DISEÑO DE UN SISTEMA PARA LA COMPARACIÓN AUTOMÁTICA DE SECUENCIAS DE PROTEÍNAS BASADO EN EL ANÁLISIS DE CLUSTERS HIDROFÓBICOS (HCA).

CINDY DAYANA SOLANO MEZA
Ingeniera de Sistemas

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FISICOMECÁNICAS
ESCUELA DE INGENIERÍA DE SISTEMAS
MAESTRÍA EN INGENIERÍA DE SISTEMAS E INFORMÁTICA
BUCARAMANGA
2016

DISEÑO DE UN SISTEMA PARA LA COMPARACIÓN AUTOMÁTICA DE SECUENCIAS DE PROTEÍNAS BASADO EN EL ANÁLISIS DE CLUSTERS HIDROFÓBICOS (HCA).

CINDY DAYANA SOLANO MEZA

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

Directores del Proyecto:

CRISTIAN BLANCO TIRADO, Químico *Ph.D.*
DARIO JOSÉ DELGADO QUINTERO, Ingeniero de Sistemas *MSc.*

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FISICOMECÁNICAS
ESCUELA DE INGENIERÍA DE SISTEMAS
MAESTRÍA EN INGENIERÍA DE SISTEMAS E INFORMÁTICA
BUCARAMANGA
2016

A mi familia, por todo su apoyo incondicional,
por estar siempre conmigo y apoyarme en todos mis proyectos y decisiones.

A mis amigos, quienes siempre confiaron en mí y
fueron un gran apoyo en todo el camino.

A todas las personas que conocí en este tiempo, porque de
ellos aprendí muchas cosas buenas para la vida.

A ti, que pudiste compartir conmigo esta experiencia.

Cindy Dayana Solano

Agradecimientos

A mi director de investigación, profesor Cristian Blanco Tirado, por sus aportes, sus conocimientos, su apoyo, su confianza y orientación en el presente trabajo de investigación.

A Darío José Delgado por ser un guía, un amigo. Sus concejos y conocimiento me ayudaron durante toda la investigación y a la terminación de la misma de forma exitosa.

Al grupo de investigación GIFTEX y todos mis compañeros, por su apoyo durante el trascurso de la maestría, y a todas aquellas personas que directa o indirectamente me ayudaron en este proceso de formación.

A mis compañeros de maestría, porque de ellos recibí apoyo en todo momento.

A Álvaro Gaitán director de CENICAFE, y Narmer Galeano, por su interés en trabajar en convenio con la Universidad industrial de Santander y el grupo de investigación GIFTEX.

A la Universidad Industrial de Santander, por formar excelentes profesionales y brindarme la posibilidad de ser parte de su comunidad.

CONTENIDO

Pág.

LISTA DE FIGURAS

Pág.

LISTA DE TABLAS

Pág.

GLOSARIO

AMINOÁCIDO: Compuesto orgánico que contiene un grupo amino (-NH_2) y un grupo carboxilo (-COOH). Los aminoácidos son los constituyentes de las proteínas, y existen 20 esenciales en las proteínas de los humanos.

BALIBASE: Base de datos que contiene alineamientos de secuencias de proteínas construidos manualmente, de alta calidad, con detalles de anotación.

***CLUSTER*:** (Inglés) Agrupación de elementos pertenecientes ha determinado grupo, las cuales presentan un tipo de características específicas en común o están asociados a una función particular.

***EIIP* (Potencial de interacción ion-electrón):** representa el promedio de estados de energía de todos los electrones de valencia para cada aminoácido en particular.

FILOGENÉTICA: relaciones de proximidad evolutiva entre las distintas especies.

GAP: Las penalizaciones por *gap* se utilizan en el alineamiento de secuencias ya sea de ADN, ARN o secuencias lineales de proteínas en la introducción de desplazamientos para encontrar coincidencias en los elementos que constituyen las secuencias en comparación.

GENOMA: Es todo el material genético contenido en las células de un organismo en particular.

HCA: (Hydrophobic Cluster Analysis) Análisis de *clúster* hidrofóbicos, técnica de comparación 2D de proteínas.

PDB: (**Protein Data Bank**) es un repositorio para datos estructurales en 3D de proteínas. Estas estructuras se han resuelto mediante metodologías experimentales de resonancia magnética nuclear (RMN) y difracción de rayos X (DRX).

PÉPTIDO: Son un tipo de moléculas formadas por la unión de varios aminoácidos mediante enlaces peptídicos.

POLIPÉPTIDO: Es el nombre utilizado para designar un péptido de tamaño suficientemente grande.

PROTEÍNA: Biomoléculas poliméricas conformadas por cadenas de monómeros denominados aminoácidos, son fundamentales en la constitución de la materia viva.

PROTEÓMICA: Estudio a profundidad de las proteínas, en particular de su relación estructura - función.

RESUMEN

TÍTULO: DISEÑO DE UN SISTEMA PARA LA COMPARACIÓN AUTOMÁTICADE SECUENCIAS DE PROTEÍNAS BASADO EN EL ANÁLISIS DECLUSTERS HIDROFÓBICOS (HCA)[1].

AUTOR: CINDY DAYANA SOLANO MEZA[2]

PALABRAS CLAVE: Análisis de Clusters Hidrofóbicos (HCA), Automatización, Comparación de proteínas, Correlación de señales, Transformada discreta de Wavelet (DWT).

Desde hace dos décadas, la secuenciación de genomas se ha convertido en un proceso de bajo costo. Cada genoma transcribe una gran cantidad de proteínas, cada una con una función específica en un organismo vivo. Entender su función es una tarea compleja debido a la cantidad de información de proteínas previamente identificas con las que se podrían comprar. Este escenario ha llevado a la implementación de métodos de comparación y análisis de secuencias que permitan extraer información estructural y funcional a partir de secuencias ya reconocidas.

Se realizan comúnmente dos tipos de análisis sobre secuencias desconocidas: la detección de la identidad u homología a través de la alineación de secuencias o el análisis estructural. Estos métodos se centran en calcular una métrica porcentual de identidad entre secuencias; si el valor de identidad es superior al 35%, se considera entre pares de proteínas estructura similar, y para identidad inferiores al 25% secuencias sin similitud. Aquellos grupos de secuencias entre el 25% y 35% de identidad, se le denomina zona de penumbras (*Twilight Zone* [1]), en donde es confuso establecer características específicas entre pares de secuencias.

En este trabajo se presenta un nuevo modelo para realizar la comparación de secuencias de proteínas con bajo porcentaje de identidad utilizando la transformada discreta de wavelet y el de análisis clúster hidrofóbicos (HCA [2]). El modelo se enfoca en un esquema de codificación de secuencias de proteínas utilizando la representación B y Q proporcionada por HCA [3,4] y los valores EIIP para cada aminoácido en la estructura primaria. Dicha información se emplea para tratar las secuencias de proteínas mediante la transformada discreta de Wavelet y el análisis de correlación. El método identifica posibles secuencias con homología estructural de manera automática donde al igual se identifica la wavelet madre con mejor aproximación a los resultados deseados.

[1] Trabajo de grado

[2] Facultad de Ingenierías Físico Mecánicas. Escuela e Ingeniería de Sistemas e Informática. Director: Cristian Blanco Tirado. Codirector: Darío José Delgado.

ABSTRACT

TITLE: DESIGN OF A SYSTEM FOR AUTOMATIC COMPARISON FOR PROTEIN SEQUENCES BASED HYDROPHOBIC CLUSTER ANALYSIS (HCA)[3].

AUTHOR: Cindy Dayana Solano Meza[4]

KEY WORDS: Discrete Wavelet Transform (DWT), Hydrophobic Cluster Analysis (HCA), Automatically Approach, Protein Sequence Comparison, signal correlation.

Over the last two decades, genome sequencing has become a low-cost process. Each genome transcripts several proteins, each one with a specific function inside a living organism. Understanding their function is a complex task due to the amount of protein information previously identified which they could be compared. This scenario has led to the implementation of comparison methods and sequence analysis that allow extracting structural and functional information from already known sequences.

Two types of analysis are commonly performed over unknown sequences: Detecting identity or homology through sequence alignment or structural analysis. These methods focus on calculating a percentage measure of sequence identity. If the identity value exceeds 35%, a similar structure between pairs of proteins is estimated, for values lower than 25% the sequences have no similarities. Those groups of sequences that fall between a 25% and 35% identity value are considered within twilight zone [1] , where is confusing to set specific characteristics between sequences.

This work presents a new model for comparison of protein sequences with low percentage of identity using the discrete wavelet transform and hydrophobic cluster analysis (HCA [2]). The model focuses on an encoding scheme of protein sequences using the B and Q representation provided by HCA [3,4] and EIIP values for each amino acid in the primary structure. This information is used to process protein sequences using the discrete Wavelet transform and correlation analysis. The method identifies automatically potential sequences with structural homology likewise, the mother wavelet with the best approximation to the desired results is identified.

[3] Research work

[4] Faculty of Physical- Mechanical Engineering. System engineering and informatics department. Advisor: Cristian Blanco Tirado. Co- advisor: Dario José Delgado

INTRODUCCIÓN

Las proteínas son estructuras biológicas constituidas por secuencias lineales de aminoácidos. Las proteínas son las responsables de la mayor parte de las funciones vitales en las células. Conocer la estructura de una proteína y asociarla a su función en el organismo es un proceso tedioso que involucra una combinación de técnicas analíticas químicas, bioquímicas y computacionales. Dos técnicas instrumentales que se utilizan para resolver la estructura de las proteínas son la Resonancia Magnética Nuclear y la Difracción de Rayos X. Actualmente existen bases de datos que contienen un gran número de proteínas clasificadas según su estructura y función, por ejemplo, Protein Data Bank (PDB) [5], Refseq [6], Uniprot [7], y SCOP [8]. Estas bases de datos sirven de referencia para resolver nuevos genomas, de manera que es posible conocer *a priori*, solo con la secuencia, la estructura y función de proteínas desconocidas que tengan algún porcentaje de identidad con alguna de las que se encuentran en la base de datos. Para esto se han desarrollado métodos computacionales de alineación de proteínas que miden su grado de similitud, con base en la secuencia de aminoácidos.

Algunas de las herramientas computacionales más utilizadas en el análisis de proteínas son FASTA [9], ClustalW [10], BLAST y PSI-BLAST [11] las cuales son generalmente eficientes pero con importantes limitaciones. Por ejemplo, estas herramientas no predicen si dos proteínas son homólogas (es decir que tienen la misma función debido al mismo origen evolutivo de los dos organismos) cuando difieren notablemente en su secuencia (cantidad de aminoácidos en la cadena que comparten la misma posición e identidad) [12]. Este problema es complejo cuando la comparación de la nueva secuencia con las de la base de datos resulta en porcentajes de identidad entre el 20% y el 35% (*twilight zone* [1]), porque aunque es posible que las dos proteínas sean homólogas es difícil establecer esa relación con los mecanismos de alineación convencionales. Por esta razón es importante desarrollar nuevas herramientas que permitan establecer efectivamente si existe homología o no entre dos secuencias de proteínas cuya identidad se encuentra en ese intervalo.

En este trabajo se muestran los resultados obtenidos al modelar la comparación de secuencias de proteínas como un sistema orgánico, donde la entrada es una o varias secuencias de proteínas, cuya comparación lineal se encuentra en el *twilight zone*, y la salida es el resultado de la comparación sistemática utilizando como motor la metodología HCA y transformada discreta de *Wavelet*.

El motor consta de las siguientes etapas: 1) la codificación en HCA para la identificación de los clústeres hidrofóbicos en las secuencias de comparación,2) El pre-procesamiento de las secuencias que implica crear el vector con la información HCA de la proteína y asignarle una señal *F(x)*, teniendo en cuenta la secuencia de *clusters* hidrofóbicos y un peso asociado con su valor EIIP; 3) el proceso de comparación mediante DWT y el análisis de correlación.

Los resultados de este trabajo de investigación se han presentado en los siguientes eventos nacionales e internacionales:

1. V Encuentro Nacional de Químicos Teóricos y Computacionales (V–ENQTC) y de la II Escuela Colombiana de Teoría y Computación en las Ciencias Moleculares (II–ECTCCM), 2014, Guatapé.
2. I Escuela Colombiana de Teoría y Computación en las Ciencias Moleculares y IV Encuentro Nacional de Químicos Teóricos y Computacionales, Cali, Colombia, 2012.
3. ISCB-LatinAmerica-2012 Conference on Bioinformatics, Santigo, Chile, 2012.

Este documento se encuentra organizado de la siguiente forma: en la sección 2 se muestra una revisión de la literatura junto con antecedentes asociados a modelos utilizados para la comparación de secuencias de proteínas. En la sección 3 se presenta un modelo propuesto como complemento para lograr el análisis de altos volúmenes de secuencias de proteínas. En la sección 4 se presenta el planteamiento del problema y su desarrollo conceptual. La sección 5 describe la metodología. La sección 6 la validación del método propuesto, y por último una sección de análisis, conclusiones y recomendaciones para trabajos futuros.

1. Proteínas

Las proteínas son moléculas biológicas que cumplen diversas funciones en los seres vivos. Algunas proteínas sirven de soporte, de tejido, de portador de señales en la célula, de transporte, como enzimas, entre otras funciones. Se sabe que la función de las proteínas depende de sus estructuras terciaria y cuaternaria, las cuales a su vez dependen de sus estructuras secundaria y primaria, esta última, la secuencia lineal. No es fácil determinar a partir de la estructura primaria de una proteína su estructura terciaria o su función [13].

