{1, 2, ..., N_e\}$ si r_4 es igual a 1 o al conjunto {1,2,3} si r_4 es igual a 2 o 3.

Un ejemplo gráfico del operador de mutación se puede apreciar en la figura 11.

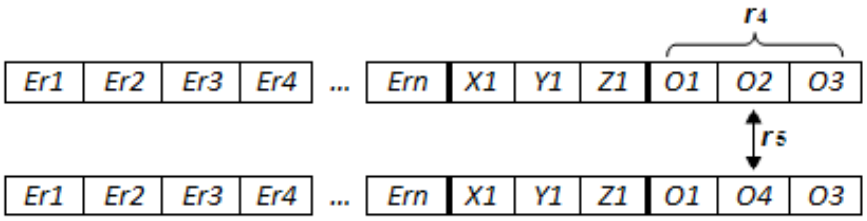


Figura 11: Ejemplo gráfico del operador de mutación. En el ejemplo de la figura,r_4 toma el valor de 3, indicando que la zona en la cual un bit va a ser modificado es la tercera. El valor de r_5, en este caso 2, indica el bit a mutar en la zona previamente seleccionada.

4.2.4.4 Criterio de Parada

El criterio de parada del algoritmo genético desarrollado está dado por el número de generaciones definido al inicio de la simulación, en donde cada nueva generación reemplaza el 10 % de los individuos de la población.

4.2.5 Resultados suministrados por el Algoritmo Genético

El algoritmo desarrollado permite crear los archivos con extensión .pdb de los k mejores complejos obtenidos en el proceso de optimización. El valor de k es indicado al inicio de la simulación. Los archivos generados en esta etapa

del proceso son los que se utilizan en la etapa de validación de resultados.

5. Validación de resultados

La prueba que generalmente se aplica a un algoritmo de docking molecular, es realizar simulaciones con complejos para los cuales su estructura cristalina es conocida [13].

La precisión de la predicción de la estructura de los complejos proteína-ligando, se puede evaluar mediante la desviación cuadrática media (RMSD) existente entre la estructura predicha y la estructura cristalina. El estándar indica que un valor RMSD menor o igual a 2 $\mathring{A}$ es aceptable y un valor por encima de esta cantidad, es considerado una falla [11].

La ecuación para el RMSD es

$$RMSD = \sqrt{\frac{\sum_{i=1}^{N_{at}} d_i^2}{N_{at}}} \tag{8}$$

donde N_{at} es el número de átomos sobre los cuales se va a calcular el RMSD y d_i es la distancia existente entre las coordenadas del átomo i en las dos estructuras, cuando éstas se superponen.

5.1 Simulaciones Realizadas

Las simulaciones realizadas en el presente proyecto, ilustran la relación existente entre el valor de energía, el valor RMSD y el tamaño de la población del algoritmo genético, tomando el número de generaciones constante. Las simulaciones muestran además, la relación entre el valor de energía y el valor RMSD a medida que aumenta el número de generaciones, para una población específica. Las simulaciones realizadas se pueden apreciar en las tablas 4 y 5. Para cada simulación se realizaron 5 corridas del programa y se obtuvo el valor de la energía promedio, el valor RMSD promedio y las

respectivas desviaciones estándar. Los valores para el porcentaje de cruce y porcentaje de mutación se fijaron en 0.8 y 0.5 respectivamente. Las simulaciones fueron realizadas con el complejo Thermolysin cuyo PDB ID es 2TMN, el cual se escogió de forma aleatoria del conjunto de prueba. La distancia entre los puntos de la malla tridimensional se fijó en 0.2 $\mathring{A}$ y el número de puntos adicionales en la caja de enlace se tomó igual a 2 puntos. La mínima rotación en los enlaces rotables fue de 30 grados, mientras que la mínima rotación en cada uno de los ejes espaciales fue de 10 grados.

Simulación No.	**Población**
1	100
2	200
3	300
4	400
5	500

Tabla 4: Simulaciones para el complejo Thermolysin, con 1000 generaciones. Se realizaron 5 corridas del programa para cada valor de la población.

Simulacion No	**No. Generaciones**
6	1000
7	2000
8	3000
9	4000
10	5000
11	6000
12	7000
13	8000
14	9000
15	10000

Tabla 5: Simulaciones para el complejo Thermolysin, con una población de 400 individuos. Se realizaron 5 corridas del programa para cada número de evaluaciones de energía.

Finalmente, se realizaron 5 simulaciones de docking molecular para cada uno de los complejos con PDB id 2CTC, 4TIM, 4TLN y 1CBX. Las simulaciones se realizaron con una población de 400 individuos y 1000 generaciones. Los demás parámetros se fijaron a los mismos valores utilizados en las simulaciones anteriores.

