HERRAMIENTA ORIENTADA A LA WEB BASADA EN UN ALGORITMO DE OPTIMIZACIÓN DE COLONIA DE HORMIGAS (OCH) PARA APROXIMAR EL PLEGAMIENTO DE UNA PROTEÍNA EN DOS DIMENSIONES (2D).

Autores:

Álvaro Andrés Obregón Carreño
John Fredy Gómez Rojas

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FÍSICO-MECÁNICAS
ESCUELA DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
BUCARAMANGA
2011

HERRAMIENTA ORIENTADA A LA WEB BASADA EN UN ALGORITMO DE OPTIMIZACIÓN DE COLONIA DE HORMIGAS (OCH) PARA APROXIMAR EL PLEGAMIENTO DE UNA PROTEÍNA EN DOS DIMENSIONES (2D).

Director:

Bsc. Alfonso Mendoza Castellanos

Codirectores:

Msc. Darío José Delgado Quintero **Phd. Rodrigo Gonzalo Torres Sáez**

Autores:

Álvaro Andrés Obregón Carreño
John Fredy Gómez Rojas

Tesis de grado presentada como requisito para optar al título de:

Ingeniero de Sistemas

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FÍSICO-MECÁNICAS
ESCUELA DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
BUCARAMANGA
2011

Dedicatoria

Agradecido con Dios quiero dedicar este proyecto a:

"Mis padres Álvaro e Hilda cuyos esfuerzos y alientos me impulsaron
a enfrentar esta etapa de mi vida, de igual manera a los familiares que con
su apoyo sin importar la distancia me ayudaron a continuar;
de igual manera dedico este triunfo a mis amigos que
me acompañaron en el proceso dando me ánimo
A todos les dedico este triunfo."

Álvaro Andrés

Dedicado a:

"Dios, por las oportunidades que me brinda cada día.
Mis padres Héctor Julio Gómez Caicedo y Carmen Tulia Rojas,
por su apoyo, comprensión y dedicación.
Mis hermanos, Alexander, Carmen Judit, Héctor Andrés y Edilma,
por las experiencias inolvidables compartidas.
Mis amigos, por su compañia y consejos.
Mis profesores, por haberme compartido más que conocimiento.
Todas aquellas personas que han influido en el transcurso de mi vida."

John Fredy

Agradecimientos

Agradezco a mis padres por su apoyo y valores infundidos en el hogar.
A Dios por permitirnos culminar todas las metas hasta el momento.
Al **Bsc. Alfonso Mendoza Castellanos**, por su dirección en este trabajo.
Al magister en informática **Darío José Delgado Quintero**, pos su guía, consejos y apoyo en este trabajo.
Al **Ing. Juan Carlos Escobar Ramírez**, por su gran colaboración.
A la **Universidad Industrial de Santander** y a la **Escuela de Ingeniería de Sistemas e Informática**, por la formación profesional recibida.
Al **Grupo de Investigación en Informática Biomédica (GIIB)** por brindarme el apoyo y la oportunidad de de realizar proyectos de este tipo.
A mis amigos que me acompaÃśaron en esta etapa.
Un agradecimiento especial a mi hermano Iván Felipe, compañero de mil batallas, en las cuales juntos siempre hemos salido victoriosos.

Álvaro Andrés

"Principalmente a Dios, mis padres, hermanos, amigos y profesores, por su aporte en mi formación académica y personal."

Además quiero agradecer de manera especial:

Al **Msc. Darío José Delgado Quintero**, por la paciencia y entrega al proyecto.
Al **Bsc. Alfonso Mendoza Castellanos**, por el apoyo incondicional.
Al **Grupo de Investigación en Informática Biomédica**, por el respaldo brindado.
Al **Ing. Juan Carlos Escobar Ramírez**, por su gran colaboración.
A la **Universidad Industrial de Santander**, por la oportunidad ofrecida.
A los compañeros de carrera por hacer parte de cada logro alcanzado.

John Fredy

Resumen

Título:

Herramienta orientada a la web basada en un algoritmo de Optimización de Colonia de Hormigas (OCH) para aproximar el plegamiento de una proteína en dos dimensiones (2D) [*]

Autores:

Álvaro Andrés Obregón Carreño
John Fredy Gómez Rojas [**]

Palabras Clave:

Optimización de Colonia de Hormigas, Plegamiento de Proteínas, Modelo Hidrofóbico-Polar, Proteína, Modelo HP, OCH, ACO , algoritmo de hormigas, optimización

Descripción:

Desde hace varias décadas se han venido utilizando técnicas computacionales para aproximar la conformación de una proteína a partir de la secuencia de aminoácidos que la compone, donde los resultados obtenidos no son tan precisos como si lo son las técnicas experimentales, sin embargo las técnicas computacionales si ofrecen aproximaciones útiles a menores costos.

Frecuentemente este problema de optimización es estudiado a partir de modelos simplificados, extensamente usados para el estudio de enfoques algorítmicos del problema de la aproximación de la estructura protéica siendo este un problema computacionalmente difícil, prueba de ello es su grado de complejidad.

Dado el planteamiento anterior los métodos de optimización heurísticos se perfilan como los más promisorios enfoques para abordar el problema donde se modela la energía libre de una cadena de aminoácidos dada y a partir de allí encontrar aquellas estructuras que minimicen dicha energía.

En este trabajo se adaptada el algoritmo de optimización de colonia de hormigas a la problemática del plegamiento proteico basados en el modelo Hidrofóbico-Polar en 2D, donde se definieron las características principales presentes en el proceso de predicción de la estructura secundaria de las proteínas planteando este problema en términos de un problema de optimización.

[*] *Trabajo de Investigación.*

[**] *Facultad de Ingenierías Físico-mecánicas. Ingeniería de Sistemas. Director: Bsc Alfonso Mendoza Castellanos. Codirectores: Phd. Rodrigo Gonzalo Torres Sáez, Msc. Darío José Delgado Quintero.*

Además se desarrollo una herramienta con interfaz web permitiendo la interacción y visualización de los resultados generados por el algoritmo implementado, esta herramienta se implanto en un servidor, facilitando el acceso a esta a través de redes locales e internet.

Abstract

Title:

Tool oriented to the web based on an Ants Colony Optimization (ACO) algorithm to approach the fold of a protein in two dimensions (2D) [*]

Authors:

Álvaro Andrés Obregón Carreño
John Fredy Gómez Rojas [**]

Key Words:

Ants Colony Optimization, Protein Folding, Hydrophobic-Polar Model, ACO, HP model.

Description:

For several decades, computational techniques have been used to approach the conformation of a protein from the sequence of amino acids that composes it, where the results are not as precise as if they are experimental techniques, nevertheless computational techniques if provide useful approximations to lower costs.

Often this optimization problem is studied based on simplified models, widely used for the study of algorithmic approaches to the problem of approximation of protein structure, this being a computationally difficult problem, the proof is its complexity.

Given the above approach heuristic optimization methods are emerging as the most promising approaches to the problem which models the free energy of a given amino acid chain and from there find the structures that minimize this energy.

In this work we adapted the algorithm of ant colony optimization to the problem of protein folding, model-based 2D Hydrophobic-Polar, where defined the main features present in the process of predicting the protein secondary structure posed this problem in terms of an optimization problem.

In addition we developed a tool with a web interface allowing interaction and visualization of results generated by the algorithm, this tool is implanted on a server, providing access to this through local networks and internet.

[*] *Research Proyect.*

[**] *Faculty of Physical-Mechanical Engineering. Systems Engineering. Director: Bsc Alfonso Mendoza Castellanos. Codirectores: Phd. Rodrigo Gonzalo Torres Sáez, Msc. Darío José Delgado Quintero.*

Glosario

- **Aminoácido:** Sustancia química orgánica en cuya composición molecular posee un grupo amino y otro carboxilo y 20 de tales sustancias son los componentes fundamentales de las proteínas.
- **Polipetído:** Nombre utilizado para designar un péptido de tamaño suficientemente grande, se puede hablar de más de 10 aminoácidos.
- **Enlace covalente:** Se produce por compartición de electrones entre dos o mas átomos.
- **Enlace peptídico:** Enlace covalente entre el grupo amino de un aminoácido y el grupo carboxilo de otro aminoácido.
- **Estructura**

  primaria: Forma de organización más básica de las proteínas. Está determinada por la secuencia de aminoácidos de la cadena proteíca, es decir, el número de aminoácidos presentes y el orden en que están enlazados por medio de enlaces peptídicos.

  secundaria: Plegamiento regular local entre residuos aminoacídicos cercanos de la cadena polipeptídica. Se adopta gracias a la formación de enlaces de hidrógeno entre las cadenas laterales (radicales) de aminoácidos cercanos en la cadena.

  terciaria: Modo en el que la cadena polipeptídica se pliega en el espacio.
- **Péptido:** Sustancia orgánica, formado de moléculas estructuralmente similares a las de las proteínas, aunque más pequeñas y más livianas.).
- **Proteína:** Macromoléculas formadas por cadenas lineales de aminoácidos.

Índice general

Índice de cuadros

Índice de figuras

1. Introducción

1.1. Aminoácidos y proteínas

Los aminoácidos son las unidades estructurales básicas de las proteínas, la unión[1][14] de estos integran una secuencia polimérica[12]. Resultados del análisis realizado a un gran número de proteínas de diversas fuentes han mostrado que todas estas están compuestas de 20 aminoácidos estándar diferentes[6] (Cuadro 1).

1.2. Niveles estructurales de las proteínas

En su estado natural o estado nativo, cada tipo de molécula tiene una estructura característica que determina en gran medida las propiedades de la proteína[6]. Esta estructura posee una distribución espacial conocida como la conformación de la proteína, es decir, la forma como los polipéptidos se pliegan en el espacio[12].