Las proteínas se representan mediante una cadena de caracteres lineal, finita y ordenada. Cada carácter representa la sigla un aminoácido en la cadena. La secuencia de las proteínas se puede determinar mediante las técnicas: 1) secuenciación directa de la proteína (utilizando espectrometría de masas, por ejemplo); 2) mediante la expresión del genoma, utilizando el gen y replicando la secuencia o 3) Conociendo el genoma y realizando la transcripción a través técnicas computacionales.

1.1. Aminoácidos

Los aminoácidos que conforman las proteínas son moléculas que contiene un grupo funcional ácido (COOH) y un grupo amino (NH_2), en donde el grupo amino está unido al carbono contiguo al grupo carboxilo, como se aprecia en la Figura 1.

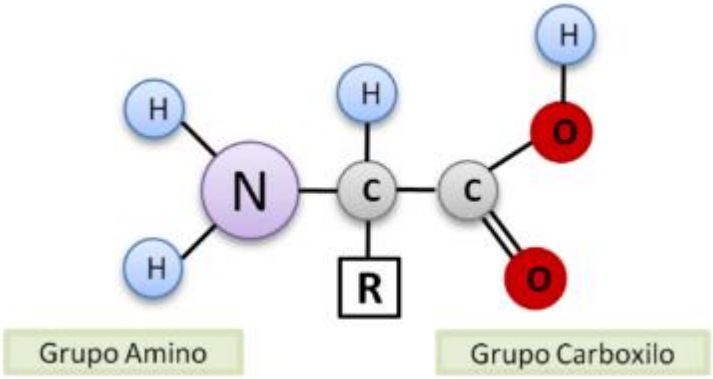

Figura 1. Estructura básica de un aminoácido con el grupo amino y carboxilo

Existen cientos de aminoácidos, sin embargo en las proteínas de los eucariotas solo se observan 21 de ellos. En los humanos solo 20 aminoácidos están presentes, los cuales son listados en la Tabla 1 [14].

AMINOÁCIDOS			
NOMBRE	**REPRESENTACIÓN 1 Letra**	**NOMBRE**	**REPRESENTACIÓN 1 Letra**
Valina	V	Alanina	A
Leucina	L	Prolina	P
Treonina	T	Glicina	G
Metionina	M	Serina	S
Triptófano	W	Cisteína	C
Lisina	K	Asparagina	N
Histidina	H	Glutamina	Q
Fenilalanina	F	Tirosina	Y
Isoleucina	I	Ácido aspártico	D
Arginina	R	Ácido glutámico	E

Tabla 1. Aminoácidos esenciales y no esenciales. Adaptado de [14]

La única diferencia entre los aminoácidos es el grupo R. Dependiendo de la naturaleza química del grupo R se derivan las propiedades de cada aminoácido, tales como acidez, hidrofobicidad, entre otras.

1.2. Niveles estructurales de las proteínas

Las proteínas presentan diferentes tipos o niveles estructurales, dependiendo de la manera como se organizan los aminoácidos en la cadena lineal, así (ver Figura 2) [14]:

- Estructura primaria: nivel básico, consiste en la secuencia lineal de aminoácidos y está determinada por el orden de los nucleótidos en el ADN o el ARN, en donde el número de estructuras posibles está relacionado con las variaciones y repeticiones de los 20 aminoácidos.

- Estructura secundaria: está asociada con las alteraciones que puede sufrir la secuencia a causa de la formación de enlaces del tipo puentes de hidrógeno, interacciones de van der Waals, hidrofóbicas e hidrofílicas. Estas interacciones hacen que la molécula se enrolle o pliegue y adopte una estructura secundaria. Debido a estas interacciones las proteínas formas dos tipos de estructuras: hélices alfa y hojas plegadas o beta.

- Estructura terciaria: corresponde con la disposición tridimensional de todos los átomos que componen la proteína y es la responsable directa de sus propiedades biológicas. En el caso de las proteínas que solo constan de una cadena polipeptídica, la estructura terciaria es su mayor nivel estructural.

- Estructura cuaternaria: La estructura cuaternaria es el producto de la asociación de varias cadenas polipeptídicas, finitas, de gran complejidad, unidas entre sí por interacciones no covalentes.

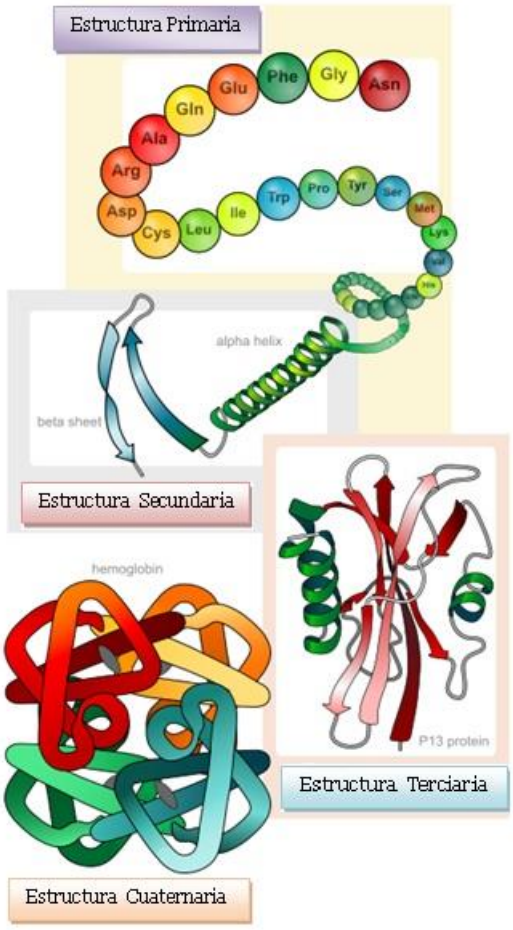

Figura 2. Representación de las 4 formas estructurales de una proteína. Adaptado de [15].

De la Figura 2 se puede establecer que la estructura terciaria o cuaternaria de una proteína está definida por la estructura secundaria y esta a su vez por la estructura primaria. Es tan importante la estructura primaria en las proteínas que con solo cambiar un aminoácido puede desnaturalizar su función. Enfermedades como el Parkinson, el Alzheimer o la de las vacas locas, se deben a mutaciones o cambio de uno de los aminoácidos en las proteínas que afectan su estructura y por ende su función en el organismo [16]. Hoy en día, encontrar la secuencia lineal de una proteína es relativamente sencillo a través del genoma, esto debido a los avances en las tecnologías de secuencias de genomas. Sin embargo encontrar su estructura a partir de esa secuencia es una tarea compleja que se logra solo mediante técnicas experimentales como Resonancia Magnética Nuclear, Difracción de Rayos X o utilizando técnicas computacionales de comparación con otras secuencias hasta

encontrar homología. En muchos casos, encontrar estructuras homólogas de proteínas es una tarea imposible de lograr.

2. Comparación de secuencias de proteínas

Una forma de identificar la funcionalidad de una proteína o su estructura es mediante la utilización de métodos de comparación lineales. Estos métodos buscan zonas de similitud significativa entre dos o más secuencias, dichas zonas permiten localizar características de interés común o diferencial, realizando una búsqueda del mayor número de coincidencias. Existen varios métodos utilizados para el análisis de secuencias a través de alineamientos y búsquedas en bases de datos de secuencias de proteínas ya conocidas [11,13,19].

2.1. Métodos tradicionales de comparación

Los métodos frecuentemente utilizados son BLAST[11], FASTA[9] y CLUSTALW [18], que se basan en el alineamiento y comparación de secuencias utilizando su estructura primaria. La similaridad de dos proteínas se define como el porcentaje de aminoácidos idénticos que existen entre ellas, en donde se observa el mínimo número de coincidencias de una secuencia problema con una secuencia previamente identificada. La identidad de dos proteínas es el porcentaje de aminoácidos idénticos que comparten dos proteínas. Estas herramientas se utilizan para encontrar patrones de diagnóstico, para caracterizar familias de proteínas, para detectar o demostrar homología entre nuevas secuencias y las familias existentes y para ayudar a predecir las estructuras secundarias y terciarias de nuevas secuencias.

Varias modificaciones se han realizado a estos algoritmos para optimizar la medida de similaridad o identidad entre secuencias. Actualmente se encuentra: PSI-BLAST con un enfoque estructural en la secuencia [11], BLAST 2 para alineamientos múltiples entre dos secuencias para detectar homología o duplicaciones internas [19], 3D BLAST para la comparación de bases de datos con respecto información de la estructura 3D de la proteína [20], entre otros.

El resultado de las técnicas de comparación tradicionales presentan un rango de confiabilidad; secuencias con un porcentaje de identidad mayor al 35% se considera que tienen una función y estructura similar. Sin embargo, aquellas en las que el porcentaje de identidad está entre el 20% y el 35% (rango de porcentaje de identidad al cual se le ha llamado zona de penumbras o *Twilight Zone* [1]*)* no son concluyentes. En algunos casos proteínas en ese rango tienen una alta homología, en otros no. Por esta razón, muchos transcriptomas muestran secuencias de proteínas cuya estructura y función es desconocida. Estudiar similitud, identidad y homología de proteínas en el *twilight zone* es una tarea difícil que requiere el desarrollo de nuevas herramientas de comparación.

2.2. Métodos alternos de comparación

Existen herramientas computacionales como HMMER y SAM [21] que utilizan cadenas ocultas de Markov (HMM). Estas herramientas se han empleado extensivamente para crear grandes bases de datos de alineamientos de secuencias como Pfam [22]. HMMER es también utilizado para refinar las alienaciones progresivas y mejorar la sensibilidad de la alineación [23].

Otra herramienta de alineamiento de secuencias es MUSCLE[24]. Este programa permite alineamientos múltiples de secuencias utilizando estimación de distancias, alineamiento progresivo utilizando una función propia diseñada por sus creadores y un proceso de refinado del alineamiento utilizando el árbol de partición restringida dependiente [24]. Por otra parte Callebaut y Mornon [25] utilizan el método de comparación de secuencias de proteínas que involucra las características estructurales 2D que se pueden establecer a partir de la estructura primaria utilizando HCA desarrollado por Gaboriaud [2,4].

2.3. Comparación de secuencias de proteínas mediante HCA

HCA se fundamente en mostrar gráficamente la proteína en 2D, y agrupar los aminoácidos de acuerdo con su hidrofobicidad. La utilidad de este método consiste en encontrar homología entre grupos de secuencias cuyos porcentajes de identidad se encuentran en el *twilight zone* [4]. El método se basa en la transformación de la secuencia lineal en una representación 2D. Esta representación de las proteínas no necesariamente representa su estructura secundaria, aunque se ha demostrado que la agrupación de aminoácidos hidrofóbicos en esta representación captura los motivos típicos de hélices α y de las hojas β. El método HCA consiste en la comparación visual de dos proteínas donde la forma de los *clusters* hidrofóbicos, su tamaño y la identidad de los aminoácidos en la cadena lineal son tenidos en cuenta, ver Figura 3 [3].

Las agrupaciones hidrofóbicas se forman por cercanía entre los aminoácidos hidrofóbicos V, I, L, F, M, Y, W. Estas agrupaciones de aminoácidos se codifican como sub-secuencias binarias que definen las regiones hidrofóbicas. A cada aminoácido hidrofóbico se le asigna el valor de uno (1) y a los no hidrofóbicos el valor de cero (0). Una agrupación termina cuando más de 4 aminoácidos no hidrofóbicos están en la cadena, o cuando aparece una prolina. Además de este tipo de codificación binaria, se encuentra también la codificación *Peitsch code* (código P), y la codificación Q como se aprecia en la Figura 4 [26].

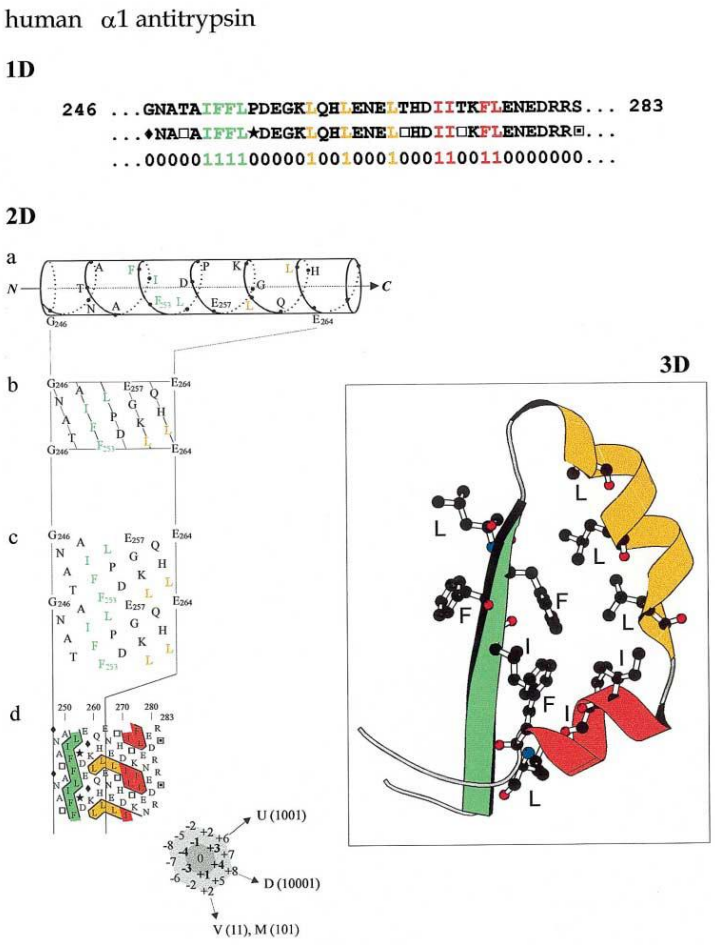

Figura 3. Representación gráfica HCA entre la secuencia y la estructura. Secuencia lineal 1D de un segmento de α- antitrypsin humana. La misma secuencia es representada en el plegamiento HCA, La secuencia es reportada como una α- hélice en un cilindro (a), el cilindro es cortado en forma paralela a su eje y desenrollado en un diagrama bidimensional (b), el cual se duplica para restaurar el ambiente completo de cada aminoácido en su representación α-hélice (c), posteriormente se pueden identificar las agrupaciones hidrofóbicas (d). Se muestra de manera coloreada las agrupaciones hidrofóbicas formando clusters. En la parte inferior derecha se puede observar la estructura 3D de esa fracción de la proteína. Los colores muestran las regiones de la cadena asociadas con la estructura secundaria de la proteína. Adaptado de [3]**.**

Secuencia lineal	KDQRNTLDLIAPSPADAQHWVQGLRKIIHHSGSMMQRQK		
Clúster Hidrofóbicos	000000**10110***0*000 0 0**11001 0 011**00000 **1 1**00000		
Código P	11	403	3
Código Q	MV	VUUV	V
(1= VILMFYW)	Clúster 1	Clúster 2	Clúster 3

Figura 4. Representación HCA de una secuencia hipotética que muestra los códigos B, P y Q

El código P considera que cada agrupación tiene un valor decimal definido por la suma de las potencias de dos del código binario de cada agrupación. Por ejemplo, la estructura 110101 corresponde en código P a 53, es decir $((1 \times 2^5) + (1 \times 2^4) + (0$

$\times 2^3$) + (1$\times 2^2$) + (0$\times 2^1$) + (1 $\times 2^0$)). El código P, permite una alternativa sencilla de descripción de las agrupaciones, por lo tanto es muy utilizado en términos de almacenamiento y clasificación computacional, especialmente para secuencias muy largas [26], en la figura 4 se observan las 3 agrupaciones hidrofóbicas con su valor P.