5.2 Resultados y Discusión

El proceso computacional de docking molecular desarrollado genera los archivos con extensión .pdb de los complejos resultantes del proceso de optimización. Un ejemplo gráfico de un complejo de Thermolysin resultante se puede apreciar en la figura 12.

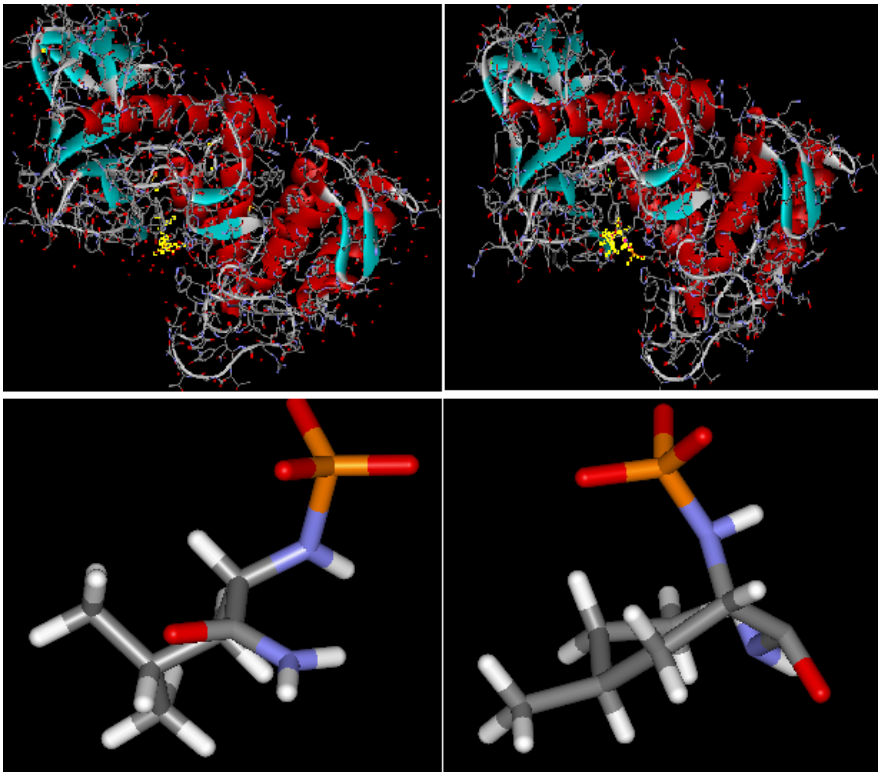


Figura 12: En la figura se puede apreciar el complejo resultante de una simulación de docking molecular para el complejo Thermolysin utilizando el algoritmo genético desarrollado con una población de 400 individuos y 1000 generaciones. En la parte derecha se encuentran el complejo Thermolysin proveniente del Protein Data Bank y su respectivo ligando proveniente del conjunto de prueba. En la parte izquierda se puede apreciar el complejo Thermolysin resultante y la estructura del ligando generada. El valor RMSD entre los complejos presentes en la figura es de 0.9480 $\mathring{A}$.

Los resultados de las simulaciones realizadas se presentan en las tablas 6 y 7, en donde se puede apreciar las variaciones de la energía y del valor RMSD de acuerdo al tamaño de la población y de acuerdo al número de generaciones, respectivamente. Las gráficas de estos resultados se pueden apreciar en las figuras 13, 14 y 15.

Sim. No.	RMSD Prom.	Desv. Est. RMSD	Energía Prom.	Desv. Est. Energía
1	1.5939	0.0816	492.6811	15.1771
2	1.5018	0.0780	481.4335	9.3229
3	1.3970	0.1009	405.2007	9.4517
4	1.4874	0.1395	391.7063	4.6501
5	1.3990	0.1241	395.0098	3.5718

Tabla 6: Resultados para las simulaciones realizadas con el complejo Thermolysin manteniendo constante el número de generaciones en 1000 y variando el tamaño de la población