Nombres y abreviaturas de los principales aminoácidos			
Número	**Nombre**	**Una letra**	**Tres letras**
1	Alanina	A	Ala
2	Cisteína	C	Sys
3	Ácido Aspartico	D	Asp
4	Ácido Glutámico	E	Glu
5	Fenilalanina	F	Phe
6	Glicina	G	Gly
7	Histidina	H	His
8	Isoleucina	I	Ile
9	Lisina	K	Lys
10	Leucina	L	Leu
11	Metionina	M	Met
12	Asparagina	N	Asn
13	Prolina	P	Pro
14	Glutamina	Q	Gln
15	Arginina	R	Arg
16	Serina	S	Ser
17	Treonina	T	Thr
18	Valina	V	Val
19	Triptófano	W	Trp
20	Tirosina	Y	Tyr

Cuadro 1: *Tabla extraída de Current Protocols in Protein Science[5].*

[1] *En una proteína los aminoácidos pueden combinarse en cualquier orden y pueden repetirse de cualquier manera, lo cual determina una secuencia específica.*

Estos 20 aminoácidos se encuentran comúnmente en las proteínas unidos por enlaces peptídicos conformando así secuencias lineales, las cuales contiene la información necesaria para generar una molécula proteica con una estructura particular. La complejidad de una estructura proteica se puede analizar de manera simpliïňĄcada si se toman en cuenta 4 niveles fundamentales de organizaci ón en las macromoléculas, los cuales se denominan: estructura primaria, secundaria, terciaria y cuaternaria.

El primer nivel estructural que se puede delimitar en una proteína, está constituido por el número y la variedad de aminoácidos que entran en su composición, como por el orden (llamado también secuencia) que se disponen estos a lo largo de la cadena polipéptidica.

El segundo nivel estructural llamado estructura secundaria[2][12] (Figura 1[3]) se refiere a la relación espacial que guarda un aminoácido respecto al anterior y al siguiente en la cadena polipéptidica. En algunos casos el polipéptido o algunas zonas de este se mantienen extendidas (Láminas-β), mientras que en otros casos se pliegan en forma helicoidal (Hélices-α).

El tercer nivel estructural llamado también estructura terciaria[13] (Figura 2[4]) se refiere a la relación espacial que guardan entre si las diferentes zonas o áreas de cada cadena polipéptidica que forman una proteína.
Cuando una proteína tiene más de una cadena polipéptidica, es posible que interactúen entre ellas, lo cual determina la estructura cuaternaria[8].

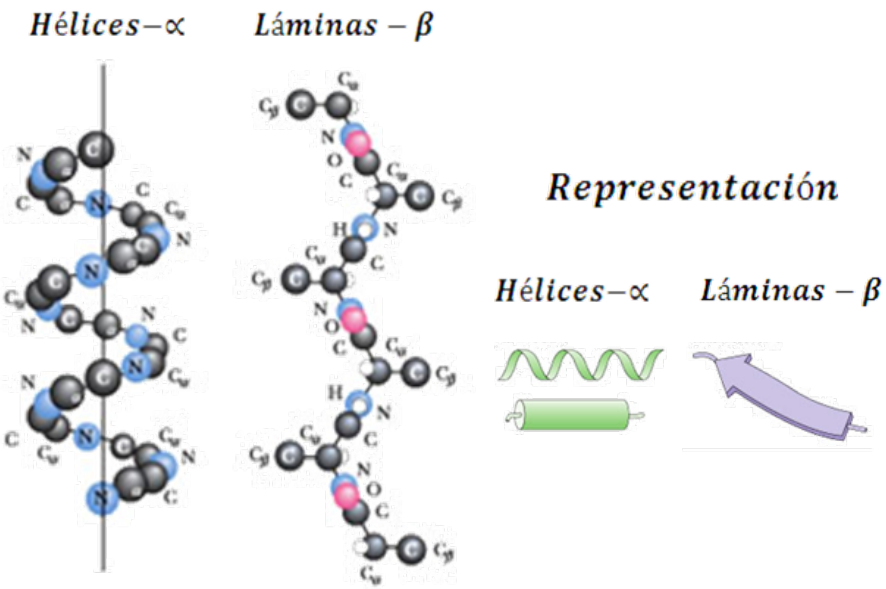

Figura 1: ***Hélice-***α*. Sólo se representa la secuencia* $N - C_\alpha - C$*. La línea vertical es el eje de la hélice-*α*.* ***Lámina-***β*. Se representa la secuencia* $N - C_\alpha - C_O$*, así como el* C_β *de los grupos* R*.*

[2] *Existe otro motivo estructural denominado Coil el cual son combinaciones de Hélices-*α *y Láminas-*β*.*

[3] *Imagen extraída de* `http://www.web.virginia.edu/Heidi/chapter5/chp5frameset.htm` *(Julio 18 de 2011).*

[4] *Imagen: Lcanthamoeba Castellanii ProïňĄlin ib (1acf), extraída de* `http://www.rcsb.org/pdb/explore.do?structureId=1ACF` *(Julio 18 de 2011).*

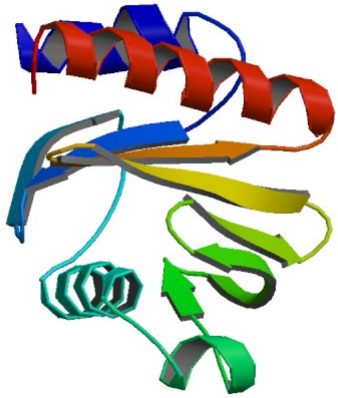

Figura 2: *Tercer nivel estructural de una proteína*

2. Marco teórico

2.1. Proteínas

Son polipéptidos longitudinales o plegados, compuestos comúnmente por 20 tipos diferentes de aminoácidos, cuya combinación determina de cierta manera la estructura general de la molécula y, esta a su vez define en gran medida la función que cumple la proteína. Esta secuencia de aminoácidos se pliega hasta formar siempre una misma estructura compacta, específica para cada tipo de proteína[2]. En condiciones normales poseen solamente una disposición espacial o conformación siendo el estado más estable que puede adoptar.

Comúnmente esta estructura es estudiada a través de técnicas experimentales, como la espectroscopia por resonancia magnética nuclear (RMN, por sus siglas en inglés) y la cristalografía con rayos X, siendo técnicas de gran costo en términos de equipos de laboratorio, tiempo en obtener resultados y complejidad en los métodos usados.

Dado el planteamiento anterior los métodos de optimización heurísticos se perfilan como los más promisorios enfoques para abordar el problema donde se modela la energía libre de una cadena de aminoácidos dada y a partir de allí encontrar aquellas estructuras que minimicen dicha energía.

2.2. El modelo HP

Modelo propuesto por Ken Dill[7] y referenciado en una gran cantidad de trabajos debido a su papel fundamental en el modelado de plegamiento de proteínas[15, 18, 10]. Es el modelo más simple utilizado para representar el plegamiento de una proteína, siendo estructurada como una cadena con dos únicos tipos: H (hidrofóbica o no polar, designada en color blanco) y P (hidrófilico o polar designada en color negro) (Figura 3[5]), los cuales describen la hidrofobicidad de una cadena de animoácidos (Tabla 2), y de esta manera forman así los enlaces intramoleculares: H-H, H-P y P-P.

Estos enlaces se forman mediante interacciones hidrofóbicas, las cuales se deben a la fuerte tendencia que posee el agua para excluir a los aminoácidos no polares (H), y no surgen tanto por una afinidad intrínseca entre estos, sino porque las moléculas de agua prefieren las interacciones que comparten unas con otras siendo estas más fuertes, en comparación con su interacción con las moléculas no polares que predominantemente residen en el interior de la proteína.

Hidrofobidicad de algunos de los principales aminoácidos	
Hidrofóbicos	Alanina, Fenilalanina, Isoleucina, Leucina, Metionina, Valina.
Polares	Cisteína, Glicina, Histidina, Lisina, Asparagina, Prolina, Glutamina, Arginina, Serina, Treonina, Triptófano, Tirosina.

Cuadro 2: *Clasificación de los principales aminoácidos según su hidrofobicidad.*

El modelo además se basa en el enfoque termodinámico que poseen el proceso de plegado en el

[5] *Imagen realizada por los autores.*

problema de la predicción de la estructura de las proteínas, donde se asume que al plegarse, las proteínas buscan minimizar la cantidad de energía utilizada por el sistema, llegando a un estado estable, llamado estado nativo[17]. Basandose además en el hecho de que las proteínas nativas al plegarse tienden a formar núcleos muy compactos dominantes impulsados por interacciones hidrofóbicas[9].

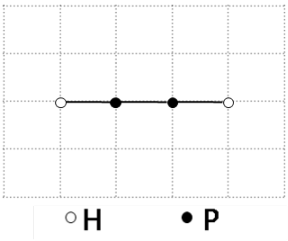

Figura 3: *Representación de una cadena de animoácidos usando el modelo HP.*

Esta secuencia de enlaces se pliega en una red cuadrada de dos dimensiones en el que en cada punto de la cadena puede girar $90°$ hacia arriba, hacia abajo (Figura 4[6]), o seguir adelante.

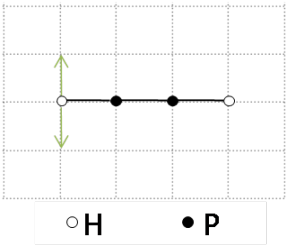

Figura 4: *Posibles movimientos para una representación de un aminoácido de tipo H.*

Con este proceso se obtiene la conformación plegada de la proteína y se procede a realizar el cálculo de la energía del plegamiento plasmado en dicha malla bidimensional, el cual, está definido como el número de contactos topológicos entre aminoácidos hidrofóbicos adyacentes no vecinos, es decir que cada aminoácido de tipo H que se encuentra frente a otro (Figura 5[6]), contribuye en -1 a la energía total esto es equivalente a maximizar el número de enlaces H-H en el modelo (Figura 6[6]). Así el cálculo del mínimo de energía libre de la conformación de plegamiento de proteínas se transforma en un problema de la optimización.