El código Q considera las agrupaciones como la concatenación de cuatro agrupaciones básicas: 11V (Vertical), 101M (mosaico), 1001U (Up) and 10001D (*Down*) (por ejemplo: 10101= 101+ 101= MM). Esta codificación es muy utilizada ya que representa una forma sencilla de para describir *clusters,* especialmente aquellos de gran longitud [26].

El único problema de HCA es que depende de un experto entrenado en encontrar visualmente la correspondencia entre los *clusters* de las dos proteínas. La comparación dependerá entonces de la habilidad del experto. A pesar de la información valiosa que se puede obtener con HCA, depender del entrenamiento de un experto y el tiempo que pueda tomar su análisis la hace muy poco útil. Gracias a que la metodología tiene un trasfondo codificado en B, P y Q, consideramos que es posible implementar un sistema computacional que utilice esta metodología. Ya en la literatura existen intentos de lograr tal automatización. Por ejemplo, Silva y colaboradores desarrollaron un algoritmo para calcular el porcentaje de identidad y homología de dos proteínas [27]. Sin embargo, esta metodología no permite hacer la comparación automática y sistemática de una proteína contra todo el PDB, por ejemplo.

Silva desarrolló un indicador numérico de la similitud de los grupos hidrofóbicos, dicha medida permite la detección automática de información en bases de datos mediante la comparación de patrones generados por codificaciones HCA. Este método se basa en la detección de agrupaciones similares de residuos hidrofóbicos en dos secuencias, conceptualmente similar a la alineación de secuencias, en este caso dicho alineamiento se realiza únicamente con aminoácidos hidrofóbicos y mediante un análisis de permutaciones [27].

Debido a que las codificaciones P, B, y Q, de HCA permiten analizar secuencias de proteínas, independientemente de la imagen bidimensional, consideramos que se puede desarrollar un sistema autoconsistente para la comparación de proteínas, utilizando como motor los códigos HCA, el algoritmo de Silva y la comparación mediante el procesamiento de señales.

2.4. Comparación a través de procesamiento de señales

La representación lineal de proteínas se puede utilizar para hacer comparaciones entre ellas. Sin embargo esta metodología, como se vio anteriormente, no es eficiente para predecir la estructura y funcionalidad de nuevas secuencias de proteínas. Recientemente se han creado nuevos métodos matemáticos para la

comparación de proteínas, por ejemplo MAFFT (Multiple sequence alignment based on fast Fourier transform) para alinear [28], RRM (Resonant recognition model) para determinar la función [29], o DWT para encontrar formas estructurales o similitud a lo largo de la secuencia [30].

MAFFT, es un método de alineamiento múltiple de secuencias basado en la transformada de Fourier, el cual permite una rápida detección de segmentos homólogos. Utiliza un sistema de puntuación, para aquellas secuencias que poseen largas inserciones o extensiones de aminoácidos, al igual que para la comparación de aquellas secuencias distantemente relacionadas pero de longitud similar [28]. Otro método que utiliza el tratamiento digital de señales enfocado en la asociación de funcionalidades utilizando la transformada de Fourier, es el caso del método RRM (Resonant Recognition Model) [31].

RRM, es un modelo matemático que analiza la interacción de una proteína y una secuencia identificada con la utilización de la transformada discreta de Fourier. Este modelo abarca la predicción de la función biológica de una proteína y su estructura 3D a partir de la secuencia lineal de aminoácidos [31]. La función puede ser considerada como una medida de la similitud entre las diferentes secuencias de la proteína en el dominio de frecuencia, cuando cada secuencia es tratada como una serie numérica. La señal más prominente de frecuencia muestra la similitud espectral de las secuencias. Esta similitud en secuencia es global porque el espectro es una contribución de todos los aminoácidos individuales en la secuencia [12].

La transformada discreta de Wavelet permite la comparación de varias señales de manera sistemática. La ventaja de esta técnica de comparación de señales, con otras disponibles, radica en la posibilidad de encontrar similaridad en dos dimensiones. Esta metodología de comparación no es nueva en proteínas. Se ha utilizado para predecir la estructura secundaria de proteínas [30]; estudios sobre la estructura primaria y su evolución [32], caracterización de motivos en proteínas [33], la comparación para la identificación de similitud en secuencia [23], entre otros. A continuación en el presente trabajo se desarrolla un modelo de comparación de proteínas tomando como ventaja la metodología HCA bidimensional.

3. Antecedentes

En este trabajo se desarrolló un sistema de comparación de secuencias proteínas constituidas por dos componentes principales: a) un sistema de pre-selección de secuencias con el fin de reducir el campo de búsqueda y b) el sistema de comparación utilizando HCA y la DWT, ver Figura 5.
El sistema de pre-selección compara y asocia una secuencia problema con un grupo de secuencias del PDB que ha sido clasificado de acuerdo con sus propiedades HCA, los aminoácidos que componen la cadena y la posición en la misma. A

continuación se describen todos los pasos que se involucran en la implementación del sistema.

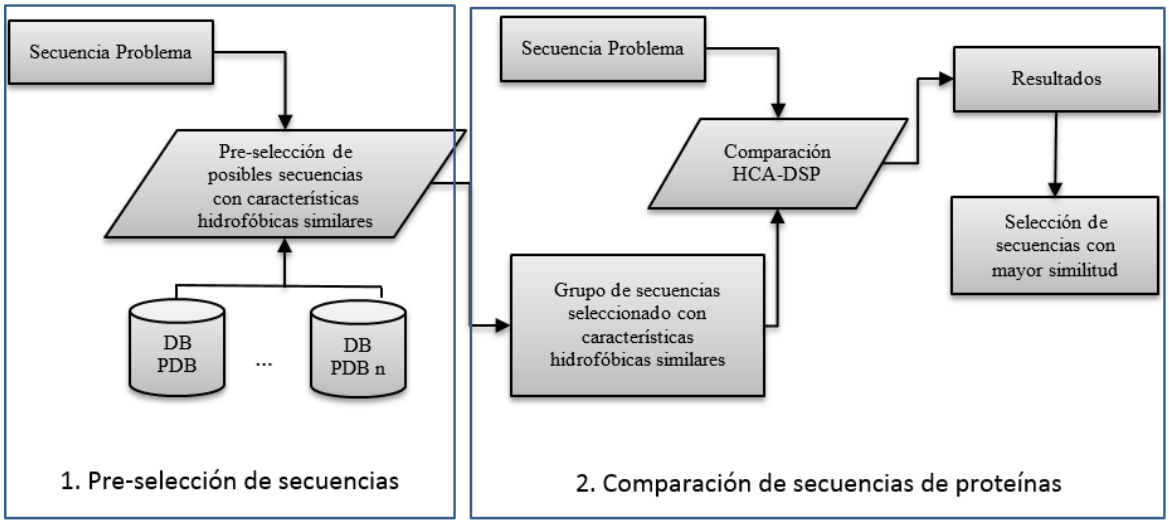

Figura 5. Modelo general para la anotación de secuencias de proteínas

3.1. Sistema de pre-selección de secuencias de proteínas

El sistema de pre-selección de secuencias involucra técnicas de codificación e inteligencia artificial. La finalidad es utilizar la base de datos PBD [5], por ser la más completa con información estructural y funcional de proteínas que existe actualmente, para hacer una clasificación rápida de las proteínas conocidas y utilizar esta nueva base de datos clasificada como método de discriminación. De esta manera es posible reducir el espacio de muestreo desde cuarenta mil secuencias a unos cuantos cientos.

El método consiste en crear, a partir de una red neuronal utilizando mapas auto-organizativos (SOM), un esquema de clasificación y búsqueda de secuencias previa a la comparación de proteínas. La clasificación se hace con base en las características de los *clusters* hidrofóbicos de cada proteína. Las proteínas se agrupan entonces de acuerdo con su distribución de *clusters* hidrofóbicos según los códigos del método HCA.

El funcionamiento del sistema de pre-selección se muestra en la Figura 6. La red SOM se entrenó utilizando las secuencias de proteínas que se encuentran en la base de datos PDB. Mediante la técnica de vector de composición de momento (VCM) y la expresión de las proteínas en los códigos P, Q y B, según el método HCA, se crea una representación de cada una de las proteínas que se agrupa en familias de acuerdo con su características hidrofóbicas.

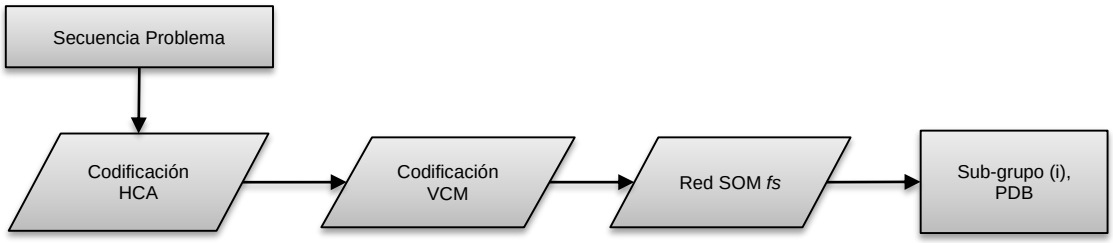

Figura 6. Diagrama de agrupación de secuencias con base a su contenido estructural lineal y HCA

3.1.1. Codificación HCA y clasificación de acuerdo con la red SOM

El proceso de pre-clasificación de secuencias inicia tomando como referencia el PDB como base de datos para crear grupos con características HCA similares. Este proceso inicia con la simplificación de la base de datos, eliminando aquellas secuencias que se encuentran repetidas [34]. Posteriormente 67379 secuencias de proteínas del PDB se codifican de acuerdo con los algoritmos P, B, Q del método HCA. La técnica de codificación es VCM, la cual permite entregar la información a la red SOM para su entrenamiento.

3.1.2. Vector composición de momento (VCM)

Para realizar la agrupación de secuencias utilizando una SOM es necesario extraer información de la cadena lineal de aminoácidos y transformarla en datos numéricos o vectores que describen el contenido de la secuencia, para lograr este propósito se utiliza la técnica de Vector Composición de Momento VCM. Esta técnica permite crear un vector representativo para cada secuencia de proteínas, el cual contiene información sobre la composición y la posición de los aminoácidos, de manera que se establece una relación funcional con el contenido estructural. Esta transformación permite que no existan dos secuencias con diferente contenido estructural e igual vector composición de momento[35].

Para el cálculo de este vector para cada proteína, se toma una secuencia *O* perteneciente a la base de datos del PDB. Las siguientes son las variables: A*i* es el aminoácido *i*, cuando los aminoácidos se ordenan de la siguiente forma: A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y, adicionando la codificación HCA V,U,D,M, #, 0 (en donde # indica agrupaciones de longitud 1, y 0 para aminoácidos que no pertenecen a una agrupación hidrofóbica) y el código binario 0,1 se obtienen los elementos (x_1, x_2,. . ., x_i,… x_{28}) del vector composición de *O*. Para un entero k ≥ 0, se define como k-ésimo vector momento de orden ($x_1^{(k)}$, $x_2^{(k)}$,..., $x_{28}^{(k)}$) como:

$$X_i^{(k)} = \frac{1}{N(N-1)\dots(N-k)} \sum_{j=1}^{k_i} n_{ij}^k \qquad (1)$$

Para *i* = 1,2,...28, donde N es la longitud de la cadena de aminoácidos, n_{ij} la j-ésima posición del i-ésimo aminoácido y k_i el número total del i-ésimo aminoácidos en la secuencia. El vector de momento de orden $x_i^{(0)}$ indica el vector composición. A partir de la ecuación se obtiene la matriz de momento para la proteína O para todos los *k*[35]. Dada la complejidad computacional para secuencias de gran longitud el vector composición de momento se representa por la concatenación del vector de momento $x_i^{(0)}$ y $x_i^{(1)}$ de la forma:

$$(\boldsymbol{X_1}, \boldsymbol{X_2}, \dots, \boldsymbol{X_{28}}, \boldsymbol{X_1^{(1)}}, \boldsymbol{X_2^{(1)}}, \dots, \boldsymbol{X_{28}^{(1)}}) \qquad (2)$$

Con este vector representativo, se garantiza que toda la información de todos los aminoácidos en la cadena es tenida en cuenta. A su vez, se pueden clasificar las proteínas de acuerdo con sus características hidrofóbicas en familias de compuestos de menor cardinalidad que la base de datos completa. Con cada vector único para una secuencia y todo el PDB codificado se utiliza la técnica de mapas auto-organizado SOM para la clasificación en subgrupos con características similares [36], ver Figura 7.