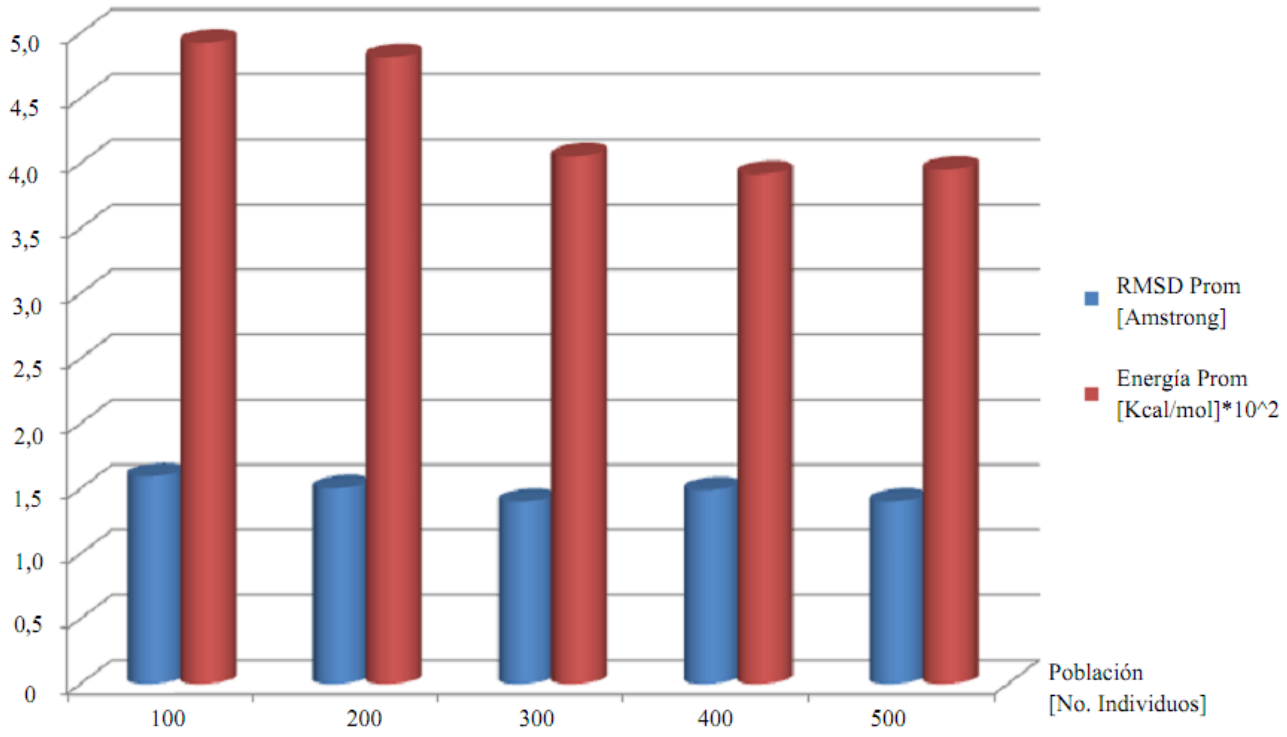


Figura 13: La gráfica muestra los resultados del valor RMSD promedio y la energía promedio de 5 simulaciones de docking molecular a medida que aumenta el tamaño de la población. El número de generaciones se fija en 1000.

En los resultados de la tabla 6 y en su correspondiente gráfica (figura 13), se puede apreciar que un incremento en la población no repercute de manera significativa en el valor de la energía promedio obtenido en las simulaciones. Sin embargo, se escoge trabajar en las simulaciones posteriores con una población de 400 individuos, debido a que fue la población para la cual se obtuvo los menores valores de energía. En esta misma gráfica también se puede apreciar que una disminución en el valor de la energía no implica una disminución en el valor RMSD, así como tampoco un aumento de la energía genera un aumento en el valor RMSD. Esta situación se presenta debido a la falta de precisión en la función de evaluación de energía, que ignora múltiples aspectos bioquímicos que marcan la diferencia entre un complejo muy similar (con un valor de RMSD menor a 2 $\mathring{A}$) con un complejo casi perfecto (con un valor de RMSD muy cercano a 0 $\mathring{A}$).

Sim. No.	**RMSD Prom.**	**Desv. Est. RMSD**	**Energía Prom.**	**Desv. Est. Energía**
6	1.4100	0.0189	432.8311	2.8493
7	1.3934	0.0070	429.8902	2.3632
8	1.3934	0.0070	429.8902	2.3632
9	1.3934	0.0070	429.8902	2.3632
10	1.3934	0.0070	429.8902	2.3632
11	1.3942	0.0076	429.6773	2.3374
12	1.3942	0.0076	429.6773	2.3374
13	1.3942	0.0076	429.6773	2.3374
14	1.3942	0.0076	429.6773	2.3374
15	1.3942	0.0076	429.6773	2.3374

Tabla 7: Resultados para las simulaciones realizadas con el complejo Thermolysin a medida que aumenta el número de generaciones. El tamaño de la población para estas simulaciones se fijo en 400 individuos.