[6] *Imagen realizada por los autores.*

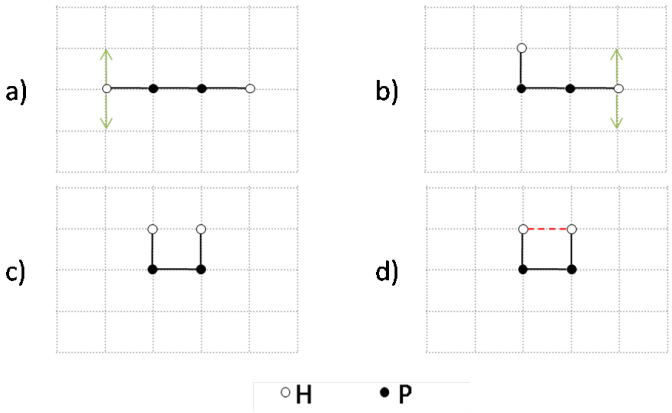

Figura 5: *Las conformaciones candidatas se pliegan mediante lo movimientos siguiendo las direcciones relativas arriba y abajo (**a**), uno a la vez en cada movimiento (**b**) los cuales indican, para cada aminoácido, su posición siguiente en la red con relación a su predecesor directo, con la que se obtiene una estructura óptima (**c**) con un enlace H-H, por lo que la energía de la estructura será -1 (**d**).*

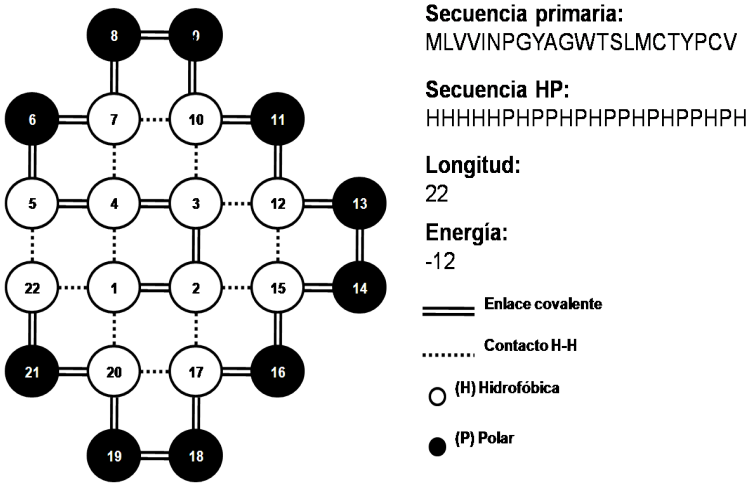

Figura 6: *Plegado óptimo de una secuencia HP, con un total de 12 contactos H-H, bajo el modelo HP 2D.*

2.3. Algoritmo de optimización de colonia de hormigas

Es una de las metaheurísticas utilizadas para explorar espacios de búsqueda, inspirada en la conducta colectiva de las hormigas planteada por Dorigo, Maniezzo y Colorni[19], basada en el comportamiento estructurado de una colonia de hormigas donde los individuos se comunican por medio de una sustancia química denominada feromona, estableciendo caminos cortos entre el hormiguero y la fuente de alimento (Figura 7[7]).

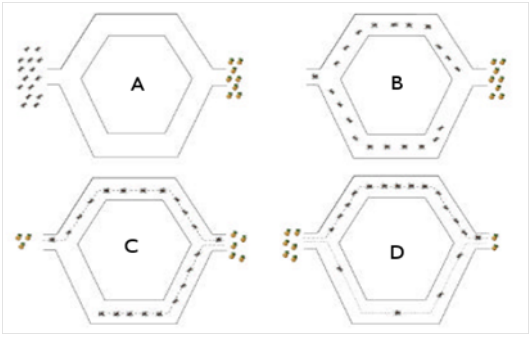

Figura 7: *En principio, las hormigas a ciegas elegirán aleatoriamente uno de las dos caminos (**B**), después de un tiempo casi todas las hormigas irán por el camino superior (**D**).*

El método consiste en simular computacionalmente la comunicación indirecta de un conjunto de agentes cooperativos, para establecer el camino más corto, donde el comportamiento de una hormiga es independiente de las demás durante la misma iteración. Este algoritmo ha sido empleado con éxito para abordar el problema de plegamiento de proteínas usando el modelo HP en 2D[16].

Los algoritmos de OCH son esencialmente algoritmos constructivos, es decir, en cada iteración, cada uno de los agentes construyen una posible solución al problema recorriendo un grafo de construcción (Figura 8[8]).

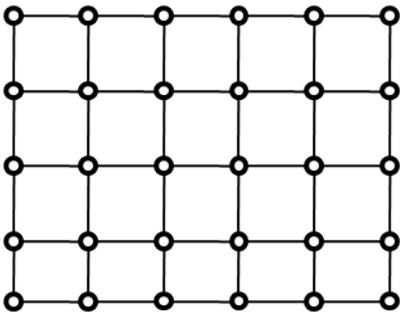

Figura 8: *Grafo de construcción.*

Cada arista del grafo representa los posibles movimientos que el agente puede realizar, y tiene asociada dos tipos de información que guían el movimiento:

- *Información de visibilidad:* Mide la preferencia de moverse desde el nodo *i* hasta el nodo *j*, es decir, recorrer la arista n_{ij} (Figura 9[8]). Los agentes no modifican esta información durante la ejecución del algoritmo.

[7] *Imagen extraída de* http://razonartificial.com/2010/01/optimizacion-basada-en-colonias-de-hormigas-en-la-naturaleza *(Julio 18 de 2011).*

[8] *Imagen realizada por los autores.*

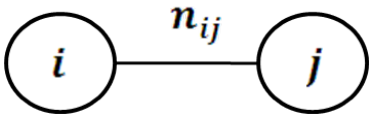

Figura 9: *Arista* n_{ij}.

- *Información de feromona artificial:* Mide la deseabilidad del movimiento de *i* a *j*. Esta información se modifica durante la ejecución del algoritmo dependiendo de las soluciones encontradas. Se nota por τ_{ij}, donde τ es una matriz con los valores del nivel de feromona en cada nodo.

La transición se basa en asignar la probabilidad de ir del nodo *i* al *j* según la ecuación:

$$P_{ij} = \begin{cases} \dfrac{(\tau_{ij})^{\alpha} \cdot (n_{ij})^{\beta}}{\sum_{l \in J_i} (\tau_{lj})^{\alpha} \cdot (n_{ij})^{\beta}} & \text{Si } j \in J_i \\ 0 & \text{Si } j \notin J_i \end{cases} \tag{1}$$

Donde *J* representa los nodos alcanzables desde el nodo *i*, α y β son parámetros definidos a priori que reflejan la importancia relativa de la feromona y la visibilidad respectivamente. Cuando $\alpha = 0$, solo la visibilidad es tenida en cuenta, o cuando $\beta = 0$, solo los rastros de feromonas son consideradas al elegir a cual nodo moverse.

Durante la construcción o al completar una posible solución, la hormiga la evalúa y modifica los rastros de feromonas en las componentes de la matriz de feromonas τ, la cual almacena los rastros de las áreas ya exploradas. Con esta información se guiará en la búsqueda a futuras hormigas. El algoritmo también puede incluir un proceso de evaporación de rastros de feromonas, y otras acciones como realizar optimizaciones locales sobre soluciones encontradas o actualizar la información global para guiar el proceso desde una perspectiva no local. La actualización y evaporación de feromonas se realizan según la ecuación:

$$\tau_{i,j}(t+1) = (1-\rho) \cdot \tau_{i,j}(t) + \rho \cdot \Delta\tau_{i,j}(t) \tag{2}$$

Donde ρ representa el coeficiente de evaporación, y $\Delta\tau_{i,j}(t) = \sum_{k=1}^{m} \Delta\tau_{ij}^{k}(t)$ con m es el número de hormigas. La cantidad inicial de feromonas en cada arista está uniformemente distribuida en pequeñas cantidades ($\tau \geq 0$). Después de cada iteración cada hormiga deja cierta cantidad de feromona $\Delta\tau_{ij}^{k}(t)$ en su recorrido, esta cantidad depende de la calidad de la solución encontrada:

$$\Delta\tau_{ij}^{k}(t) = \begin{cases} \dfrac{Q}{L^{k}(t)} & \text{Si } (i,j) \in T^{k}(t) \\ 0 & \text{Si } (i,j) \notin T^{k}(t) \end{cases} \tag{3}$$

Siendo $T^{k}(t)$ el camino recorrido por la hormiga *k* durante la iteración *t*, $L^{k}(t)$ la longitud de regreso y *Q* un parámetro constante.

3. Planteamiento del problema

De manera natural un péptido se pliega a medida que va siendo sintetizada y la cantidad de formas que puede adquirir, está determinada por la cantidad de aminoácidos que posea y en este sentido la forma que toma es única y se denomina también estado nativo; este plegamiento se realiza en un sólo proceso con tal precisión que la proteína resultante es funcional; pero dicho proceso aun no es muy conocido.

La cantidad de posibles combinaciones es enorme; sí se estimara el número de posibles plegamientos para una cadena corta de unos 100 aminoácidos, y asumiendo que solo posee dos posibles conformaciones o posibles movimientos en el espacio para cada residuo, existirían aproximadamente 10^{30} formas para dicha cadena; y si tan solo se requieren 10^{-11} segundos para pasar de una conformación a otra se necesitaría un tiempo del orden de 10^{11} años[11]; una cifra totalmente inverosímil considerando el promedio de vida actual del ser humano.

Encontrar la forma más energéticamente estable o forma nativa, en un espacio de búsqueda tan extenso hace de este un problema donde los algoritmos de inteligencia artificial poseen el potencial para abordar problemas de tipo np-duros en los cuales la heurística es más importante que una búsqueda determinista.

Este problema se puede plantear matemáticamente de la siguiente manera: dado el modelo HP propuesto por Dill en 1985[7] donde identifico interacciones hidrofóbicas de los aminoácidos y los clasifico como H (hidrofóbicos) y P (polares), y estos se ubican en una malla L en dos dimensiones definido como el plano cartesiano compuesta de puntos r_i que identifican las coordenadas (x, y); sobre el cual se desea realizar los plegamientos, y sea $S = \{s_1, s_2, ..., s_n\}$ una secuencia lineal donde cada s_i puede tomar el valor de H o P , sea C el conjunto de posibles conformaciones espaciales en 2D que puede adoptar la cadena S, entonces se define a $c \in C$ como la función $c : [1, ..., n] \to L$ que sería una conformación particular de S, donde $S \in L$.

$$\forall\ 1\ \leq i < n : (c[i]\ y\ c[i+1]) \text{ son vecinos} \tag{4}$$

$$\forall\ 1\ \leq j \leq n : (c[i]\ \neq\ c[j]) \tag{5}$$

Dada la secuencia $S \in \{H, P\}$ de longitud n y un plegamiento de la secuencia c en L, el cálculo de energía de la conformación obetnida al terminar el plegamiento c se realiza de la siguiente manera:

$$E(c) = \sum_{1\ \leq\ i+1\ <\ j\ \leq\ n} \beta(s_i, s_j)\delta(r_i, r_j) \tag{6}$$

En donde la función β evalúa las interacciones entre los tipos de aminoácidos en la secuencia S,*ver ecuación* 7 . Y la función δ valida si la interacción entre dos aminoácidos es una interacción de vecindad no adyacente, *ver ecuación* 8.

$$\beta(s_i, s_j) = \begin{cases} -1 \text{ Si } s_i \wedge s_j = H \\ 0 \text{ de otra manera} \end{cases} \tag{7}$$

$$\delta(r_i, r_j) = \begin{cases} 1\ \text{ Si }\ \|r_i - r_j\| = 1 \\ 0 \text{ de otra manera} \end{cases} \tag{8}$$

La ecuación 7, busca los aminoácidos no adyacentes que sean Hidrofóbicos. Y la ecuación 8 busca aquellos aminoácidos no adyacentes que sean vecinos, siendo $\|r_i - r_j\|$ la norma de la diferencia de la distancia que existe entre dos puntos en L. Si les distancia es igual a 1, se dice que los dos puntos son vecinos.