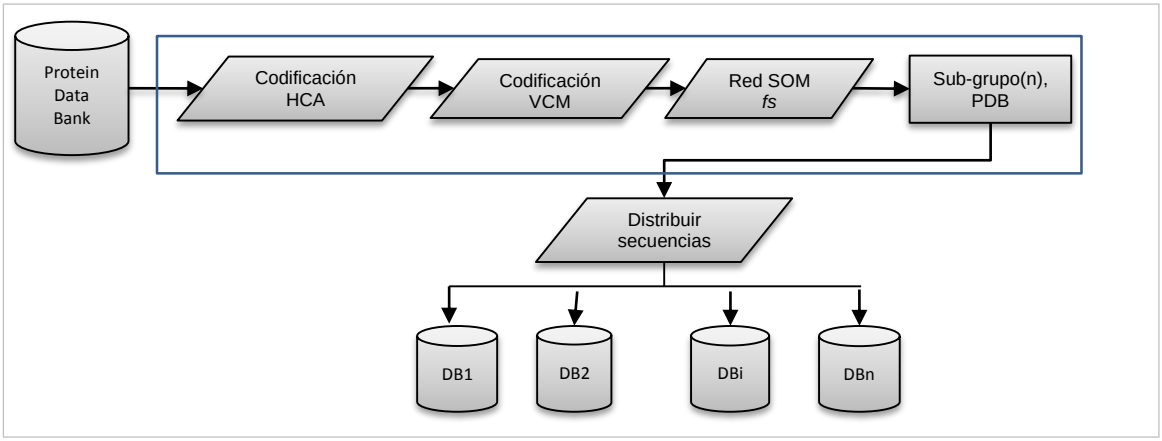

Figura 7. Grafico codificación del PDB en subgrupos de acuerdo a su contenido estructural lineal y la información HCA de la secuencia.

3.1.3. Redes SOM

Las redes SOM (self-organizing map) o mapa auto-organizado, es un modelo de red neuronal no supervisada que se ha utilizado con éxito en una amplia variedad de aplicaciones en la ingeniería. El objetivo del algoritmo es tomar vectores prototipos que representen el conjunto de datos de entrada y al mismo tiempo realizar una aplicación continua desde el espacio de entrada a una red.

Esta red consiste en una capa de entrada y una capa de Kohonen. La capa de Kohonen esta usualmente diseñada como un arreglo bidimensional de neuronas

que mapea entradas N-dimensionales a dos dimensiones, preservando el orden topológico [37]. La red es entrenada con un algoritmo de aprendizaje competitivo no supervisado, en donde la red debe ser alimentada con un gran número de muestras de entrenamiento que representen las características de los vectores esperados durante el mapeo.

Matemáticamente en una red SOM se define los siguientes componentes: el vector de entrada $x = (x_1, x_2, \dots x_p)'$ (vector de entrenamiento), $w = (w_{l1}, w_{l2}, \dots w_{lp})'$ el vector de pesos asociado con el nodo l donde w_{lj} indica el peso asignado a la entrada x_j al nodo l , donde se tienen k nodos y p es el número de variables. Cada objeto de del conjunto de datos de entrada es presentado a una red en orden aleatorio. La capa de aprendizaje de Kohonen es un algoritmo que encuentra el nodo más próximo en cada caso de entrenamiento y mueve el nodo ganador cerca al caso de entrenamiento [36]. El nodo es movido cierta proporción de distancia entre este y el caso de entrenamiento. La proporción está dada por la rata de aprendizaje.

Para cada objeto j en el conjunto de entrada, la distancia $d_i$entre el vector de peso y la señal de entrada es almacenada. Así la competencia comienza y el nodo con la distancia más pequeña es el ganador. Los pesos del nodo ganador son entonces actualizados utilizando alguna regla de aprendizaje. Los pesos de los nodos no ganadores no se modifican, usualmente se utilizada la distancia Euclidiana para comparar cada nodo con cada objeto a pesar de que se podría utilizar otra métrica [36]. Una vez la red se encuentra entrenada, se obtiene un total de 900 grupos en los que se encuentran un aproximado de 70 secuencias por cada grupo. Cada secuencia de PDB es guardada en su respectivo grupo utilizando la red entrenada.

3.2. Resultados y Análisis

Para verificar que los resultados de nuestra propuesta de modelo de comparación sistemático de proteínas funciona adecuadamente, se utilizó la secuencia de la hemoglobina alfa humana, se codificó utilizando HCA, se buscó la afinidad con una familia de secuencias, según la red SOM entrenada y posteriormente se comparó contra cada una de las secuencias en esa familia de proteínas, utilizando la metodología propuesta por Silva [27], ver Figura 8.

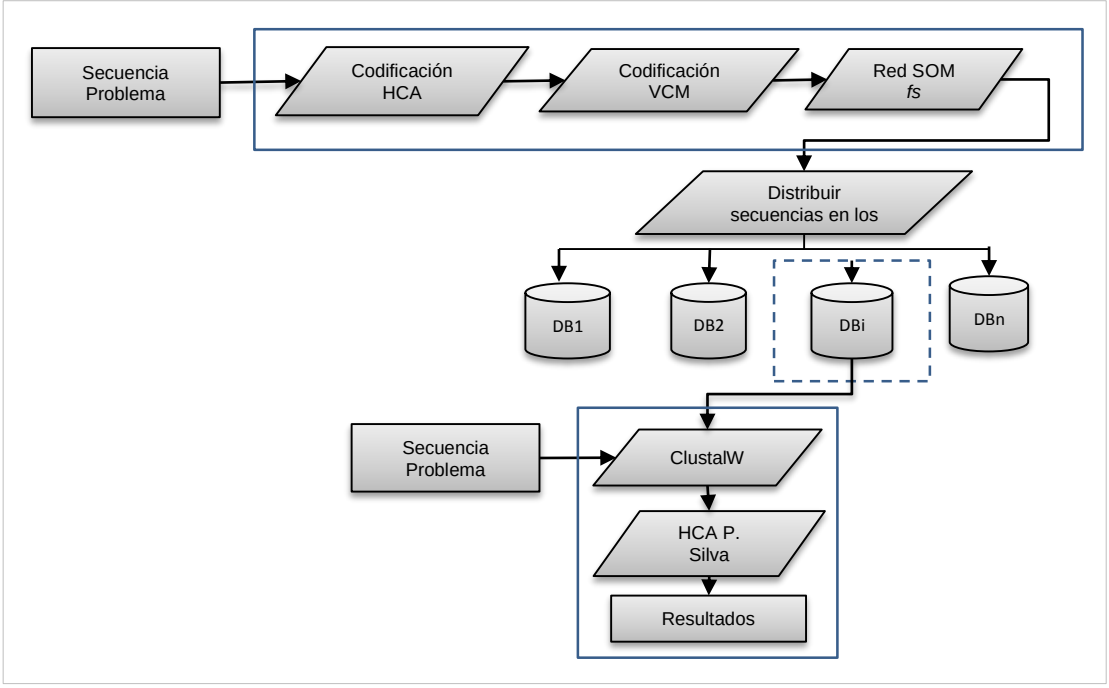

Figura 8. Representación del sistema para el análisis de una secuencia objetivo con un análisis final entre el grupo seleccionado utilizando HCA de P. Silva

La secuencia seleccionada fue analizada utilizando la propuesta de P. Silva en donde se indica la similitud en secuencia de acuerdo con su contenido hidrofóbico, y el análisis secuencias BLAST. El enfoque de Silva permite la detección de agrupaciones similares de residuos hidrofóbicos en dos secuencias, utilizando un valor de medición de la relación de alineamiento de los aminoácidos hidrofóbicos (HCA Score). Para evaluar el alineamiento del algoritmo, Silva asigna un valor de umbral para la medición acertada sin exceso de gaps, las secuencias que se encuentren sobre el umbral serán tenidas en cuenta.

Para la secuencia Hemoglobina Humana – Alfa, Uniprot P69905 se encontró como grupo de semejanza el número 300 de los 900 grupos en los que se encuentra clasificado el PDB. Los resultados de la comparación con el grupo se presentan en la tabla 9. El análisis con BLAST que se presenta para las 5 primeras secuencias con alto porcentaje de identidad.

Secuencia	BLAST % Identidad	E- Value	HCA-Score	Umbral
3ia3	100%	9e-105	1.00	Above threshold
1o1o	99%	6e-103	1.00	Above threshold
1rvw	99%	9e-103	1.00	Above threshold
1y0c	99%	9e-103	1.00	Above threshold
4mqc	99%	2e-103	1.00	Above threshold

Tabla 2. Descripción y resultados de la secuencia P69905 contra el grupo 300 de la clasificación del PDB.

El grupo contiene un total de 60 secuencias de las cuales se obtiene un mínimo porcentaje de identidad de 41% con e-value de 7e-35, 95% del grupo de secuencias posee un valor de identidad superior al 50% y poseen un HCA-score superior a 0.7 sobre el umbral.

4. Planteamiento del problema

La predicción de similitud y homología de secuencias de proteínas ha sido ampliamente estudiada en los últimos años [29,35]. Debido a la aparición de metodologías de secuenciación de genomas, cada vez más sofisticados y precisos[38], la cantidad de información sobre nuevas proteínas es abrumadora. En muchos casos, la información que se obtiene de estos genomas no se explota adecuadamente debido a la incapacidad que tenemos actualmente para determinar la estructura y función de muchas proteínas contenidas en los genomas. Hasta ahora los métodos más utilizados para buscar homología de proteínas son los de comparación lineal. Se ha demostrado que estos métodos no son adecuados para encontrar homología cuando la identidad entre proteínas oscila entre el 25 y 35% o zona de penumbra (*twilight zone*[1]).

Para abordar el grupo de secuencias en el *twilight zone* se utilizó el método de comparación HCA [4,40]. Recientes estudios han demostrado que esta metodología de comparación de proteínas en 2D es útil para encontrar homología entre proteínas que se encuentran en la zona de penumbras [4,5,39]. Sin embargo la utilidad de esta técnica está limitada por la experticia de los investigadores en su utilización. No existe actualmente una metodología robusta que permita realizar la comparación de proteínas de manera sistemática utilizando la herramienta HCA.

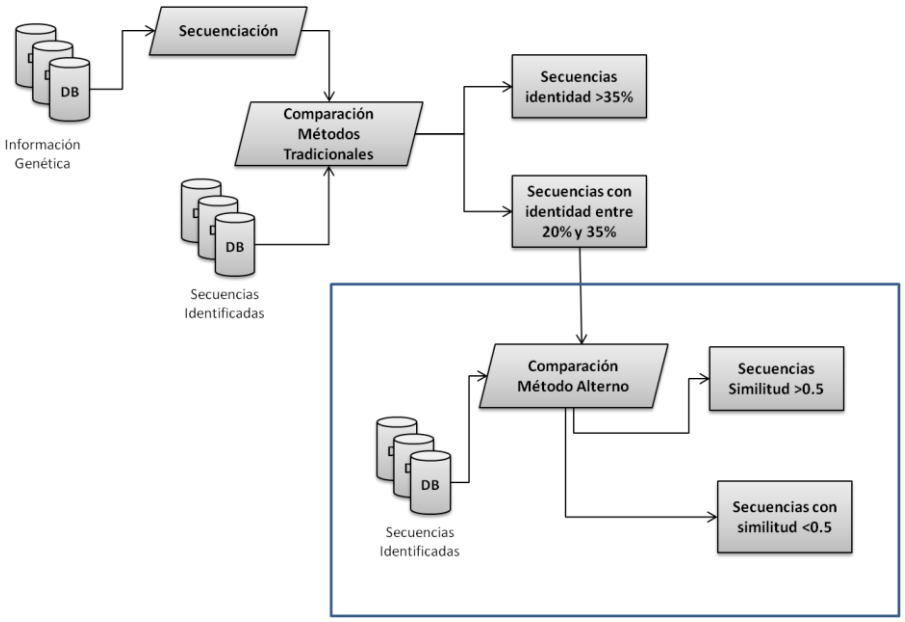

Figura 9. Sistema orgánico de comparación de proteínas utilizando HCA.

En este trabajo se propone un método alterno de comparación de proteínas con enfoque en aquellas secuencias que presentan baja identidad, ver Figura 9. El desarrollo de este trabajo se centró en la aplicación de la transformada discreta de *Wavelet* DWT para la comparación de secuencias utilizando HCA de forma automática. El método consta de varias etapas entre las cuales se incluyen: la codificación HCA de la secuencia de las proteínas para identificar agrupaciones hidrofóbicas, la caracterización de las agrupaciones para su transformación en una señal numérica, el procesamiento de las señales con la DWT y un posterior análisis de correlación, para detectar similitud en secuencia. Cada uno de estos pasos es descrito a continuación en la metodología.