En la tabla 7 y en su correspondiente gráfica (figura 14), se puede observar que a medida que avanza la simulación, el valor de la energía disminuye, debido a que la tendencia del algoritmo genético desarrollado es permanecer estable hasta que encuentre una solución mejor, es decir, un complejo con una menor energía. Por otra parte, se puede apreciar que la gráfica del valor RMSD, en la figura 15, no presenta una tendencia de disminución a medida que avanza la simulación del proceso de docking molecular. Esta situación se presenta debido a que el parámetro de optimización del algoritmo genético es la energía y no el valor RMSD. Adicionalmente, se puede observar que si

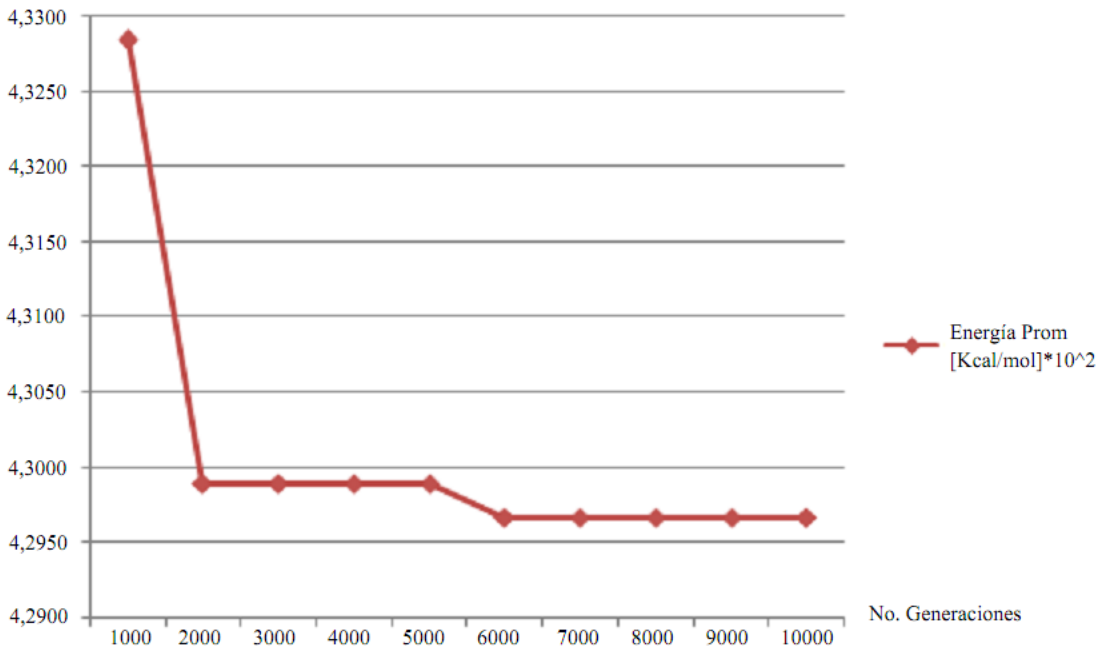


Figura 14: La gráfica muestra los resultados de la energía promedio a medida que avanza el número de generaciones en una simulación de docking molecular. El número de individuos en la población es de 400.

lo que se desea es minimizar el valor RMSD, la función de energía a utilizar debería tener una relación directa con este valor, de tal forma que la minimización de la energía conduzca a la minimización del valor RMSD. De los resultados obtenidos se aprecia que la función de energía utilizada no presenta una relación directa con el valor RMSD y por lo tanto, la tendencia de este último valor no se puede determinar de forma precisa. Sin embargo, es importante resaltar que los valores de RMSD obtenidos en las simulaciones son satisfactorios teniendo como referencia los valores presentes en la literatura, los cuales plantean que un resultado de docking molecular que presente un valor de RMSD menor a 2 $\mathring{A}$ es considerado un éxito [11].

Finalmente, los resultados obtenidos en las simulaciones realizadas con los complejos con PDB ID 2CTC, 4TIM, 4TLN y 1CBX son de una calidad aceptable, considerando que los valores más altos obtenidos no superan los 3.6 $\mathring{A}$ y los resultados más bajos están muy cercanos a 2 $\mathring{A}$. Los resultados se pueden apreciar en la tabla 8.