En el mismo sentido se puede decir que una proteína se pliega llegando a la conformación que menos energía disipe, por tanto lo que se busca es minimizar la función de energía, *ver ecuación* 6, y de esta forma encontrar la conformación c que aproxime el plegamiento más estable de la proteína.

En el mismo sentido sea $E : C \rightarrow \mathbb{R}$ la función que asocia un valor de energía a las posibles conformaciones espaciales $c \in C$ que pueda tener una cadena S, lo que se busca es encontrar $c^* \in C$ tal que $\forall\, c \in C : E(c) \geq E(c^*)$, de esta manera el problema de encontrar la confirmación que disipe la menor cantidad de energía posible, se puede plantear de la siguiente manera

$$\begin{aligned} &\text{mín } E(c) \\ \text{Sujeto a: } &\forall\ 1 \leq i < n\ \|c_i - c_{i+1}\| = 1 \\ &\text{y } \forall\ i \neq j :\ c_i \neq c_j \end{aligned} \tag{9}$$

En donde la primera restricción ilustra que la distancia entre aminoácidos adyacentes siempre debe ser la misma. Y la segunda restricción, que ningún aminoácido puede ocupar el mismo espacio en la malla que ya este ocupado por otro aminoácido.

4. Planteamiento de la solución

Para llevar a cabo el análisis a la problemática y una posterior adaptación del algoritmo OCH al problema de aproximar el plegamiento de proteínas, se inicio con una adaptación al problema del agente viajero, el cual consiste en buscar el mejor circuito cerrado, en términos de distancia, de una lista de n ciudades, dado que este fue uno de los primeros problemas para los cuales fue adaptado el algoritmo OCH.

En este ejemplo se evidenció el potencial que tiene el algoritmo OCH en la búsqueda de soluciones aproximadas a problemas cuyos tiempos de solución son computacionalmente costosos, y deja al descubierto los parámetros libres (variables) que este algoritmo ofrece, estos parámetros lo dotan del potencial para acercarse a la solución deseada, pero tienen que ser ajustados para que ofrezcan una mayor estabilidad y los resultados sean cercanos a los óptimos.

A continuación se realizó una adaptación en términos de las dos funciones involucradas en la dinámica del algoritmo OCH enunciadas como la visibilidad (distancia) y la recurrencia (feromonas), estos aspectos no son propios del problema del plegamiento proteico sino de la heurística utilizada, dado que el problema es de optimización. Haciendo énfasis en la tendencia actual de adaptar metodologías de solución que no son propias de un mismo campo, un ejemplo de esto es el de solucionar el problema del agente viajero mediante los SOM[3] (mapas auto-organizativos por sus siglas en inglés), se procedió al estudio de las adaptaciones disponibles en la literatura[4].

Durante este estudio se determinó que en la última década el algoritmo OCH había adquirido un particular interés por la comunidad en la búsqueda de una solución al plegamiento de proteínas, durante este tiempo diversas comunidades han propuesto nuevas adaptaciones, una de las cuales es reconocida por ofrecer una de las aproximaciones más estables, explicada posteriormente.

4.1. Adaptación del algoritmo de Optimización de Colonia de Hormigas

El algoritmo de OCH es constructivo en el tiempo, es decir, la mejor solución encontrada no se visualiza sino hasta el final de la ejecución del algoritmo, por ejemplo en el agente viajero, sí el número de ciudades es n entonces cada hormiga realizará n movimientos antes de dar a conocer un camino, y en cada uno de esos movimientos tendrá en cuenta la visibilidad y la recurrencia para dar el siguiente paso.

En la adaptación el mejor camino estará dada por la cadena de dirección (cadena compuesta de las letras U=arriba, D=abajo, R=derecha y L=izquierda), esta cadena es relativa al punto (0,0) del plano cartesiano $\mathbb{R}^2$ por tanto el camino realizado por cada una de la hormigas será una configuración particular de la cadena de aminoácidos en su representación del modelo HP[7].

Con la cadena de dirección dispuesta para determinar el camino y coste del mismo solo queda analizar cómo se comportan los parámetros nativos o principales del algoritmo de colonia de hormigas: la visibilidad y la recurrencia o feromonas.

4.1.1. Análisis de parámetros

- Visibilidad

La viabilidad de un movimiento está representada en una malla (Figura 10[9]), empezando por descartar los lugares por los cuales una hormiga ya ha estado durante la construcción del mismo camino y luego determinando el mejor movimiento contando con la información recolectada de los puntos a, b, c, d , e ,f y g, ya que esta retornara un valor característico en cualquiera de las siguientes opciones: sí hay o no un aminoácido, y si lo hay, que tipo de aminoácido (H o P) está en esa posición, este valor numérico es usado en el cálculo final de la mejor opción del camino.

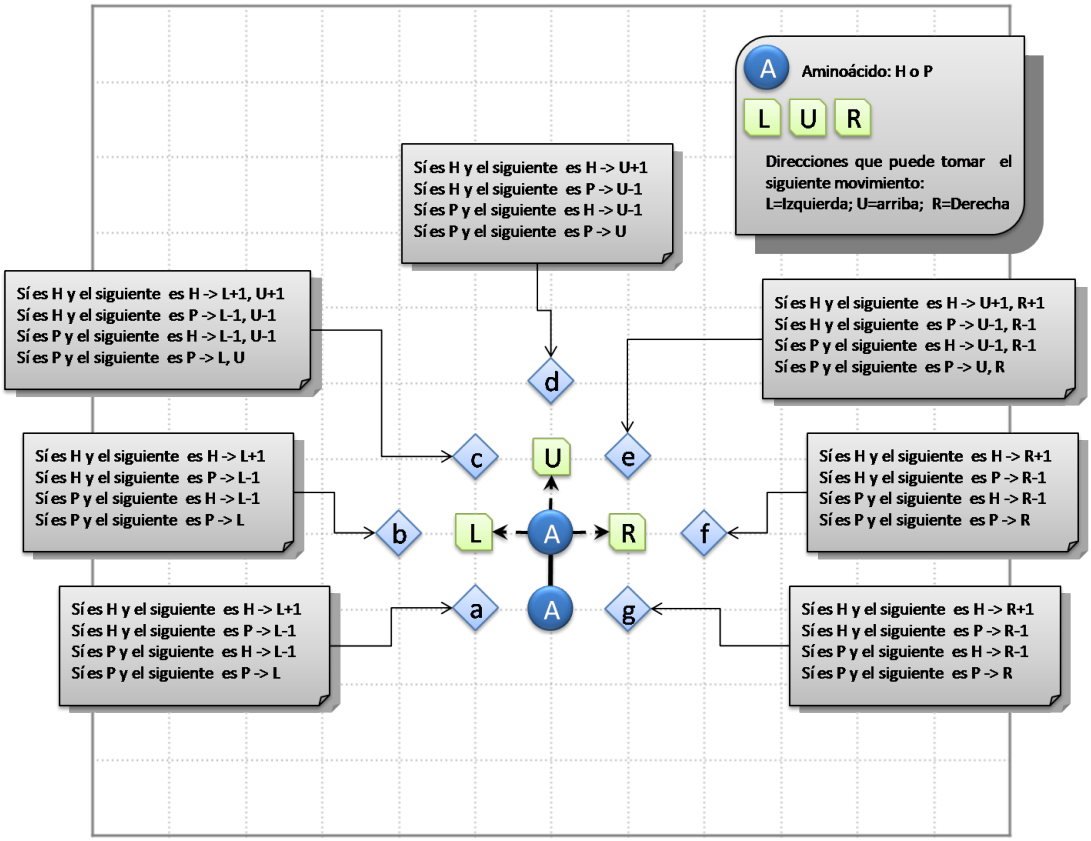

Figura 10: *Esta malla dará como resultado un valor ponderado para cada punto viable.*

- Recurrencia

La recurrencia o feromona, representa la sinergia del algoritmo OCH la cual es característica del mismo ya que el número de agentes buscadores (hormigas) tendrán la misma información y por tanto la convergencia indicará una recurrencia mucho mayor en un camino particular que en de los demás, debido a esto la ponderación de caminos se hace importante, ya que de una buena ponderación de caminos dará como resultado la convergencia correcta del algoritmo (entiéndase como una correcta convergencia el encontrar un mínimo, local o total, que tenga la menor cantidad de energía disipada).

[9] *Imagen realizada por los autores.*

4.2. Algoritmo

Pseudocódigo 1

1: Definición de los parametros: α, β, $SecuenciaHP$, $Cantidaddehormigas$, $Cantidaddeiteraciones$, ρ.
2: **mientras** $\neq Cantidaddeiteracionesor \neq Convergencia$ **hacer**
3: **para** $i = 0$ hasta $Cantidaddehormigas$ **hacer**
4: Hormiga(i).getCamino();
5: **fin para**
6: ActualizarFeromona();
7: Subrutina de detección de mínimos locales
8: **fin mientras**

4.3. Descripción de los parámetros libres y condiciones iníciales

4.3.1. Cantidad de hormigas

Existe una influencia directa entre la cantidad de hormigas y la complejidad computacional, es decir, entre más hormigas son usadas más rutas serán construidas, y más depósitos de feromonas serán calculadas generando carga computacional acorde a la cantidad de hormigas, por otro lado si se usan pocas hormigas se tendrá una baja exploración y por consecuencia poca información del espacio de búsqueda explorado.

4.3.2. Cantidad de iteraciones

La cantidad de iteraciones está directamente ligado con la exploración que puedan realizar las hormigas, es decir, pocas iteraciones implicaría una exploración muy pobre y demasiadas involucraría cálculos innecesarios.

4.3.3. Feromona inicial

Durante el paso de inicialización, todas los valores de las feromonas son inicializadas en un valor constante o aleatorio positivo, el cual indicará el área fértil para la exploración.