5. Metodología

El método HCA de Gaboriaud [2] demuestra que secuencias con patrones de distribución muy similares en los residuos hidrofóbicos (detectados en la representación helicoidal 2D de la secuencia de la proteína) son frecuentemente homologas estructurales [41]. El sistema de comparación propuesto toma como enfoque el análisis HCA para analizar aquellas secuencias en el *Twilight zone.*

- El proceso de comparación automática de secuencias de proteínas propuesto involucra varios pasos: La codificación de las proteínas en HCA, en este paso se involucra el alineamiento de secuencias, la identificación de los *cluster* hidrofóbicos y su codificación en Q y B.

- El pre-procesamiento de la secuencia que implica crear el vector la información HCA de la proteína y asignarle una señal *f(x)* teniendo en cuenta la secuencia de *clusters* hidrofóbicos y un peso asociado con su valor EIIP.

- El proceso de comparación utilizando la trasformada discreta de *Wavelet* y el análisis de correlación.

- El análisis de resultados de acuerdo a la validación utilizando como referencia la base de datos del Balibase, tomando los mejores resultados y la visualización utilizando HCA.

Estos pasos se describen en la Figura 10. En primer lugar, realizar el alineamiento y codificación HCA para transformar las agrupaciones hidrofóbicas de la secuencia de la proteína en una señal numérica sin perder su información funcional o estructural, seguido del análisis multi-resolución con la DWT, para ello se utiliza una *Wavelet* madre y un nivel de descomposición determinado, y por último el análisis de correlación cruzada entre las dos secuencias de estudio. Las etapas de la comparación se describen a continuación.

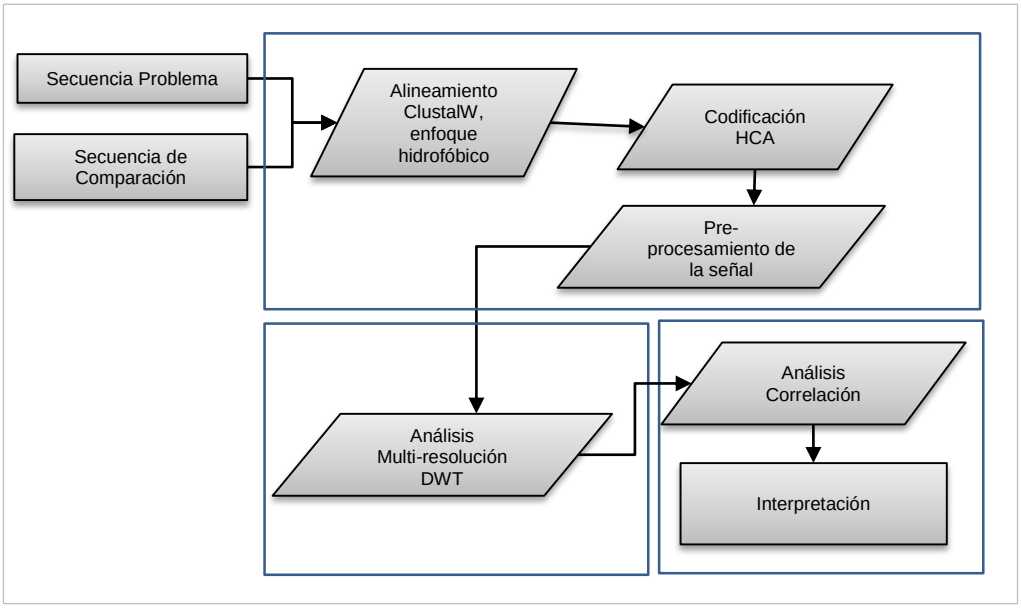

Figura 10. Diagrama de flujo del sistema automático de comparación de secuencias de proteínas utilizando HCA y DWT

5.1. Alineamiento de secuencias

En esta etapa se alinean las secuencias de acuerdo con una matriz de sustitución donde los aminoácidos hidrofóbicos y la prolina tienen un peso de 1, si y solo si se encuentran en la misma posición en la secuencia, mientras que los demás

aminoácidos tienen un peso de 0.7. Como se aprecia en la Figura 11. Esta matriz y la secuencia lineal de la proteína son las entradas del algoritmo ClustalW [42], el que está encargado de determinar el grado de identidad entre las dos proteínas. Como se observa en la matriz todas las iteraciones fuera de la diagonal son iguales a 0, excepto para los aminoácidos hidrofóbicos, cuyos intercambios por aminoácidos del mismo conjunto se muestran equivalentes [27].

	"V"	"I"	"L"	"F"	"Y"	"W"	"M"	"A"	"B"	"C"	"D"	"E"	"G"	"H"	"K"	"N"	"P"	"Q"	"R"	"S"	"T"	"X"	"Z"	"*"
"V"	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
"I"	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
"L"	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
"F"	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
"Y"	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
"W"	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
"M"	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
"A"	0	0	0	0	0	0	0	0.7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
"B"	0	0	0	0	0	0	0	0	0.7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
"C"	0	0	0	0	0	0	0	0	0	0.7	0	0	0	0	0	0	0	0	0	0	0	0	0	0
"D"	0	0	0	0	0	0	0	0	0	0	0.7	0	0	0	0	0	0	0	0	0	0	0	0	0
"E"	0	0	0	0	0	0	0	0	0	0	0	0.7	0	0	0	0	0	0	0	0	0	0	0	0
"G"	0	0	0	0	0	0	0	0	0	0	0	0	0.7	0	0	0	0	0	0	0	0	0	0	0
"H"	0	0	0	0	0	0	0	0	0	0	0	0	0	0.7	0	0	0	0	0	0	0	0	0	0
"K"	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.7	0	0	0	0	0	0	0	0	0
"N"	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.7	0	0	0	0	0	0	0	0
"P"	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
"Q"	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.7	0	0	0	0	0	0
"R"	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.7	0	0	0	0	0
"S"	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.7	0	0	0	0
"T"	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.7	0	0	0
"X"	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.7	0	0
"Z"	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.7	0
"*"	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figura 11. Matriz de sustitución aplicada en ClustalW para la identificación de zonas hidrofóbicas

El alineamiento de las secuencias se realiza con la finalidad de identificar las posibles zonas hidrofóbicas en común entre las secuencias comparadas, para lo que se utiliza una matriz de sustitución como PAM o BLOSUM [43]. Estas matrices nos permite asignar una puntuación a cada posible sustitución o la conservación posible de la cadena de aminoácidos [44] .

Un ejemplo del alineamiento obtenido se observa en la Figura 12. En este caso se realiza la comparación de dos secuencias, la hemoglobina humana-alfa y la Lupin leghemoglobin- alfa con porcentaje de identidad <15%.

```
>P1;sp|P69905|HBA_HUMAN

-MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFD----LSHGSAQ
VKGHGKKVADALTNAVAHVDD-----MPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTL
AAHLPAEFTPAVHASLDKFLASVSTVLTSKYR---
*
>P1;sp|P02240|LGB2_LUPLU

MGALTESQAALVKSSWEEFNANIPKHTHRFFILVLEIAPAAKDLFSFLKGTSEVPQNNPE
LQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKG-VADAHFPVVKEAILKTI
KEVVGAKWSEELNSAWTIAYDELAIVIKKEMNDAA
*
```

Figura 12. Representación de la información obtenida del alineamiento entre la Secuencia hemoglobina humana-alfa y la Lupin leghemoglobin- alfa utilizando la matriz de alineamiento propuesta.

5.2. Codificación en secuencia utilizando HCA y el EIIP

La transformación de la secuencia en una señal numérica, requiere de 2 pasos relevantes: Por un lado extraer las agrupaciones en código B y Q, ver Figura 13, y por el otro asignar un peso a cada aminoácido, relacionado con sus propiedades fisicoquímicas, para obtener una señal numérica.

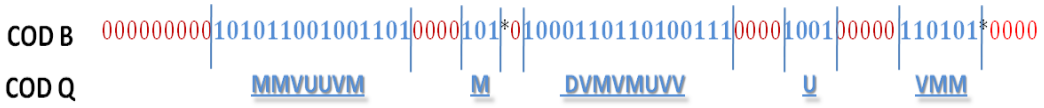

Figura 13. Representación de la codificación Q y B de una secuencia hipotética

Cada secuencia en HCA posee n cantidad de agrupaciones hidrofóbicas. Cada agrupación de aminoácidos $\{ Grupo_1, Grupo_2, \dots Grupo_i, \dots Grupo_n \}$ tiene una representación en Q: 11(V) ,101(M) ,1001(U) y 10001(D) como se explicó en 2.3, conformadas por aminoácidos hidrofóbicos y no hidrofóbicos. Identificadas las zonas hidrofóbicas y codificadas en Q, es necesario asociar un peso específico para la codificación en una señal posterior análisis utilizando la DWT.

Cada amino ácido en una estructura lineal puede ser representado por diferentes valores numéricos que indican propiedades y características físicas que son relevantes para la actividad biológica. Existen modelos como el c-p-v, que representa la composición, polaridad y volumen molecular de los aminoácidos de la secuencia [45], el modelo EIIP (el potencial de interacción ion-electrón) que representa el promedio de estados de energía de todos los electrones de valencia en cada aminoácidos [12], o la hidrofobicidad de cada aminoácido Kyte-Doolittle [46]. Para este trabajo decidimos utilizar EIIP porque discriminan los aminoácidos de acuerdo con una propiedad física específica, por la disponibilidad de los datos y porque en trabajos anteriores se ha demostrado que son suficientes el análisis de secuencias con tratamiento digital de señales [12,23,43].

Mediante la introducción de los valores del EIIP de todos los aminoácidos, cada proteína se puede representar como un vector numérico compuesto de respuestas funcionales, característica importante del método propuesto. Cada aminoácido independiente de su posición en la secuencia se puede representar por un único valor (Ver tabla 2). Los valores numéricos se derivan de la polarización de los aminoácidos como consecuencia de la movilidad de los electrones en la proteína, la cual es relevante para su actividad biológica [47]. Una vez que se logra un mapeo numérico para una secuencia de proteína, esta puede ser tratada como una señal [46].

Valores EIIP para los aminoácidos			
Aminoácido	**EIIP**	**Aminoácido**	**EIIP**
Leu	0.0000	Tyr	0.0516
Ile	0.0000	Trp	0.0548
Asn	0.0036	Gln	0.0761
Gly	0.0050	Met	0.0823
Val	0.0057	Ser	0.0829
Glu	0.0058	Cys	0.0829
Pro	0.0198	Thr	0.0941
His	0.0242	Phe	0.0946
Lys	0.0371	Arg	0.0959
Ala	0.0373	Asp	0.1263

Tabla 3. Descripción de los valores EIIP para cada uno de los aminoácidos, adaptado de [12]

Cada uno de los valores asignados a cada aminoácidos es normalizado para su posterior análisis de señales, ver Anexo numeral A. Para cada agrupación en Q, se calcula el peso a asignar con el cálculo de la combinatoria de todas las posibles conformaciones en la cadena contemplando aminoácidos hidrofóbicos y no hidrofóbicos. Los valores promedios calculados para cada posible conformación de la codificación Q se describen en la tabla 4, en dónde n corresponde a los aminoácidos hidrofóbicos y los m a los aminoácidos.

Clusters hidrofóbicos codificación Q y valores EIIP						
Código B	Código Q	Combinaciones	Peso calculado de cada combinatoria	Peso para cada clúster utilizando EIIP. (Valor normalizado)	Peso promedio para dos aminoácidos en Q	Peso para cada clúster utilizando EIIP. (Valor normalizado)
11	V	*n*×*n*	0.082571	-1,161895004	(D / V) y (U/M)	1,45717E-12
101	M	*n*×*m*×*n*	0.135725	-0,387298335	(D / M)	0,387298335
1001	U	*n*×*m*×*m*×*n*	0.188879	0,387298335	(D / U)	0,774596669
10001	D	*n*×*m*×*m*×*m*×*n*	0.242032	1,161895004	(V / M)	-0,774596669
					(U / V)	-0,387298335

Tabla 4. Descripción de los valores EIIP calculados para 4 agrupaciones Q con HCA

Con los valores calculados en la tabla 4 se construye un nuevo vector numérico para cada secuencia. La señal final codificada se conformará solo de aquellas zonas que tengan agrupaciones hidrofóbicas comunes, ver Figura 14.

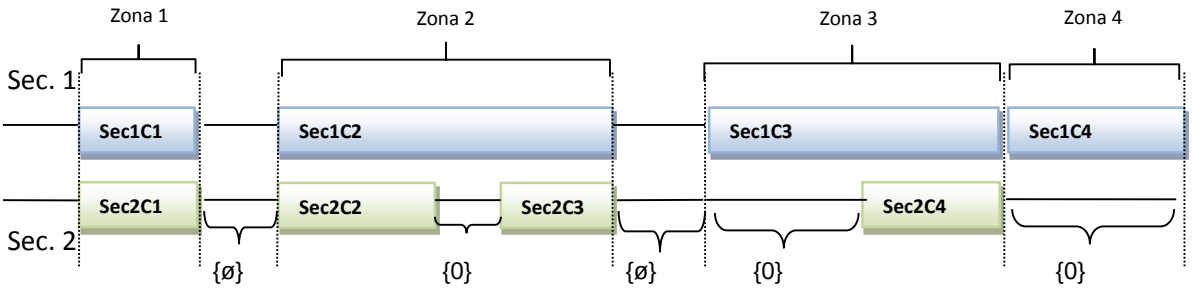

Figura 14. Representación de la selección de las agrupaciones hidrofóbicas que se pueden presentar en la transformación a una señal numérica para las proteínas

Se identifican 4 zonas específicas:

- Zona 1: es aquella en donde las agrupaciones hidrofóbicas coinciden en tamaño, los dos segmentos son codificados y quedan de la misma longitud.
- Zona 2: es aquella donde no coincide ninguna zona hidrofóbica, estos segmentos de aminoácidos serán ignorados en la secuencia.
- Zona 3: es aquella donde varias agrupaciones no coinciden en longitud, se puede dar el caso en que dos agrupaciones coincidan con una sola agrupación de comparación, las zonas sin coincidencia serán reemplazadas con el valor de cero. Y por último,
- Zona 4: es zona aquella donde una agrupación de longitud superior a otro contenido, estas zonas sin coincidencia serán completadas con el valor de cero.