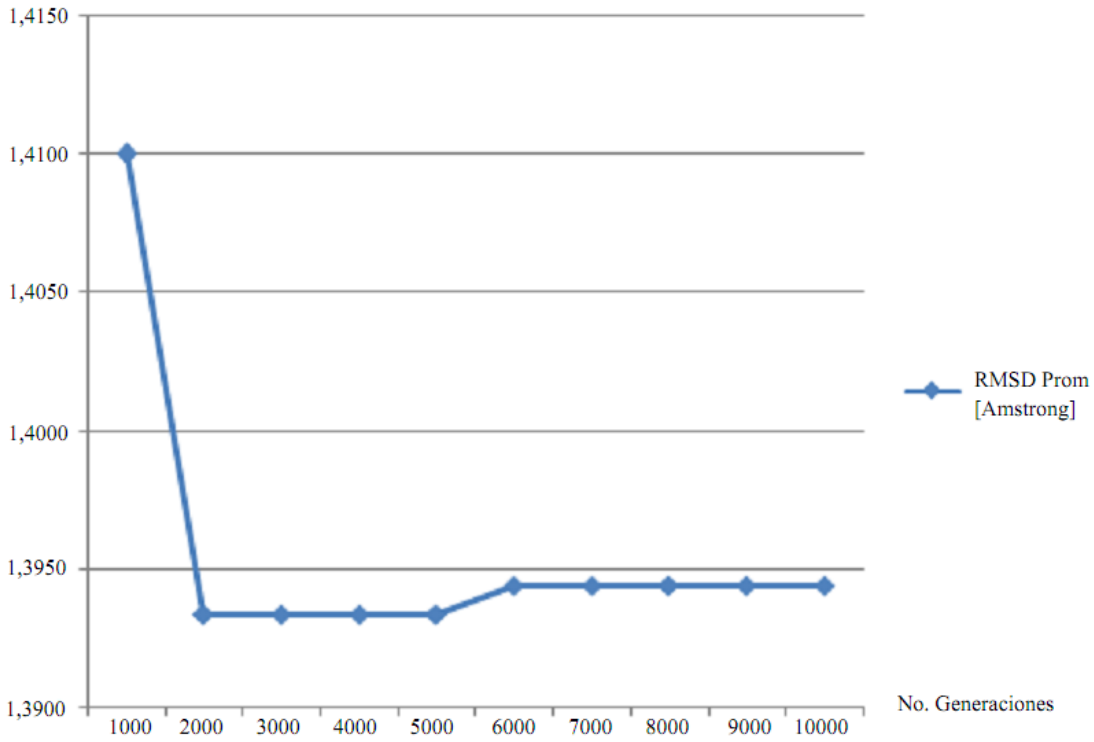


Figura 15: La gráfica muestra los resultados del valor RMSD promedio a medida que avanza el número de generaciones en una simulación de docking molecular. El número de individuos en la población es de 400.

PDB ID	RMSD Prom.	Desv. Est. RMSD	Energía Prom.	Desv. Est. Energía
2CTC	2.8349	0.0470	476.3569	15.3386
4TIM	3.5308	0.0578	233.6109	16.5723
4TLN	2.3156	0.1818	204.5225	36.9741
1CBX	3.3019	0.1621	610.5378	11.6276

Tabla 8: Resultados para las simulaciones de docking molecular realizadas con los complejos 2CTC, 4TIM, 4TLN y 1CBX, utilizando el algoritmo genético desarrollado con una población de 400 individuos y 1000 generaciones.

6. Conclusiones y Trabajos Futuros

A continuación se presentan las conclusiones más relevantes del presente proyecto, así como las recomendaciones para trabajos futuros.

6.1 Conclusiones

La metodología propuesta en el presente trabajo de investigación permite realizar con éxito un proceso de docking molecular directo y semiflexible entre una proteína y un ligando. Los pasos y procedimientos propuestos permitirían ahorrar tiempo y esfuerzo a aquellas personas que deseen iniciar labores de investigación en el campo de docking molecular, disminuyendo el tiempo dedicado a la adecuación y pre-procesamiento de los datos, logrando así, una mayor dedicación a los procesos de optimización requeridos.

El algoritmo genético diseñado en el presente proyecto, permite realizar simulaciones de docking molecular directo y semiflexible, puesto que permite minimizar el valor de la energía a medida que el número de generaciones aumenta. Adicionalmente, el cruce parcial del algoritmo diseñado permite reducir el tiempo de cómputo requerido debido a que sólo un porcentaje de la población es reemplazado por nuevos individuos provenientes de las operaciones de cruce y mutación.

Teniendo en cuenta los resultados obtenidos en las simulaciones de docking molecular realizadas, se puede concluir que dada la imposibilidad de utilizar como criterio de minimización el valor RMSD, debido a que para nuevos medicamentos no existe una estructura cristalina conocida, es de gran importancia que la función de energía presente una relación directa con el valor RMSD, de tal forma que una minimización de la energía conduzca necesariamente a una minimización en el valor RMSD.