4.3.4. Heurística (α y β)

Sí $\alpha = 0$, aquellos nodos con una mejor preferencia heurística tiene una mayor probabilidad de ser escogidos, haciendo el algoritmo muy similar a un algoritmo voraz probabilístico clásico, de lo contrario α comenzará a ser un parámetro ponderador de la heurística.

Sí $\beta = 0$, sólo se tiene en cuenta los rastros de feromona para guiar el proceso constructivo, lo que puede causar un rápido estancamiento, esto es, una situación en la que los rastros de feromona asociados a una solución ligeramente superior al resto sea el más usado provocando que las hormigas siempre construyan las mismas soluciones normalmente óptimos locales o caminos que se construyeron en una iteración anterior.

4.3.5. Acumulación y afianzamiento de las feromonas ρ

Este es un parámetro que indica la tasa dentro de la fórmula de feromonas del algoritmo, se encuentra en el rango entre $[0, 1]$, y sí este valor es muy cercano a 1 indicará un afianzamiento de las feromonas alto y una evaporación muy baja, y sí es muy bajo indicará lo contrario.

Los anteriores son los parámetros básicos usados en el funcionamiento del algoritmo OCH, estos parámetros no deben ser escogidos con arbitrariedad, sino acorde al problema que se desea abordar, es decir, dichos parámetros poseen el potencial de hacer que algoritmo converja a una mejor solución, este es el potencial que ofrece el algoritmo de colonia de hormigas sin embargo este potencial también es una desventaja ya que estos parámetros libres dejan con una posibilidad enorme que en sí, es otro problema de búsqueda y optimización, por lo que para este trabajo se optó por realizar una pequeÃśa exploración limitando los rangos de valores de cada uno de los parámetros.

4.3.6. Secuencia

La secuencia debe estar dada en términos del modelo HP de Dill, es decir, la cadena aminoácidos debe ser previamente convertida en otra que la simplifique a sólo dos estados: (H) hidrofóbico o (P) polar[7].

4.4. Determinación de la mejor opción para el siguiente movimiento.

La determinación del siguiente movimiento en el algoritmo está dada por la ecuación 7 la cual incluye la visibilidad y el nivel de feromonas, además los parámetros libres del algoritmo α y β, siendo estos los lo guían en la aproximación a una solución.
Estos parámetros son los exponentes de la visibilidad y el nivel de feromonas respectivamente, por lo que un número mayor no indica una ponderación mayor ya que esta depende enteramente de la relación base-exponente, donde el siguiente movimiento esta dado por el cálculo en la malla anteriormente mencionada que otorgará un valor ponderaro representando la visibilidad, este valor contiene la información que se usará en el momento de escoger el siguiente movimiento, teniendo en cuenta los alrededores del mismo, es consecuente decir que esta malla se aplica en todas direcciones en las cuales se realizará un posible movimiento.

Uno de los factores que determinan estos parámetros es la feromona, información que todas las hormigas poseen y usan junto con la visibilidad para el cálculo del siguiente movimiento, es decir, todas las hormigas no sólo usan la visibilidad sino también tienen en cuenta el valor de la feromona la cual indica la recurrencia (cantidad de hormigas que han pasado por el mismo punto) que tienen el camino a lo largo de la ejecución del algoritmo.

Teniendo estos dos factores se posee un punto de referencia que permitirá determinar el siguiente movimiento a realizar para cada hormiga, y cuyo conjunto de movimientos conforman un camino, aunque este por sí mismo no indica la calidad del algoritmo si no hasta que sea ponderado y de esta manera observar el nivel de optimización que tiene el camino, es donde el cálculo la relación energía-área se hace importante ya que brinda un referente en el momento de realizar la actualización del nivel de feromona. Este cálculo de energía se realiza sumando los contactos topológicos entre los aminoácidos hidrofóbicos que se encuentren uno frente a otro, siempre y cuando no se encuentren unidos, es decir a una distancia en el plano cartesiano relativa de 1 (uno) y el área es el utilizada por el camino, donde cada contacto contribuye en -1 a la energía total.

En consecuencia la convergencia del algoritmo viene dada por la convergencia de las hormigas, es decir, todas ellas determinarán un camino viable, sin embargo este puede tomar distintas direcciones equivalentes que el algoritmo reconoce como diferentes pero simplemente son mismo camino rotado en múltiplos de $90°$ (Figura 11[10]), por lo que una correcta convergencia en este caso representada por el valor de la energía obtenido a lo largo de las iteraciones. El enfoque del

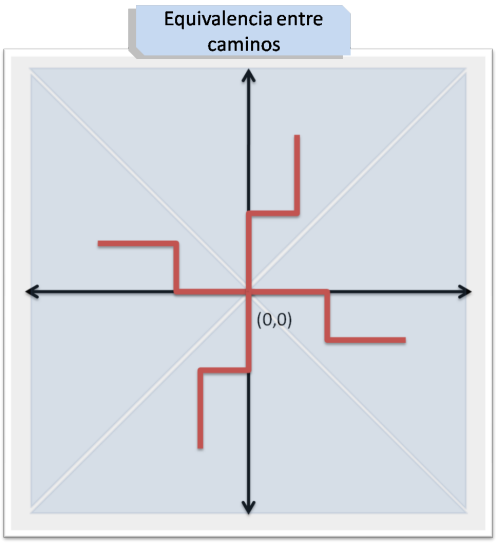

Figura 11: *Regiones de exploración y equivalencia entre caminos.*

algoritmo de OCH es elitista, esto es porque solamente se actualiza el mejor camino, además se puede considerar un algoritmo con una muy fuerte exploración ya que al ser independiente de la dirección realiza más movimientos libres ayudando a evitar estancamientos en mínimos locales.

4.5. Control de mínimos locales

El control de mínimos locales se realiza mediante la detección de estancamientos (movimientos oscilatorios del objetivo en un rango específico) mediante el análisis elitista de los caminos construidos por las hormigas, y se realiza para evitar que los mínimos locales sean confundidos como valores de convergencia, entonces se compara el valor del mínimo local con el mejor camino detectado hasta ese punto y sí dichos valores son iguales se considera una convergencia del algoritmo, de lo contrario se realizara una evaporación intensiva de las feromonas equivalente al número de iteraciones del estancamiento en el mínimo local.

4.6. Búsqueda de los parámetros libres

La búsqueda de los valores para ajustar los parámetros libres se realizó limitando el rango de los valores de los mismos y realizando una búsqueda clásica (una que no tengan en cuenta alguna heurística), en un banco de datos previamente preparado a través de la combinación de valores en los parámetros libres aplicados al algoritmo, de esta forma evidenciar el comportamiento del algoritmo y su estabilidad.

[10] *Imagen realizada por los autores.*

Este banco de datos se generó con una sencilla combinación de los valores que conforme se iban iterando realizaban pruebas a la exploración del algoritmo, la cual estaba representada por la mayor área cubierta por las hormigas en pocas iteraciones. Esta forma de determinar una solución en este algoritmo se denomina enfoque de exploración.

La combinación de los valores de los parámetros se realizó de la siguiente manera:

4.6.1. Los valores de α y β

Estos valores se ajustaron teniendo en cuenta la tendencia observada de los términos en la relación base-exponente.

Para validar la tendencia de la gráfica en el intervalo abierto $(0, 1)$ se plantea la siguiente relación.
Notando que:
La función $y = x^a \in \mathbb{R}^+$ entonces
$a \in \mathbb{R}^+$ y sí $a \in (0, 1) \Rightarrow a = \frac{R}{T} \Rightarrow x^{\frac{R}{T}} \equiv \sqrt[T]{a^R}$ con $R < T$
ahora, si consideramos que $T >> R$
podemos despreciar a R quedando la función de la siguiente forma:
$y = \sqrt[T]{a}$ De esta manera validamos la tendencia de una función de tipo raízen el intervalo abierto $(0, 1)$

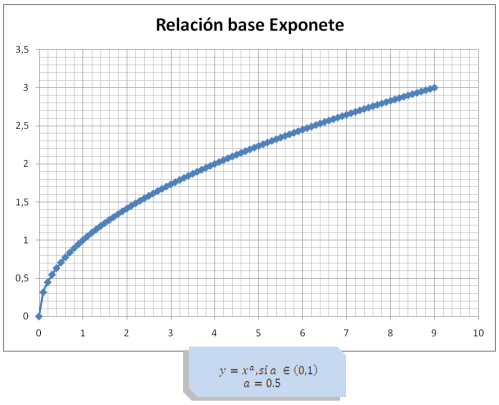

Figura 12: *Relación de tipo radical (raíz).*

Para validar esta tendencia (Figura 12[11]), en el intervalo $(1, 3]$ se plantea la siguiente relación
Notando que: La función $y = x^a \in \mathbb{R}^+$ (Figura 13[11]) entonces
$a \in \mathbb{R}^+$ y sí $a > 1 \Rightarrow a = \frac{R}{T}$
ahora consideramos a
$R > T \Rightarrow x^{\frac{R}{T}} \equiv \sqrt[T]{x^R} \equiv (\sqrt[T]{x})^R \Longrightarrow$ estableciendo la igualdad $\sqrt[T]{x} = M$
por ser la operación potencia la predominante,
se concluye que
$y = M^R$
Teniendo en cuenta esta tendencias, se optó por variar tanto el parámetro α como β en el intervalo $(0, 3]$, a una tasa de 0.01, la cual fue seleccionada de comparar el $\Delta_1 = 0{,}01$ con el $\Delta_2 = 1 \cdot 10^{-6}$

[11] *Imagen realizada por los autores.*

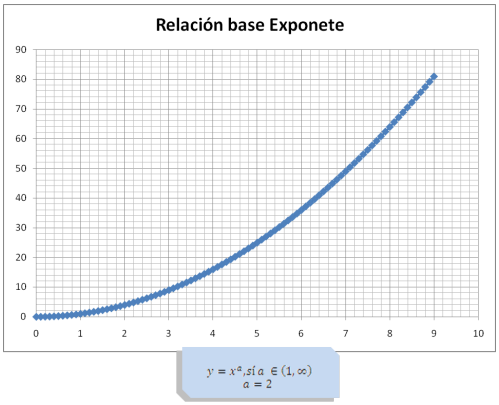

Figura 13: *Relación de tipo exponencial .*

y para verificar que sea cual sea el Δ seleccionado no tendría un error entre ellos significativo equivalente a $1 \cdot 10^{-5}$ y por lo tanto se planteo la siguiente demostración:

Hipótesis:
Partiendo de la función $y = x^a$ con $a \in 0 < a < 3$ en los $\mathbb{R}^+$ se pretende comprobar que la magnitud entre el punto O y el punto B (Figura 14[12]) utilizando un $\Delta_1 = 0{,}01$ no tendrá una diferencia mayor a $1 \cdot 10^{-5}$ con respecto a la sumatoria de las magnitudes de separación $\Delta_2 = 1 \cdot 10^{-6}$.