Así, al final se tendrá dos vectores con la información final codificada para la comparación. Los dos vectores siempre serán de igual longitud, independiente de la longitud original de la cadena de aminoácidos de cada proteína. La longitud de los vectores dependerá de la cantidad de agrupaciones hidrofóbicas en secuencia que posean las secuencias de interés, ver Figura 15.

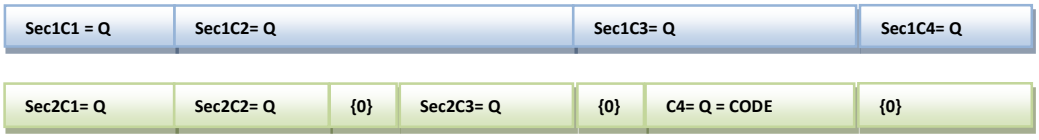

Figura 15. Representación de la información tomada de una secuencia con la nueva codificación

El código B de cada de cada una de las agrupaciones permite encontrar la cantidad de aminoácidos que pertenecen a cada una de las codificaciones Q que conforman una agrupación hidrofóbica, es decir, por cada aminoácido que pertenece a la formación de una de las agrupaciones Q será creada la nueva cadena de símbolos en Q. Si un aminoácido pertenece a dos formaciones de las agrupaciones Q, el valor

en esa posición será reemplazado por el promedio calculado las dos formaciones Q a las que pertenece, ver Figura 16.

Figura 16. Representación de una identificación de componentes de una agrupación hidrofóbica

Cuando se identifica la cadena de símbolos de todas las agrupaciones Q con la propuesta de codificación, es posible reemplazar cada uno de los valores numéricos calculado en las Tabla 3. Una vez se reemplazan todos los valores numéricos, se obtienen los dos vectores numéricos para la comparación utilizando la DWT y al análisis de correlación cruzada. A continuación de describe el algoritmo con todos los procesos involucrados en el proceso de comparación, ver Figura 17.Algoritmo de Comparación.

Entradas *Sec1, Sec2*

Sea $AA = \sum\{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}$ los aminoácidos que conforman cada una de las secuencias.

Salidas *Resultado*

Alineamiento de secuencias*Sec1, Sec2*

Sec1' , Codificación Q de *Sec1, Sec2* , sea *Q= {U,V,M,D}*

Sec2', Codificación B de *Sec1, Sec2,* sea *B= {0,1}*

S_1= Crear nueva codificación para *Sec1,*

- $S_1=\{s_11, s_12, s_12, \ldots s_1n\}$
- *n= Longitud de* S_1

  Sea $s_11$el primer componente del primer *clúster* contenido en *Sec1* en la nueva codificación numérica.

S_2= Crear nueva codificación para *Sec2,*

- $S_2=\{s_21, s_22, s_22, \ldots s_2n\}$
- *n= Longitud de* S_1

  Sea s_21 el primer componente del primer *clúster* contenido en *Sec2* en la nueva codificación numérica.

Análisis multi-resolución DWT

- Nivel de descomposición 4
- Selección de una Wavelet , 37 tomadas de prueba

$S_1' = \{Ca4,Cd4,Cd3Cd2,Cd1\}$

$S_2' = \{Ca4,Cd4,Cd3Cd2,Cd1\}$

Análisis de correlación

Resultados , Vector de correlación entre los coeficientes de descomposición en DWT para las dos secuencias de comparación

Resultados = {x1,x2,x3,x4,x5}

Variable	Descripción
Sec1	Archivo con las proteínas a comparar
Sec2	Archivo con el grupo de secuencias a comparar
Test	Archivo de alineamiento por par de secuencias a comparar
Clus1	Archivo con las agrupaciones HCA para la secuencia 1
Clus2	Archivo con las agrupaciones HCA para la secuencia 2
Resultado	Archivo con el vector de correlación superior a 0.5 en alguno de sus coeficientes, sino estará vacío.

Figura 17. Algoritmo para la comparación de secuencias de proteínas bajo el análisis HCA y la DWT

5.3. Transformada discreta de *Wavelet,* análisis multi-resolución

La transformada de *Wavelet* ha sido aplicada en múltiples trabajos relacionados con el análisis de secuencias de proteínas, alineamiento de secuencias, predicción estructural y análisis de similitud [12,29,30,43]. Con esta metodología las proteínas se representan por valores numéricos y se analizan utilizando un modelo físico-matemático.

En la literatura se puede encontrar que entre los modelos frecuentemente utilizados se encuentran la transformada de Fourier [28,31,44,45] y el análisis a través de la transformada de *Wavelet* [30,43,46,47]; al igual algunos nuevos enfoques con nuevas propuestas como el uso de estudio tiempo-frecuencia Wigner-Ville para encontrar sitios biológicamente activos para ciertas proteínas [46]. Sin embargo sigue siendo la transformada de *Wavelet* de mayor interés en el análisis de proteínas por ser más eficiente y rápida en la captura de la esencia de los datos que la transformada de Fourier [53].

La transformada *Wavelet* ha tenido diferentes enfoques en el estudio de proteínas; utilizando la transformada continua de *Wavelet* (CWT) se logró la identificación de estructuras secundarias [30], la detección y caracterización de motivos repetidos en una secuencia de proteína y en los datos estructurales [29], y la detección de sitios activos biológicamente potenciales en secuencias pero con la dificultad de que no revela componentes frecuenciales asociados a RRM y con la dificultad que existe para interpretar el espectrograma [46].

En otros trabajos utilizan la transformada discreta de *Wavelet* para la detección de estructuras secundarias [30], la predicción de núcleos hidrofóbicos [54], la predicción de tipos de membranas en proteínas [55], la detección de motivos sub-estructurales en la secuencia lineal [56], la predicción de la localización y topología de hélices en proteínas de membrana [46,53,54,56]. Esta metodología ha mostrado ser muy eficiente en la captura de componentes ocultos de datos biológicos, lo que permite una mejor relación entre las representaciones matemáticas y la función de los sistemas biológicos [53].

Con respecto a nuestro trabajo, DWT ha sido utilizada para la comparación de secuencias y la detección de similitud por otros investigadores con buenos resultados [12,23,46]. En este trabajo, una proteína es transformada en una señal $f(x)$; el valor del primer residuo x_1 corresponde al primer valor de la primera agrupación hidrofóbica en la secuencia, así para j agrupaciones hidrofóbicas se tendrán n puntos en secuencia dependiendo de la cantidad de agrupaciones en secuencia y su longitud. Así, dos secuencias de proteínas S_1 y S_2 respectivamente están compuestas de n puntos de datos en su codificación, así se tiene $S_1 = \{s_1 1, s_1 2, \dots, s_1 n\}$ y $S_2 = \{s_2 1, s_2 2, \dots, s_2 n\}$ para ser analizadas a través de la DWT, como las señales $F1(x) = \{x1, x2, .., xn\}$ y $F2(y) = \{y1, y2, \dots, yn\}$.

La transformada discreta de *Wavelet* consiste en la descomposición de una señal discreta $f(x) \in L^2(R)$ en un conjunto jerárquico de funciones ortogonales de aproximación y detalle, A_m y d_m respectivamente, siendo m el nivel de descomposición. Las funciones de aproximación contienen los componentes de baja frecuencia de la señal y pueden ser extraídos a partir de un filtro pasa-bajo (h). Para el caso de las funciones de detalle, contienen los componentes de alta frecuencia y pueden ser extraídos a partir de un filtro pasa alto $g(n)$, ver figura 18.

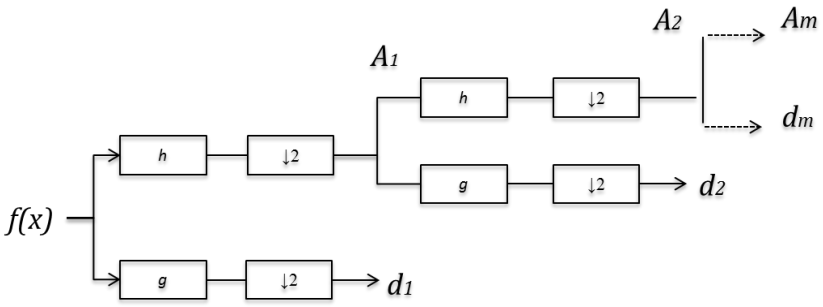

Figura 18. Diagrama de descomposición de una señal utilizando la DWT.

El proceso inicia desde el ingreso de la señal original a través de los dos filtros para obtener en el primer nivel de descomposición las señales A_1 y d_1 . A_1 es descompuesta nuevamente por los dos filtros para obtener A_2 y d_2. El proceso de descomposición se repite hasta el nivel de descomposición deseado (m).

Con todos los coeficientes de aproximación y detalle calculados es posible realizar un proceso inverso de reconstrucción de la señal. Este proceso involucra otro conjunto de filtros, uno pasa bajo h'y uno pasa alto g',ver figura 19. La selección tanto de h, g, h' y g' se define por la selección de la *Wavelet* madre que se utilice. Las *Wavelet* madre se encuentran clasificadas en familias; las *Wavelet*s asimétricas (familia Daubechies), las *Wavelet* simétricas (familia Biortogonal) y las casi simétricas (familia Coif).

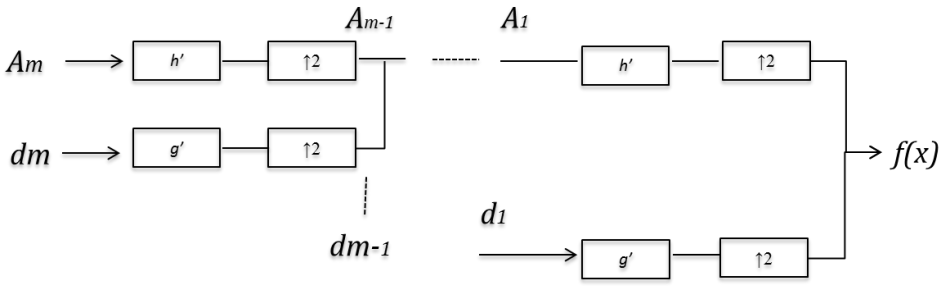

Figura 19. Diagrama de reconstrucción de una señal utilizando la DWT

Para todas las *Wavelet*s madre posible a utilizar en el análisis propuesto es necesario realizar una prueba de comparación, para seleccionar aquella que presente el mejor resultado de comparación para las secuencias de estudio. Los coeficientes obtenidos por cada *Wavelet* serán reconstruidos y comparados a través del análisis de correlación cruzada.

En el siguiente ejemplo se toman las secuencias hemoglobina humana-alfa y la Lupin leghemoglobin- alfa después del posterior alineamiento mostrado en la Figura 12. En la figura 20 y 21 se representan los coeficientes obtenidos para cada una de las señales respectivamente.

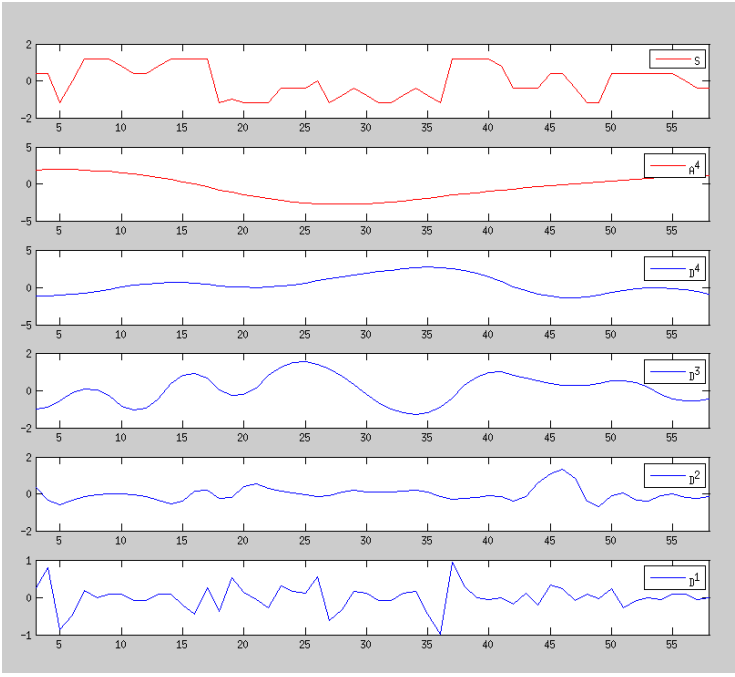

Figura 20. Coeficientes Wavelet. Representación de los coeficientes reconstruidos de detalle (D1, D2, D3, D4) y aproximación (A4) para los 4 niveles de descomposición utilizando la DWT con la Wavelet madre Bior3.1, (S) denota la señal original de la proteína transformada en sus agrupaciones hidrofóbicas y los valores calculados con el EIIP para la secuencia hemoglobina humana-Alfa.