La selección de la función de evaluación de energía es una de las etapas más importantes en el desarrollo de un proceso de docking molecular. Por esta razón, es importante tener presente que si se desea desarrollar un proceso de docking molecular que realice las simulaciones en un corto tiempo computacional, se debe escoger una función de evaluación de energía poco detallada sacrificando así precisión en los resultados obtenidos. Por el contrario, si se desea un proceso de docking molecular con una alta precisión en los valores RMSD, se debe seleccionar una función de energía que involucre diferentes aspectos bioquímicos y por lo tanto, el tiempo de cómputo requerido será de varios ordenes mayor que en el caso anterior. En el presente proyecto se optó por realizar simulaciones de docking molecular en un menor tiempo computacional.

6.2 Trabajos Futuros

En el presente trabajo de investigación se utilizó como función de evaluación de la energía, aquella suministrada por el campo de fuerza AMBER. En futuros trabajos, se pueden utilizar funciones que incluyan nuevos aspectos bioquímicos, así como también se puede trabajar con evaluaciones en consenso entre diferentes funciones de energía.

Entre las diferentes modificaciones que se pueden realizar al prototipo de docking molecular desarrollado se encuentran las siguientes: agregar las moléculas de agua que pertenecen al sitio activo de la proteína, encerrar los aminoácidos del sitio activo en una malla en forma de esfera, hacer una reconstrucción incremental del ligando flexible, realizar un proceso de docking rígido con diferentes configuraciones de un mismo ligando, utilizar métodos de optimización híbridos.

Un futuro trabajo debe contemplar la paralelización del algoritmo de docking molecular desarrollado, de manera que permita incrementar la precisión de la función de evaluación de energía sin aumentar de forma significativa el tiempo de cómputo requerido para llevar a cabo la simulación.

Bibliografía

[1] Cardellá L, Hernández R. *Bioquímica Médica. Tomo I Biomoléculas*, volume I. Editorial Ciencias Médicas, La Habana, 1999.

[2] Nelson D, Cox M. *Principles in Biochemistry.* W. H. Freeman, 4 edition, April 2004.

[3] Xu J, Stevenson J. Drug-like index: A new approach to measure drug-like compounds and their diversity. *Journal of Chemical Information and Computer Sciences*, 40(5):1177–1187, 2000.

[4] Lipinski C, Lombardo F, Dominy B, Feeney P. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, 23(1-3):3 – 25, 1997. In Vitro Models for Selection of Development Candidates.

[5] Muegge I, Heald S, Brittelli D. Simple selection criteria for drug-like chemical matter. *Journal of Medicinal Chemistry*, 44(12):1841–1846, Jun 2001.

[6] Larson R, editor. *Bioinformatics and Drug Discovery.* Humana Press, 2005.

[7] Hltje H, Sippl W, Rognan D, Folkers G. *Molecular Modeling: Basic Principles and Applications.* Wiley VCH, 2 edition, 2003.

[8] Xiao Y, Williams D. Molecular docking using genetic algorithms. In *SAC '94: Proceedings of the 1994 ACM symposium on Applied computing*, pages 196–200, New York, NY, USA, 1994. ACM.

[9] Yang J, Kao C. A family competition evolutionary algorithm for automated docking of flexible ligands to proteins. *Information Technology in Biomedicine, IEEE Transactions on*, 4(3):225–237, Sept. 2000.

[10] Amaro R, Baron R, McCammon J. An improved relaxed complex scheme for receptor flexibility in computer-aided drug design. *Journal of Computer-Aided Molecular Design*, 22(9):693–705, Sep 2008.

[11] Brooijmans N, Kuntz I. Molecular recognition and docking algorithms. *Annual Review of Biophysics and Biomolecular Structure*, 32:335–373, Jan 2003.

[12] Leach A. *Molecular Modelling. Principles and applications.* Prentice Hall, 2001.

[13] Lengauer T, editor. *Bioinformatics- from genomes to drugs. Vol 1: Basic technologies. Vol 2: Applications. (Methods and principles in medicinal chemistry Vol. 14).* Wiley-VCH, 2002.

[14] Xiao Y, Williams D. A comparison of GA and RSNR docking. In *Evolutionary Computation, 1994. IEEE World Congress on Computational Intelligence., Proceedings of the First IEEE Conference on*, pages 802–806 vol.2, Jun 1994.

[15] Westhead D, Clark D, Murray C. A comparison of heuristic search algorithms for molecular docking. *Journal of Computer-Aided Molecular Design*, 11(3):209–228, 1997.

[16] Engelbrecht A. *Computational Intelligence.* An Introduction. Wiley, 2003.

[17] Halperin I, Ma B, Wolfson H, Nussinov R. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins*, 47(4):409–443, June 2002.

[18] Cao T, Li T. A combination of numeric genetic algorithm and tabu search can be applied to molecular docking. *Computational Biology and Chemistry*, 28(4):303 – 312, 2004.