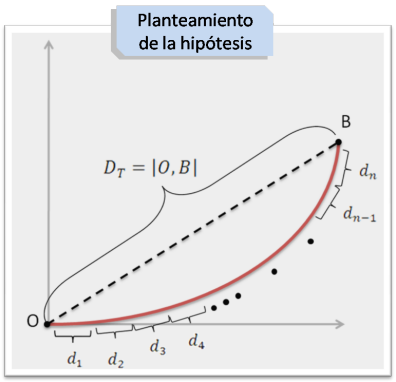

Figura 14: *Esquema del planteamiento de la hipótesis.*

Cálculo de D_T
$D_T \approx \sum_{i=1}^{n} d_i$
$O = (0, 0)$ y $B = (\Delta_1, f(\Delta_1))$
Se plantea el vector $\vec{OB} = (\Delta_1, (\Delta_1)^a)$ y se halla su magnitud que es igual:
$\left|\vec{OB}\right| = \sqrt[2]{(\Delta_1)^2 + (\Delta_1)^{2a}}$

Cálculo de la $\sum_{i=1}^{n} d_i$
Sabiendo que esta sumatoria forma parte en la deducción de la ecuación de longitud de arco, partiremos de este principio: $\lim_{\Delta x_i \to \Delta_2} \sum_{i=1}^{\infty} \sqrt[2]{1 + (f'(x))^2}$

[12] *Imagen realizada por los autores.*

Como $\Delta_2 = 1 * 10^{-6} = 0{,}000001 \approx 0 \Rightarrow \lim_{\Delta x_i \to 0} \sum_{i=1}^{\infty} \sqrt[2]{1 + (f'(x))^2}$ (Figura 15[13])

Sabiendo que: $f'(x) = a \cdot x^{a-1}$ y el límite de la sumatoria es por definición una integral, y planteada de la siguiente manera:

$$\int_0^{\Delta_1} \sqrt[2]{1 + a^2 \cdot x^{(2(a-1))}} dx$$

Tabla de datos			
Valor de a	$\Delta_1 = 0{,}01$	$\Delta_2 = 1 \cdot 10^{-6}$	$error\ \lvert\Delta_1 - \Delta_2\rvert$
1	$0,014142136$	$0,014142136$	0
$1,5$	$0,010056041$	$0,010049876$	$6,1652 \cdot 10^{-6}$
2	$0,010000667$	$0,0100005$	$1,66639 \cdot 10^{-7}$
$2,5$	$0,010000008$	$0,010000005$	$2,81249 \cdot 10^{-9}$
3	$0,01$	$0,01$	$4 \cdot 10^{-11}$

Cuadro 3: *Datos de la demostración.*

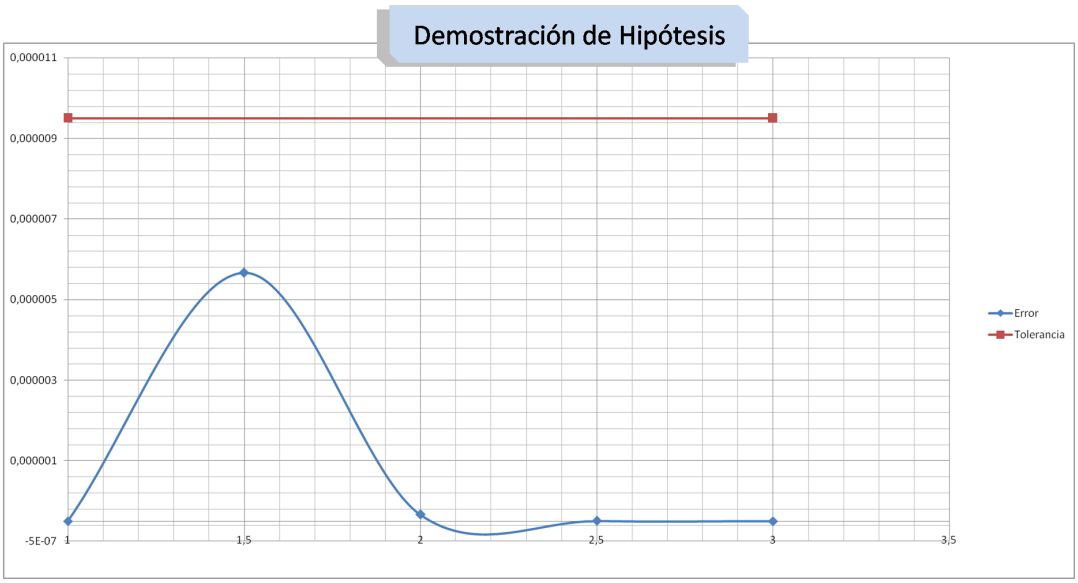

Figura 15: *Comportamiento del error con las distintas medidas calculadas de acuerdo a lo planteado en la hipótesis.*

4.6.2. Los valores de ρ (ro) o afianzamiento de las feromonas

El valor ρ es usado para determinar en qué tasa aumentará o se evaporará el valor de la feromona que enriquece un camino de acuerdo a la siguiente ecuación $\tau(t+1) = (1-\rho) * \tau(t) + \rho * \Delta$, donde $\tau(t)$, representa el valor de la feromona y delta representa una tasa constante que impedirá

[13] *Imagen realizada por los autores.*

que una multiplicación por cero anule los términos.
De esta expresión podemos inferir que ρ es un término que en esencia pondera porcentualmente el valor de la feromona en un t anterior de manera pues que el rango de valores que pueden tomar en un intervalo abierto son $(0, 1)$ y para el desarrollo del algoritmo OCH se vario con un Δ de 0.1.

4.6.3. Cantidad de hormigas

La cantidad de hormigas para este problema puede ser determinado teniendo en cuenta la forma en la que cada hormiga construye un camino, ya que cada una de ellas comienza en el punto de origen de la malla $(0, 0)$ y realiza el primer movimiento de acuerdo a un valor de igual probabilidad en cualquier dirección, este valor se obtineine por el método tradicional de Montecarlo, para distribuir el área de exploración, ahora se considera n hormigas y m regiones de exploración entonces, basados en la consistencia de los números pseudo-aleatorios del sistema (demostrada estadísticamente a través de las pruebas de $Chi - cuadrado$), no es un criterio a considerar debido a que todos los posibles movimientos en primera instancia tendrían la misma probabilidad, sin embargo para poder explorar las m regiones es necesario, como lo enuncia el principio de Dirichlet, también llamado principio del palomar, poseer más hormigas que regiones, entonces se tiene que $n > m$.

4.6.4. Número de iteraciones

Este número se eligió como límite superior de parada en la ejecución del algoritmo con un calor de 20.000 iteraciones debido a que será la última instancia de parada, puesto que este varía dependiendo de la velocidad con la que converja el algoritmo.

5. Resultados

En la validación de este tipo de algoritmos se usan secuencias estándar (Cuadro 4) de las cuales se conoce el mejor nivel de energía alcanzado, y donde el desempeño de las diversas metodologías ha sido evaluado, siendo esto una referencia que se tiene en cuenta al momento de realizar las pruebas pertinentes.
En el análisis de los resultados obtenidos en la exploración se usaron estas secuencias y se contrastaron los valores esperados (teóricos) con los valores obtenidos (experimentales) (Cuadro 5).

Los resultados para las pruebas preliminares de exploración y validación de la variación escogida de los parámetros libres (Cuadro 5), y evolución del algoritmo a lo largo de las distintas pruebas (Cuadro 6), con los mejores resultados obtenidos.

Secuencias estándar			
Número	**Longitud**	**Energía**	**Secuencia**
1	20	-10	HHHPPHPHPHPPHPHPHPPH
2	20	-9	HPHPPHHPHPPHPHHPPHPH
3	24	-9	HHPPHPPHPPHPPHPPHPPHPPHH
4	25	-8	PPHPPHHPPPPHHPPPPHHPPPPHH
5	36	-14	PPPHHPPHHPPPPPHHHHHHHPPHH PPPPHHPPHPP
6	48	-22	PPHPPHHPPHHPPPPPHHHHHHHHH HPPPPPPHHPPHHPPHPPHHHHH
7	50	-21	HHPHPHPHPHHHHPHPPPHPPPHPP PPHPPPHPPPHPHHHHPHPHPHPHH
8	60	-34	PPHHHPHHHHHHHHPPPHHHHHHH HHHPHPPPHHHHHHHHHHHHPPPPH HHHHHPHHPHP
9	64	-42	HHHHHHHHHHHHPHPHPPHHPPH PPHPPHHPPHHPPHPPHHPPHHPPH PHPHHHHHHHHHHHH

Cuadro 4: *Secuencias estándar usada en la evaluación del desempeño de algoritmos heurísticos.*

En Cuadro 5 podemos observar los resultados parciales obtenidos con la ejecución del algoritmo mediante el mecanismo usado para determinar los parámetros libres, hay que tener en cuenta que el número de iteraciones fue dispuesto con el fin de identificar una combinación de parámetros libres aceptable a través de la comparación de la exploración de las hormigas , evaluada en términos de las energías obtenidas comparadas con un bajo número de iteraciones.

Después de obtener los parámetros libres mediante el enfoque de exploración (Cuadro 5) se procedió a incrementar el número de iteraciones a 20.000, buscando una convergencia en el mismo o limitando el tiempo de ejecución al número máximo de iteraciones.
Estos resultados (Cuadro 6) son el producto de la ejecución del algoritmo con los parámetros libres anteriormente encontrados, los cuales evidencian el potencial del algoritmo OCH aplicados a este tipo de problemas, con el propósito de comparar los resultados con los reconocidos en la literatura se realiza la siguiente tabla.