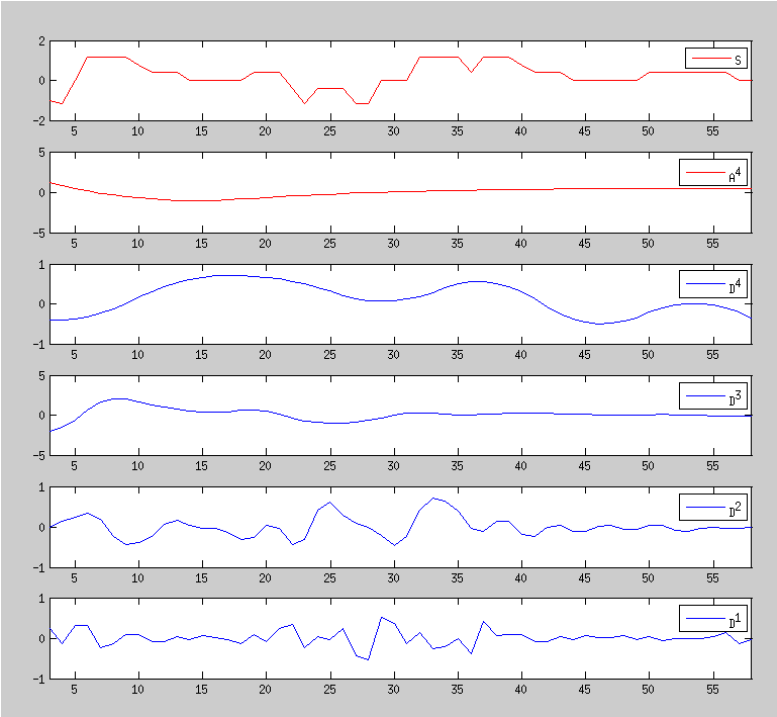

Figura 21. Coeficientes Wavelet. Representación de los coeficientes reconstruidos de detalle (D1, D2, D3, D4) y aproximación (A4) para los 4 niveles de descomposición utilizando la DWT con la Wavelet madre Bior3.1, (S) denota la señal original de la proteína transformada en sus agrupaciones hidrofóbicas y los valores calculados con el EIIP para la secuencia Lupin leghemoglobin- alfa.

El ejemplo se realiza para todo el sistema de comparación iniciando con el alineamiento, y los resultados de los coeficientes de detalle y aproximación para cada secuencia. La abscisa representa la posición de cada uno de los componentes hidrofóbicos para cada agrupación a lo largo de la secuencia y la ordinaria representa la magnitud de los coeficientes de la DWT. Para estas dos secuencias se calcula el vector de correlación para identificar el grado de similitud respecto a su contenido hidrofóbico entre ellas.

5.4. Análisis de Correlación

Los coeficientes de correlación cruzada se calculan en cada nivel para establecer y cuantificar la similitud entre las dos secuencias de proteínas comparadas. Utilizando la teoría de señales biomédicas[59], se considera que dos señales son fuertemente correlacionadas si el coeficiente de correlación supera $\pm$0.7,y débilmente correlacionadas si la correlación está en el rango $\pm$0,7 y $\pm$0,5, y sin correlación para valores inferiores a 0.5 [23]. Los coeficientes de correlación cruzada se definen como:

$$\rho^{12}(j) = \frac{\frac{1}{N}[\sum_{n=0}^{N-1} S_2\,(n) S_1(n-j)]}{\frac{1}{N}[\sum_{n=0}^{N-1} S_1^2\,(n) \sum_{n=0}^{N-1} S_2^2\,(n)]^{1/2}} \qquad j = 0, +1, +2, \ldots \tag{3}$$

Donde n es la longitud de la señal y j es el número de retrasos. El máximo valor absoluto del coeficiente de correlación para cada nivel de descomposición es asignado como el puntaje de similaridad para las dos proteínas en ese nivel. De la comparación de cada coeficiente se crea un vector de similaridad en secuencia a diferentes escalas. El máximo valor de correlación es 1 indicando 100% similitud [56].

Estudios previos han mostrado que las proteínas que se encuentran altamente relacionadas tienen una correlación fuerte, superior a 0.7 y aquellas distantemente relacionadas pero con funciones biológicas similares tienen una escala de correlación en algún nivel, y aquellas secuencias sin función biológica relacionada no presentan ninguna escala de correlación [12,23]. Estos tres rangos para la similitud en secuencia: fuertemente correlacionados, débilmente correlacionados y sin correlación se tomarán en la comparación propuesta. El proceso de comparación depende fundamentalmente de la codificación de la secuencia.

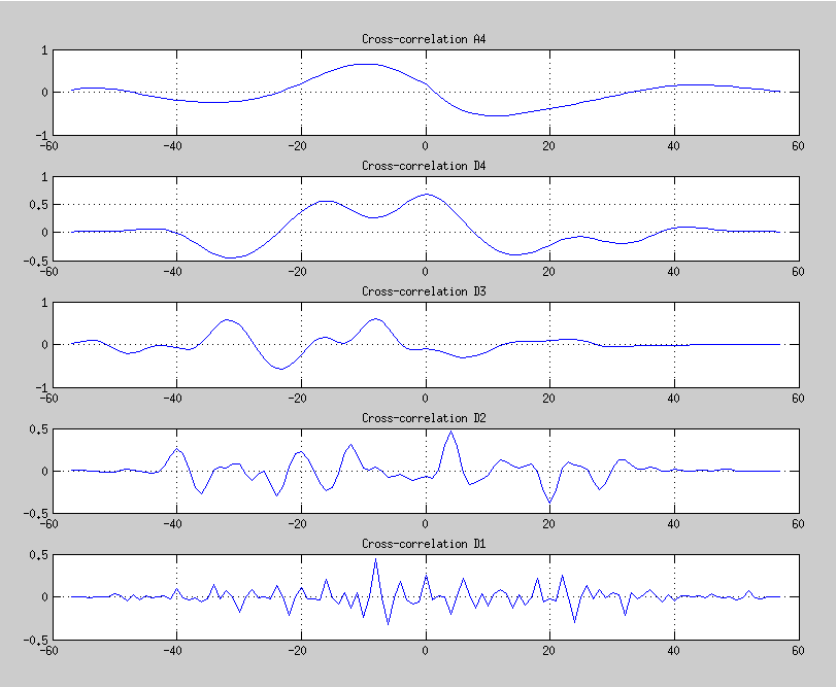

Figura 22. Coeficientes de correlación cruzada entre la secuencia la hemoglobina humana-Alfa y Lupin leghemoglobin- alfa. La abscisa es la posición de cada uno de los componentes de cada agrupación hidrofóbica en Q y la ordinaria es la magnitud de la correlación. Se utiliza la Wavelet Bior3.1.

La Figura 22 representa los coeficientes de correlación cruzada para el análisis de las secuencias. El vector de similitud que se obtiene es [0.6622 0.6811 0.6119

0.4748 0.4426] en el orden A4, D4, D3, D2, D1 respectivamente, donde se revela correlación alta en 3 de sus niveles esto nos indicar que existe similitud estructural entre estas dos secuencias, es necesario recordar que es necesario verificar el nivel de descomposición de la señal. En el estudio realizado en estas dos secuencias se identifica utilizando HCA, homología a través de sus *cluster* hidrofóbicos cubren el 65% asignando homología HCA de 80%[4]. En nuestro método es posible detectar esa similitud en secuencia a través del análisis de correlación entre los coeficientes.

Para la selección de los parámetros del análisis propuesto se realizó una prueba y una validación del trabajo utilizando como referencia la base de datos del BAliBase en diferentes casos.

6. Análisis de Resultados

Las pruebas experimentales realizadas en la presente investigación, fueron realizadas utilizando la base de datos BaliBase, base de datos estándar con información de secuencias con proteínas identificadas con estructura utilizada como banco de prueba para sistemas de alineación [60]. Tomando dos de sus agrupaciones se desea realizar la selección de la *Wavelet* madre para la comparación propuesta a un nivel de descomposición 4, encontrado en la literatura como la mejor aproximación de descomposición con la DWT [12, 23, 60] y la posterior validación con la *Wavelet* que presenta la mejor aproximación de similitud.

Las alineaciones en el BaliBase se clasifican por la longitud de la secuencia y la similitud principalmente [23]. La base de datos actualmente consta de 142 alineaciones de referencia, que contiene más de 1000 secuencias. Las alineaciones se dividen en cuatro conjuntos de referencia jerárquicos. De interés en las pruebas para la identificación de la *Wavelet* a seleccionar, se utiliza el grupo de referencia 3. Este grupo contiene 3 subgrupos de secuencias clasificados por su longitud, y posee secuencias con menos del 25% de identidad en residuos y secuencias altamente relacionadas, ver tabla 4.

Para la validación del método con la *Wavelet* madre seleccionada se toma como referencia el grupo Ref_1. Esta grupo contiene alineamientos de secuencias (menos de 6) equidistantes, es decir el porcentaje de identidad entre dos secuencias está dentro de un rango especificado. Todas las secuencias son de longitud similar, con lo no hay grandes inserciones o extensiones, ver Tabla 5. Tanto para el grupo Ref_3 y Ref_1 se describen las características de referencia en el BAliBase y los resultados obtenidos.

Referencia Ref -3	Cantidad de Secuencias	Tamaño	Max. % Identidad	Min. % Identidad
1idy	27	Short	81	1
1ubi	22	Short	70	5
2pia	20	Medium	70	13
kinase	25	Medium	100	14
1ajsA	28	Long	68	8
1pamA	19	Long	78	16

Tabla 5. Descripción de las secuencias tomadas del BaliBase Ref_3

En el experimento para la selección de la *Wavelet* se toman 6 grupos de secuencias con diferente longitud y diferente porcentaje de identidad del grupo Ref_3. Para cada grupo de secuencias se realiza el análisis multi-resolución utilizando la transformada discreta de *Wavelet*. Cada familia *Wavelet* posee su conjunto de funciones para diferentes señales, generando diferentes resultados. En este trabajo se probaron 37 *Wavelets*, DB1, DB2, DB3, DB4, DB5, DB6, DB7, DB8 Bior1.1, Bior1.3, Bior1.5, Bior2.2, Bior2.4, Bior2.6, Bior3.1, Bior3.3, Bior3.5, Bior3.7, Bior3.9, Bior4.4, Bior5.5, Rbio1.3, Rbio1.5, Rbio2.2, Rbio2.4, Rbio2.6, Rbio2.8, Rbio3. 1, Rbio3.3, Rbio3.5, Rbio3.7, Coif1, Coif2, Coif3, Coif4, Coif5, para encontrar aquellas que se ajusten en el análisis propuesto.

Todas las señales fueron analizadas a un nivel de descomposición 4, sin exceder el máximo valor de descomposición de acuerdo a la longitud de la señal, $\log_2 N$ siendo N la longitud de una señal. La Figura 16 contiene la información de los resultados obtenidos para cada familia *Wavelet*. El análisis de coeficientes se calcula en cada nivel de descomposición. Para el criterio de similitud en secuencia se toma como referencia aquellos pares de secuencias en las que se presentan al menos dos coeficientes con correlación débil o un coeficiente con correlación fuerte sobre el total de secuencias en el grupo.

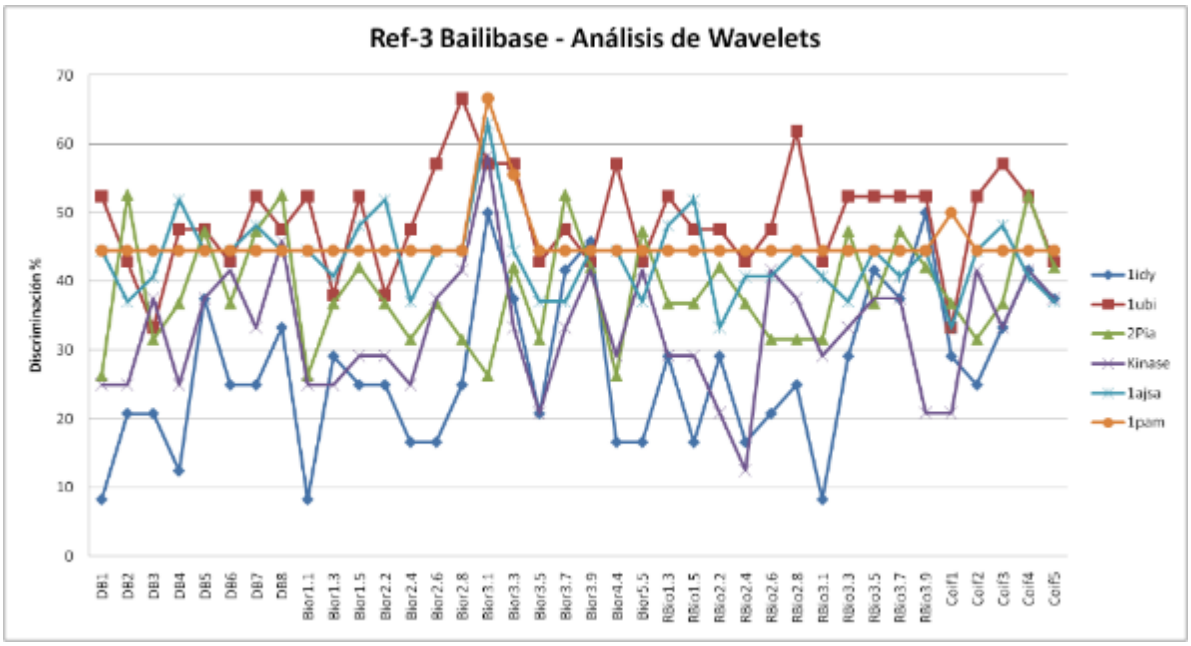

Figura 23. Resultados obtenidos para el análisis 6 grupos de secuencias tomados del BaliBase a un nivel de descomposición 4. El eje x representa cada una de las Wavelets analizada y el eje Y la discriminación a partir de las secuencias encontradas con similitud sobre la cantidad de secuencias en el grupo.