[19] Sung W. Employing improved GA to promote molecular docking efficiency for drug design. In *Bioinformatics and Biomedical Engineering, 2008. ICBBE 2008. The 2nd International Conference on*, pages 37–40, May 2008.

[20] Zsoldos Z, Szabo I, Szabo Z, Johnson A. Software tools for structure based rational drug design. *Journal of Molecular Structure: THEOCHEM*, 666-667:659 – 665, 2003.

[21] Porter C, Bartlett G, Thornton J. The catalytic site atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Research*, 32(suppl_1):D129–133, January 2004.

[22] Cornell W, Cieplak P, Bayly C, Gould I, Merz K, Ferguson D, Spellmeyer D, Fox T, Caldwell J, Kollman P. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society*, 117(19):5179–5197, 1995.

[23] Bursulaya B, Totrov M, Abagyan R, Brooks C. Comparative study of several algorithms for flexible ligand docking. *Journal of Computer-Aided Molecular Design*, 17:755–763, 2003.

[24] Dreo J, Pétrowski A, Siarry P, Taillard E. *Metaheuristics for Hard Optimization: Methods and Case Studies*. Springer, December 2005.

[25] Jones G, Willett P, Glen R. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *Journal of Molecular Biology*, 245(1):43 – 53, 1995.

[26] Jones G, Willett P, Glen R, Leach A, Taylor R. Development and validation of a genetic algorithm for flexible docking. *Journal of Molecular Biology*, 267(3):727 – 748, 1997.

[27] Magalhes C, Barbosa H, Dardenne L. A genetic algorithm for the ligand-protein docking problem. *Genetics and Molecular Biology*, 27(4): 605–610, 2004.

[28] Schuler L, Daura X, Gunsteren W. An improved GROMOS96 force field for aliphatic hydrocarbons in the condensed phase. *Journal of Computational Chemistry*, 22(11):1205–1218, 2001.

[29] Liu B, Chen H, Huang H, Hwang S, Ho S. Flexible protein-ligand docking using particle swarm optimization. volume 1, pages 251–258 Vol.1, Sept. 2005.

[30] Oduguwa A, Tiwari A, Fiorentino S, Roy R. Multi-objective optimisation of the protein-ligand docking problem in drug discovery. In *GECCO '06: Proceedings of the 8th annual conference on Genetic and evolutionary computation*, pages 1793–1800, New York, NY, USA, 2006. ACM.

[31] Chen K, Li T, Cao T. Tribe-pso: A novel global optimization algorithm and its application in molecular docking. *Chemometrics and Intelligent Laboratory Systems*, 82(1-2):248 – 259, 2006.

[32] Wang L, Weng Z, Liang Y, Wang Y, Zhang Z, Di R. Design and implementation of parallel lamarckian genetic algorithm for automated docking of molecules. *High Performance Computing and Communications, 10th IEEE International Conference on*, pages 689–694, Sept. 2008.

[33] Kang L, Li H, Jiang H, Wang X. An improved adaptive genetic algorithm for protein-ligand docking. *Journal of Computer Aided Molecular Design*, 23(1):1–12, Jan 2009.

[34] Thiriot E, Monard G. Combining a genetic algorithm with a linear scaling semiempirical method for protein-ligand docking. *Journal of Molecular Structure: THEOCHEM*, 898(1-3):31 – 41, 2009.

[35] Software autodock. Disponible en: http://autodock.scripps.edu/. Fecha de consulta: Mayo de 2009.

[36] Software dockvision. Disponible en: http://dockvision.com/. Fecha de consulta: Junio de 2009.

[37] Verdonk M, Cole J, Hartshorn M, Murray C, Taylor R. Improved proteinligand docking using GOLD. *PROTEINS*, 52:609–623, 2003.

[38] Software dock. Disponible: http://dock.compbio.ucsf.edu/. Fecha de consulta: Mayo de 2009.

[39] Software flexx. Disponible: http://www.biosolveit.de/flexx/. Fecha de consulta: Junio de 2009.

[40] Clark M, Cramer R, Opdenbosch N. Validation of the general purpose tripos 5.2 force field. *Journal of Computational Chemistry*, 10(8):982–1012, September 1989.

[41] Wang R, Lu Y, Wang Shaomeng. Comparative evaluation of 11 scoring functions for molecular docking. *Journal of Medicinal Chemistry*, 46(12):2287–2303, 2003.

[42] Cheng T, Li X, Li Y, Liu Z, Wang R. Comparative assessment of scoring functions on a diverse test set. *Journal of Chemical Information and Modeling*, 49(4):1079–1093, 2009.

[43] Prez C, Ortiz A. Evaluation of docking functions for protein-ligand docking. *Journal of Medicinal Chemistry*, 44:3768–3785, 2001.