Resultados parciales							
Número	**Energía esperada**	**Energía obtenida**	**Cantidad hormigas**	**Número de iteraciones**	α	β	ρ
1	-8	-5	10	200	1,9	0,8	0,5
2	-9	-6	10	220	1,7	1	0,9
3	-9	-6	10	200	0,6	0,8	0,7
4	-10	-10	10	200	0,2	0,8	0,1
5	-14	-9	10	230	1,8	0,6	0,9
6	-21	-13	10	210	1,8	1,8	0,7
7	-22	-15	10	230	1	1,9	0,5
8	-34	-24	10	230	0,8	0,6	0,5
9	-42	-23	10	220	0,4	1,2	0,7

Cuadro 5: *Resultados parciales obtenidos de la exploración realizada por el algoritmo.*

Resultados obtenidos							
Número	**Energía esperada**	**Energía obtenida**	**Cantidad hormigas**	**Número de iteraciones**	α	β	ρ
1	-8	-8	10	11500	1,9	0,8	0,5
2	-9	-9	10	9800	1,7	1	0,9
3	-9	-8	10	1000	0,6	0,8	0,7
4	-10	-10	10	200	0,2	0,8	0,1
5	-14	-9	10	4400	1,8	0,6	0,9
6	-21	-15	10	2550	1,8	1,8	0,7
7	-22	-15	10	2980	1	1,9	0,5
8	-34	-25	10	13000	0,8	0,6	0,5
9	-42	-27	10	20000	0,4	1,2	0,7

Cuadro 6: *Resultados obtenidos en la ejecución del algoritmo.*

En la comparación de los resultados (Cuadro 7) se muestra cómo el algoritmo realizó una aproximación exitosa en los casos donde la energía fue inferior (casos $-8,-9,-10$) evidenciando un correcto funcionamiento del algoritmo, sin embargo, el mismo parece quedarse corto cuando la exigencia obedece a longitudes de secuencia mayores a 36 aminoácidos representados de la forma HP.

Las secuencias usadas para comparar los resultados obtenidos al implementar del algoritmo OCH fueron tomadas de una tabla reconocida como estándar (Cuadro 4).

Comparación de resultados			
Número	**Energía esperada**	**Shmygelska & Hoos [1]**	**Algoritmo propuesto**
1	-8	-8	-8
2	-9	-9	-9
3	-9	-9	-8
4	-10	-10	-10
5	-14	-14	-9
6	-21	-21	-15
7	-22	-22	-15
8	-34	-34	-25
9	-42	-42	-27

Cuadro 7: *Comparación de los resultados obtenidos con los valores encontrados en la literatura.*

5.1. Caso de éxito

A pesar del corto tiempo usado en la búsqueda de los parámetros comparados con la complejidad del problema, con la ejecución del algoritmo se obtuvo un resultado exitoso (Figura 16[14]) donde los valores de los parámetros usados (Cuadro 8) llevaron al algoritmo a converger satisfactoriamente hacia la solución esperada, la cual muestra los aminoácidos involucrados en el plegamiento y los enlaces entre ellos dispuestos en la malla ($plano\ cartesiano$) dispuesta para el replegado proteico.

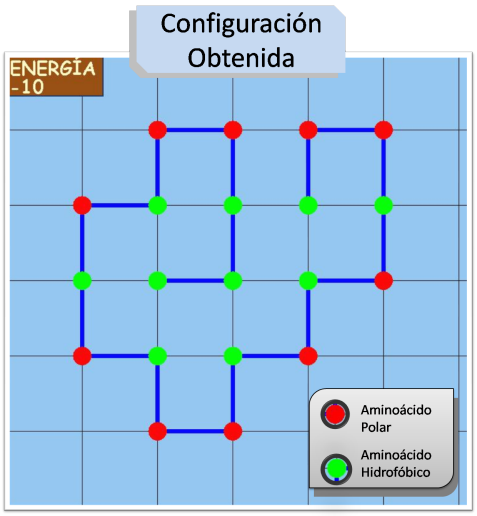

Figura 16: *Plegamiento resultante de la convergencia exitosa del algoritmo.*

[14] *Imagen realizada por los autores.*

Parámetros usados	
Secuencia	HHHPPHPHPHPPHPHPHPPH
Energía esperada	-10
Energía obtenida	-10
Número de hormigas	10
Iteraciones	185
α	0,2
β	0,8
ρ	0,1

Cuadro 8: *Datos con los cuales se obtuvo el caso de éxito.*

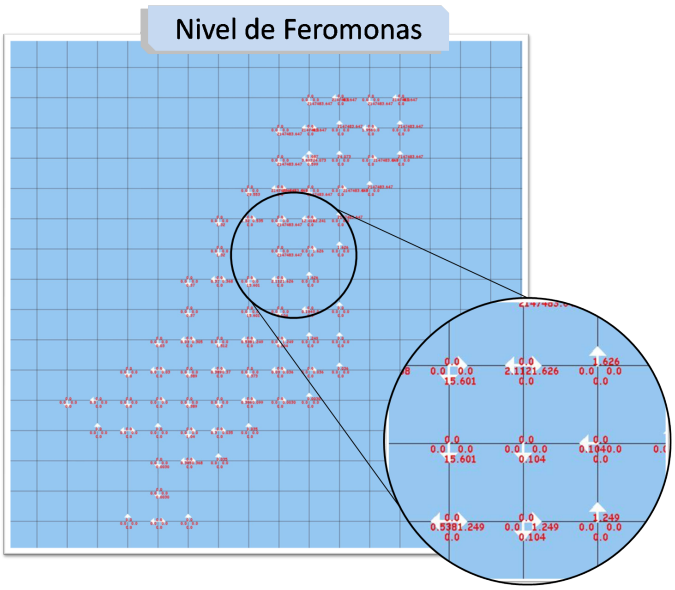

Figura 17: *Zona de exploración de las hormigas distinguidas por las flechas que indican la dirección que tomaron las hormigas al pasar por alguna coordenada y el nivel de feromonas en cada punto.*

En la Figura 17[15] podemos ver el área que exploraron las hormigas durante la ejecución del algoritmo, esta área se encuentra sobre la malla ($plano\ cartesiano$) dispuesto para el plegamiento proteico, para ayudar al análisis del algoritmo esta representación cuenta con flechas que indican los caminos desde cada vértice de la malla, estas están acompaÃśadas de un valor numérico que indica el nivel de feromonas en esas direcciones.

[15] *Imagen realizada por los autores.*

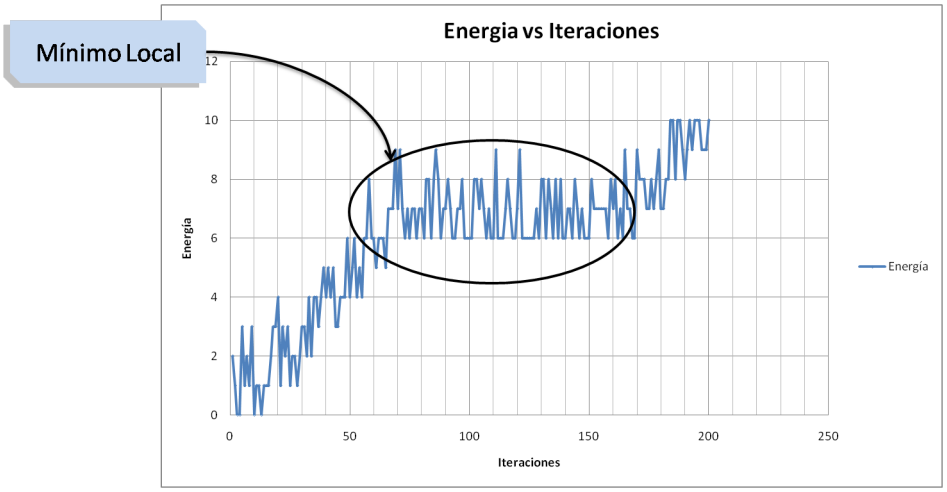

Figura 18: *Comportamiento del algoritmo a lo largo de la convergencia.*

En este caso en particular (Figura 18[16]) se observa como el algoritmo se estanca entre las iteraciones 70 y 150 la cual se detecta como una convergencia parcial, pero debido a los picos en el valor de la energía, el algoritmo encontró valores mínimos locales y los trato de corregir, logrando un caso exitoso llegando a una convergencia estable.

[16] *Imagen realizada por los autores.*

6. Interfaz

Con el propósito de mostrar los resultados generados por el algoritmo y permitir una interacción con el usuario que desea emplearlo con sus propios datos se creó una interfaz orientada a la web, que permita el ingreso de los parámetros libres, así como la secuencia HP y obtenga de manera visual los resultados generados para su posterior interpretación.

6.1. Aspectos considerados

- Tecnología en el servidor donde se implantará la herramienta.
- Disposición e impacto de la herramienta en el servidor.
- Dinámica de la herramienta (datos de entrada - datos de salida)
- Información de uso

Con estos aspectos es posible analizar la viabilidad de desarrollar la herramienta e implementarla en un servidor.

6.2. Análisis

- Tecnología en el servidor donde se implantará la herramienta.

  El servidor en el cual se implantaría la herramienta actualmente es del Grupo de Investigación de Informática Biomédica (GIIB), posee el software Apache Tomcat o Jakarta Tomcat [17] que implementa las especificaciones de los Servlets, Applets y de Java Server Page (JSP), de esta manera se acotan las posibilidades de lenguajes de programación disponibles a usar, a Java [18].

- Disposición y control de impacto de la herramienta en el servidor.

  El servidor ofrece una disposición constante para la petición y respuesta de recursos web dispuestos en el mismo, sin embargo no dispone de recursos suficiente para soportar una carga real de ejecución de algoritmos, puesto que si la cantidad de peticiones simultáneas realizadas a la herramienta es superior a 140 harían colapsar el servidor por memoria insuficiente para ejecutar de manera satisfactoria todas las peticiones realizadas.

- Dinámica de la herramienta.

  La dinámica de la herramienta contiene dos fases:

  - Datos de entrada: Es donde se le proporciona la información necesaria coherente y válida representada en valores numéricos para los parámetros y alfabéticos (H y P) para la secuencias y de esta manera se pueda ejecutar el algoritmo a satisfacción .
  - Datos de salida: Es donde el resultado obtenido de la ejecución del algoritmo es visializado representado en dos ventanas, una con la estructura plegada y la otra con el nivel de feromonas alcanzado en el área de exploración.

[17] Apache y el logotipo de Apache son marcas registradas de The Apache Software Foundation
[18] Java es una marca registrada de Oracle y sus afiliados.

- Información de uso.

  Esta debe ser clara con el propósito de que el usuario contextualice la información que se pide en la página y la use de manera eficiente.

Analizando estos aspectos se optó por utilizar el lenguaje de programación Java para desarrollar Java Server Pages (JSP) y Java Applets, el primero para el manejo del formulario donde se ingresarán los datos de entrada e interacción de las páginas web, y el Java Applet para desplegar interacciones de tipo gráfico y de ejecución de algoritmos de Java en el equipo del usuario y así liberar la carga de ejecución al servidor.
Se puede inferir que la arquitectura más adecuada a utilizar es la de Cliente-Servidor (Figura 19[19]), la cual consiste básicamente en un cliente que realiza peticiones a un servidor donde se encuentra implantada la herramienta la cual dará resultados a dichas peticiones.

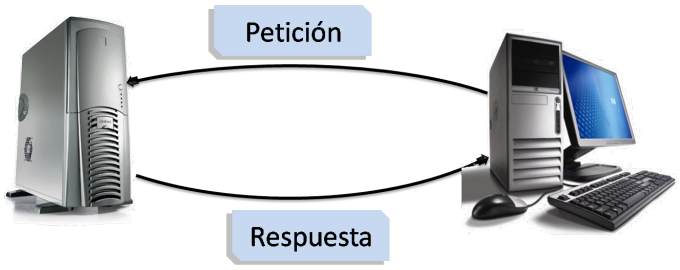

Figura 19: *Arquitectura Cliente-Servidor.*

Compilando lo anterior se establece una interacción por parte del usuario con la herramienta y los recursos representadas en la Figura 20[19] y la Figura 21[19].

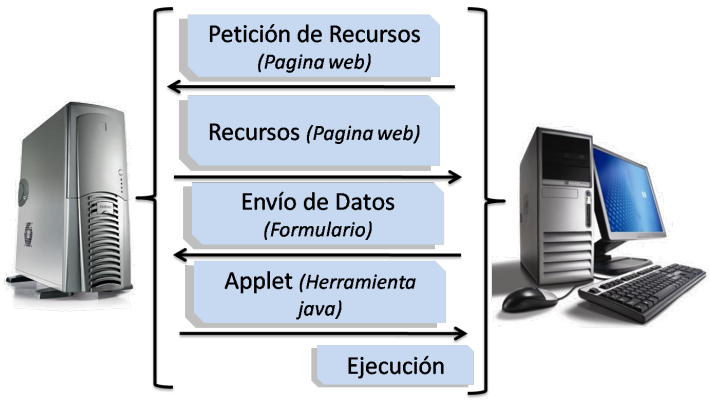

Figura 20: *Interacción de la herramienta en la arquitectura Cliente-Servidor.*

[19] *Imagen realizada por los autores.*

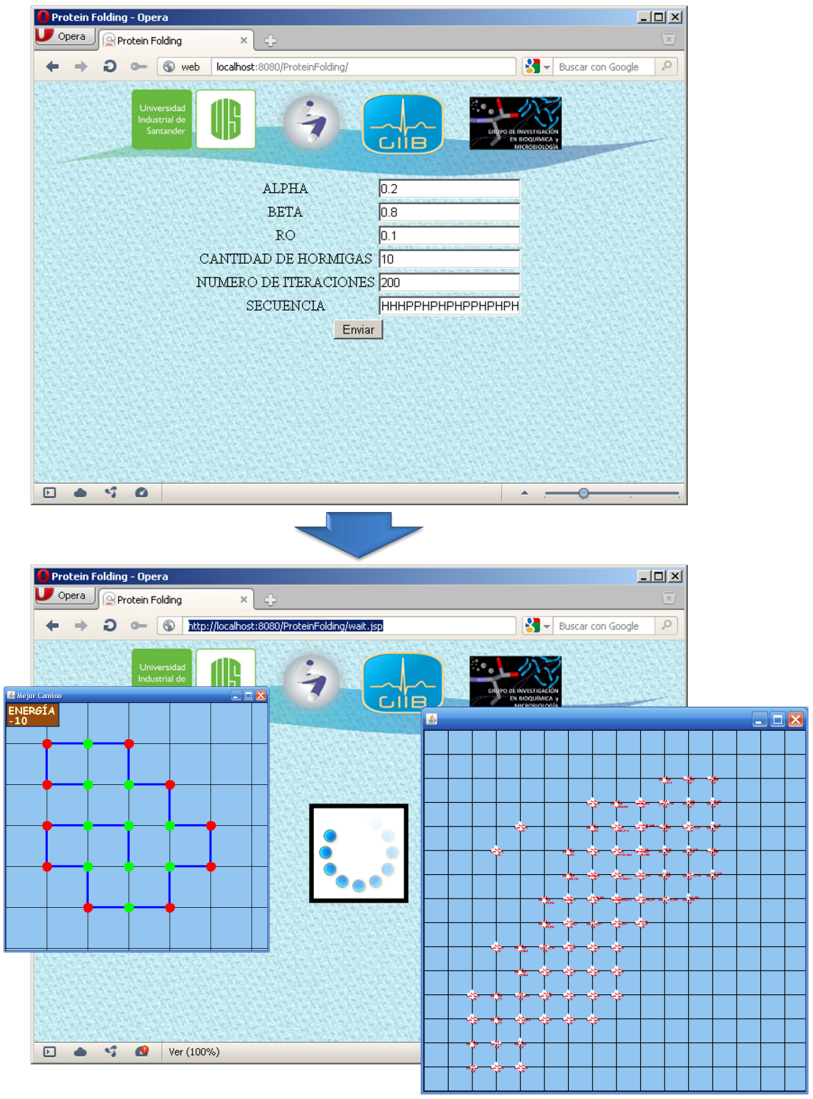

Figura 21: *Interacción de la herramienta web con el usuario final.*

7. Conclusiones

Apoyarse en metodologías heurísticas para abordar problemas cuya solución por medio técnicas tradicionales resultan muy costosas, es una forma de aprovechar tanto el potencial que nos brindan estas metodologías como la oportunidad de realizar proyectos interdisciplinarios que permiten un mejor aprovechamiento de los conocimientos adquiridos en diversas áreas del conocimiento.

Además la utilización de este tipo de algoritmos que no son comúnmente utilizados para aproximar soluciones a estas problemáticas, ayuda en la diversificación en los métodos de búsqueda de la solución usados.

El enfoque elitista utilizado en el algoritmo OCH brinda la posibilidad de terminar el algoritmo en cualquier momento y resaltar el mejor valor encontrado hasta ese momento.

De acuerdo con resultados obtenidos en la exploración (Cuadro 5) se observa el potencial que tiene el algoritmo OCH en cuanto a la exploración del universo de soluciones de un problema que no posee una solución por algoritmo lineal viable.

8. Recomendaciones

Los resultados aquí obtenidos se ven limitados por el tiempo estimado para la realización del proyecto, sin embargo son lo suficientemente aproximados, si se comparan con aquellos algoritmos que han tenido años de desarrollo (Cuadro 7), dado esto, una continuación en el uso de la metodología explorada, es ajustar los parámetros libres, los cuales permitan una mayor aproximación en secuencias con longitudes relativamente grandes.

Continuar con la investigación en este campo debido a la interdisciplinaridad que ofrece la bioinformática, con el fin de fomentar el trabajo en equipos que no competan conocimientos de una única ciencia además de ofrecer problemáticas cercanas a la realidad.

Realizar pruebas de ejecución en equipos cuya capacidad computacional este adaptada para ejecutar este tipo de algoritmos sin reparo en la utilización de recursos.

Bibliografía

[1] Rosalía Aguirre-Hernández Alena Shmygelska and Holger H Hoos. An ant colony optimization algorithm for the 2d hp protein folding problem. 2002.

[2] C.B. Anfinsen. Principles that govern the folding of protein chain. science 181, 223-230. 1973.

[3] Lucas Brocki. Kohonen self-organization map for the traveling salesperson problem. 2002.

[4] Dr Martyn Amos Chris Wilton. Testing an ant colony optimization algorithm for the two-dimensional hydrophobic-polar protein folding. 2003.

[5] Dunn B. M. Speicher D. W. Wingfield P. T. Coligan, J. E. and H. L. Ploegh. *Current protocols in protein science*, 2000.

[6] Voet D. and G. Voet J. Biochemestry, second edition. 1995.

[7] K. A. Dill. Theory for the folding and stability of globular proteins. biochemistry, 24. 1985.

[8] T. Igor F. Protein structure prediction bioinformatic approach. 2002.

[9] K. Yue K. M. Fiebig D. P. Yee P. D. Thomas K. A. Dill, S. Bromberg and H. S. Chan. Principles of protein folding: A perspective from simple exact models. *Prot. Sci., 4:561Ű602.*, 1995.

[10] K. F. Lau and K. A. Dill. A lattice statistical mechanics model of the conformation and sequence spaces of proteins. *Macromolecules, 22:3986Ű3997.*, 1989.

[11] Cyrus. Levinthal. Are there pathways for protein folding? *Journal de Chimie Physique et de Physico-Chimie Biologique 65: 44Ű45.*, 1968.

[12] Antonio P.D. Bioquímica, segunda edición. 1988.

[13] J. G. Pertó. Fundamentos de bioqímica. universitat de valéncia. 2007.

[14] Darío José Delgado Quintero. Predicción de la estructura 3d de proteínas usando técnicas basadas en inteligencia artificial. 2011.

[15] S. Will R. Backofen and P. Clote. Algorithmic approach to quantifying the hy-drophobic force contribution in protein folding. *Altman R (ed), Pacific Symposium on Biocomputing, Honolulu, pages 93Ű106*, 2000.

[16] Hoos H. Shmygelska A. An ant colony optimisation algorithm for the 2d and 3d hydrophobic polar protein folding problem. *BMC Bioinformatics.*, 2005.

[17] J. Skolnick. Putting the pathway back into protein fonding. *Proc Natl Acad Sci USA 102, 2265-2266.*, 2005.

[18] G. Shi T. Jiang, Q. Cui and S. Ma. Protein folding simulations of the hydrophobic-hydrophilic model by combining tabu search with genetic algorithms. *The Journal of chemical physics, 119(8):4592Ű4596.*, 2003.

[19] A. Colorni V. Maniezzo, M. Dorigo. Algodesk: An experimental comarison of eight evolutionary heuristics applied to the qap problem. *European Journal of Operation Reseach, 81. pp. 188-204*, 1995.