[44] Ferrara P, Gohlke H, Price D, Klebe G, Brooks C. Assessing scoring functions for proteinligand interactions. *Journal of Medicinal Chemistry*, 47(12):3032–3047, 2004.

[45] Wang R, Lu Y, Fang X, Wang S. An extensive test of 14 scoring functions using the pdbbind refined set of 800 proteinligand complexes. *Journal of Chemical Information and Computer Sciences*, 44(6):2114–2125, 2004.

[46] Gilson M, Zhou H. Calculation of protein-ligand binding affinities*. *Annual Review of Biophysics and Biomolecular Structure*, 36(1):21–42, 2007.

[47] Wang R, Fang X, Lu Y, Yang C, Wang S. The PDBbind database: methodologies and updates. *Journal of medicinal chemistry*, 48(12):4111–4119, June 2005.

[48] Porter C, Bartlett G, Thornton J. The catalytic site atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Research*, 32(suppl_1):D129–133, January 2004.

[49] Berman H, Westbrook J, Feng Z, Gilliland G, Bhat T, Weissig H, Shindyalov I, Bourne P. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, January 2000.

[50] Software openbabel. Disponible: http://openbabel.org/wiki/Install. Fecha de consulta: Junio de 2009.

Anexos

A. Algoritmo Genético desarrollado

Variable	Descripción
A_p	Archivo de la Proteína
A_l	Archivo del ligando
A_c	Archivo del complejo
$NumG$	Número de generaciones
C_i	Complejo i-ésimo
Ec_i	Energía del complejo i-ésimo
E	Vector que almacena la energía de cada complejo
P	Población (Estructura que almacena cada uno de los complejos)
m	Tamaño de la población
P_r	Porcentaje de individuos a reemplazar
P_c	Porcentaje de cruce
P_m	Porcentaje de mutacion
H	Complejos Hijos en el cruce
H'	Complejos Hijos mutados

Tabla A.1: Variables del Algoritmo

La función *Energía* recibe como parámetro la población inicial y retorna un vector con los valores de energía de cada uno de los individuos.

La función *Cruce* recibe como parámetros la población actual y el porcentaje de individuos a reemplazar. Esta función retorna el número de hijos correspondiente al número de individuos a reemplazar.

La función *Mutacion* recibe como parámetros los hijos provenientes del cruce y el porcentaje de mutación y retorna los hijos mutados y no mutados. La mutación se da con el porcentaje de mutación indicado.

La función *Reemplazo* recibe como parámetros la población actual y los

hijos que van a reemplazar a los peores individuos de la presente población y retorna la nueva población resultante.

Algoritmo 1 Algoritmo Genético Docking Molecular

Entradas A_p, A_l
Salidas A_c
Crear población inicial $P = \{C_1, C_2, ..., C_m\}$
$n = 0$
mientras $n < NumG$ **Haga**
Evaluar energía de la población
$E[Ec_1, Ec_2, ..., Ec_m] = Energia(P)$
$H = Cruce(P, P_r, P_c)$
$H' = Mutacion(H, P_m)$
$P = Reemplazo(P, H')$

Fin mientras
Crear archivo del mejor individuo
$A_c = CrearArchivo(P(1))$

B. Cálculo de la energía de un complejo

El cálculo de la energía de los complejos generados en las simulaciones de docking molecular semiflexible, se obtiene mediante la suma de la energía interna del ligando y la energía de interacción entre éste y la proteína.

La función de evaluación de energía que permite obtener los anteriores valores, es aquella suministrada por el campo de fuerza AMBER [22], la cual está dada por:

$$E_{total} = \sum_{enlaces} K_r(r - r_{eq})^2 + \sum_{angulos} K_\theta(\theta - \theta_{eq})^2$$
$$+ \sum_{dihedros} \frac{V_n}{2}\left[1 + cos\left(n\phi - \gamma\right)\right] + \sum_{i<j}\left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\epsilon R_{ij}}\right]$$

donde las variables K_r, r_{eq}, K_θ, θ_{eq}, V_n, n, γ, A_{ij}, B_{ij} y ϵ son los parámetros suministrados por el campo de fuerza AMBER y las variables r , θ , ϕ, R_{ij} y q representan la longitud de un determinado enlace covalente, el ángulo existente entre dos enlaces covalentes, el valor de un determinado ángulo dihedro y la carga de un determinado átomo, respectivamente.

Finalmente, teniendo en cuenta que los complejos trabajados en el presente proyecto de investigación no presentan enlaces covalentes entre la proteína y el ligando, para el cálculo de la energía de interacción entre estas dos moléculas, sólo se utilizan los términos correspondientes a la energía de van der Waals y a la energía de las interacciones electrostáticas.