Secuencia	Wavelet	% Discriminación
1idy	Bior3.1	50%
	Bior3.9	50%
	RBio3.9	45.80%
ubi	Bior2.8	66.70%
	RBior2.8	61.90%
	Bior2.6, Bior3.1,Bior3.3,Bior4.4,Coif3	57.14%
pia	DB2,DB8,Bior3.7,Coif4	52.63%
kinase	DB8	45.83%
	DB6,Bior2.8,Bior3.9,Bior5.5,RBio2.6	41.66%
	Coif2,Coif5	
ajsa	Bior3.1	62.96%
	Bior2.2,RBio1.5,DB4	51.85%
pam	Bior3.1	66.66%
	Bior3.3	55.55%
	Coif1	50%

Tabla 6. Resultados obtenidos para los 6 grupos seleccionados del BaliBase Ref_1 para el estudio de coeficientes con correlación representativa

Las *Wavelets* representativas se evidencian para Bior2.8, Bior3.1 y Rbior2.8, ver Figura 23, tabla 6, en donde se obtiene los grupos con mayor cantidad de pares de secuencia con porcentajes de correlación relevantes. En la validación del método se toma la *Wavelet* Bior3.1 como base para analizar 6 grupos de secuencias del BaliBase Ref_1; Los grupos de secuencias tomados del Ref_1 se toman con promedio de identidad en diferentes rangos, ver tabla 7.

Grupo Ref _1	Cantidad de Secuencias	Max. Identidad	Min. Identidad	Promedio Identidad
1gpb	5	61	43	47
1fkj	5	52	38	44
1ycc	4	40	25	29
1ldg	4	30	23	27
1ajsA	4	23	11	15
3grs	4	22	11	14

Tabla 7. Descripción de las secuencias tomadas del BaliBase Ref_1 para el análisis de comparación propuesto

Este grupo de secuencias se analiza con tres métodos de comparación diferentes. El primero, es la comparación a través del método propuesto por Silva [27], con enfoque en el análisis hidrofóbico en secuencia. El segundo método se basa en el alineamiento de aminoácidos a nivel local, con mayor parecido en secuencia BLAST [62], y por último el enfoque propuesto en este trabajo.

El enfoque de Silva permite la detección de agrupaciones similares de residuos hidrofóbicos en dos secuencias, utilizando un valor de medición de la relación de alineamiento de los aminoácidos hidrofóbicos (HCA Score). Para evaluar el alineamiento del algoritmo, Silva asigna un valor de umbral para la medición acertada sin exceso de gaps, las secuencias que se encuentren sobre el umbral serán tenidas en cuenta. El valor del HCA score dará un aproximado en similitud hidrofóbica guía para analizar los resultados obtenidos por el método propio propuesto. Los resultados obtenidos por la herramienta HCA se presentan en la Tabla 8.

Referencia Ref-1	Longitud promedio Secuencias	Porcentaje de identidad BaliBase	HCA Score Máximo	HCA Score Mínimo	HCA Score Promedio	Umbral
1gpb	416	>35%	0.84	0.78	0.81	ABOVE TRESHOLD
1fkj	127	>35%	0.73	0.68	0.70	ABOVE TRESHOLD
1ycc	51	20% y 40%	No significant HCA score	No significant HCA score	No significant HCA score	----
1ldg	154	20% y 40%	0.65	0.60	0.62	ABOVE TRESHOLD
3grs	270	<25%	No significant HCA score	No significant HCA score	No significant HCA score	----
1ajsA	224	<25%	No significant HCA score	No significant HCA score	No significant HCA score	----

Tabla 8. Resultados obtenidos utilizando el análisis propuesto por P. Silva para el análisis de secuencias utilizado

En los resultados obtenidos, tres grupos poseen pares de secuencias con agrupaciones de alto contenido hidrofóbico, 1gpb, 1fkj y 1ldg. Estos tres grupos se analizan con la herramienta propuesta en esta investigación para establecer características que pueden ser asociadas a un grado de correlación. Al igual los 6

grupos se analizan con la herramienta tradicional BLAST para observar los diferentes resultados al contemplar todo el contenido de la proteína, y como los métodos alternos pueden aportar más información en el análisis de comparación.

BLAST es una herramienta utilizada para detectar posible homología en secuencia a partir del alineamiento local de aminoácidos. Utiliza un algoritmo heurístico, lo que no permite garantizar con exactitud sus resultados. Blast brinda un valor de identidad entre secuencias, y un e-valué que describe el número de hits que se puede esperar encontrar cuando se busca en un grupo de secuencias. Cuanto menor sea el valor de e-valué, más significante es el match, es importante tener en cuenta que este valor también se ve afectado por la longitud en secuencia, ya que las secuencias muy cortas tienen mayor probabilidad de encontrarse entre un grupo de secuencias.

Grupo Ref _1	Porcentaje de identidad BaliBase	Max. Identidad	Min. Identidad	Promedio Identidad	E-value
1gpb	>35%	49%	44%	47%	0.0
1fkj	>35%	54%	31%	39%	5e-26
1ycc	20% y 40%	35%	27%	31%	9e-21
1ldg	20% y 40%	31%	26%	29%	2e-43
1ajsA	<25%	No significant similarity found	No significant similarity found	No significant similarity found	No significant similarity found
3grs	<25%	No significant similarity found	No significant similarity found	No significant similarity found	No significant similarity found

Tabla 9. Resultados obtenidos utilizando un herramienta tradicional de comparación de proteínas.

Los resultados obtenidos con la herramienta BLAST no permiten encontrar identidad en aquellas secuencias que se encuentran por debajo del 25% indicado en el BaliBase, las secuencias con identidad superior se puede observar que tienen un aproximado de identidad semejante lo cual permite analizar los resultados de nuestro método, ver Tabla 9.

Para aquellas secuencias que presentan un porcentaje de identidad superior a 35% en el BaliBase se obtiene al menos un coeficiente de correlación fuerte (>0.7) o la mayoría de sus coeficientes con correlación débil (0.5).Es de comparar que el grupo 1gpb y 1fkj que obtuvieron alto valor de HCA_score poseen coeficientes de correlación superior a 0.5 en casi todas sus escalas ver Figura 24.

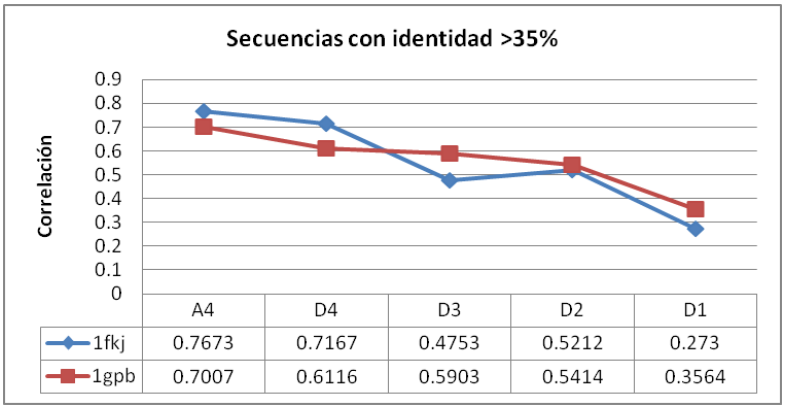

Figura 24. Coeficientes de correlación para secuencias con porcentaje de identidad superior a 35% de acuerdo a la información entregada en el BaliBase

Para obtener una representación del contenido hidrofóbico de este grupo de secuencias se realiza el gráfico HCA de las secuencias con mayor nivel de correlación, en este caso se toma como ejemplo las secuencias del grupo 1fkj35 (P45523 y P0A9L3) observar su alineamiento en secuencia y la cantidad de agrupaciones hidrofóbicas. Varios segmentos indican zonas hidrofóbicas que pueden ser relevantes en el análisis estructural de las dos proteínas.

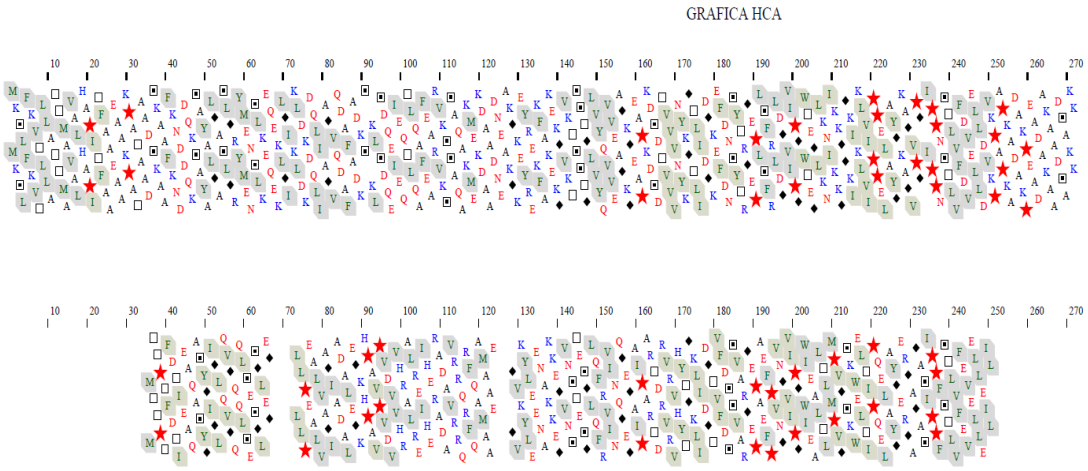

Figura 25. Representación HCA de las secuencias de proteínas pertenecientes al grupo 1fkj35 (P45523 y P0A9L3), Imagen obtenida mediante adaptación de software DrawHCA [63]**.**

Para analizar aquellas secuencias con porcentaje de identidad entre el 20% y el 40%, se observa que la comparación con el método de Silva, el grupo 1ldg contiene alto contenido hidrofóbico, y en nuestro método las dos secuencias contienen coeficientes de correlación altos. Cabe resaltar que es necesario verificar siempre en el procedimiento la longitud en secuencia y el alineamiento que tuvo el proceso para la comparación, ver Figura 26.

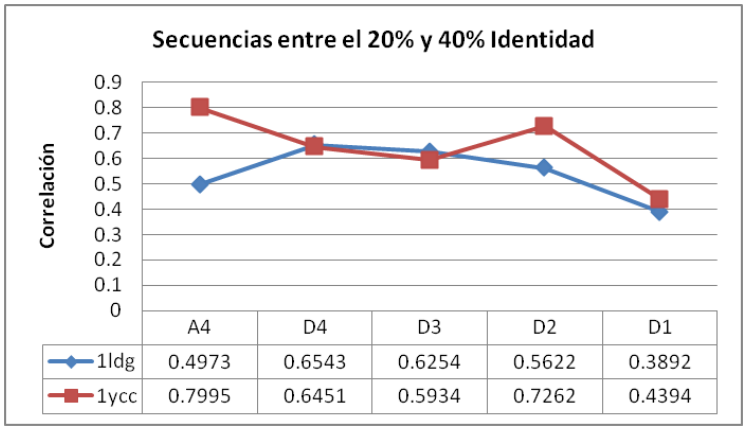

Figura 26. Coeficientes de correlación para secuencias con porcentaje de identidad entre el 20% y 40% de acuerdo a la información entregada en el BaliBase

Al igual para este grupo se realiza el gráfico HCA de las secuencias con mayor nivel de correlación, en este caso se toman las secuencias del grupo 1ldg (Q27743 y P14245) que presenta un HCA_score representativo. De los grupos hidrofóbicos identificados.

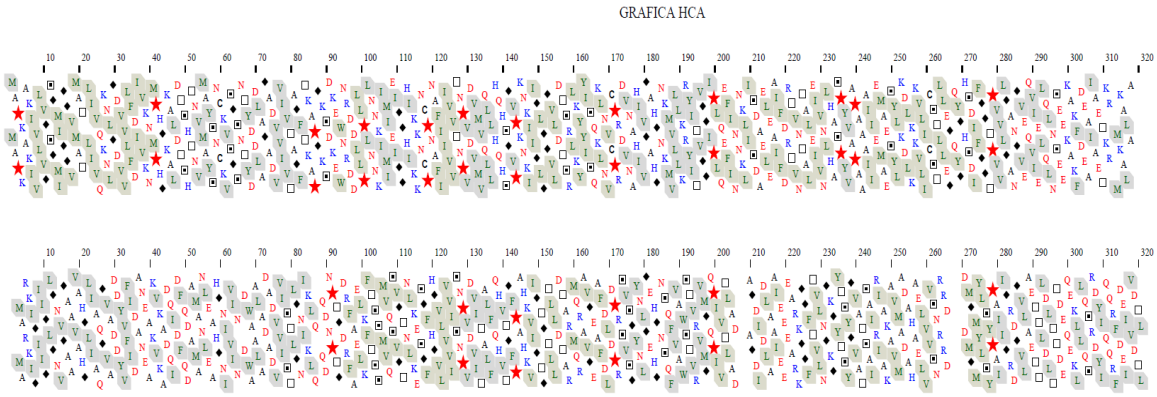

Figura 27. Representación HCA de las secuencias de proteínas pertenecientes al grupo 1ldg (Q27743 y P14245), Imagen obtenida mediante adaptación de software DrawHCA [63]**.**

Para aquellas secuencias con identidad inferior al 25%, se observan escalas con correlación superior a 0.5 en uno de sus niveles. Este indicador requiere de una verificación de la similitud detectada en el coeficiente indicado, verificar el alineamiento de secuencia y si es posible hallar similitud significativa, ver Figura 28.

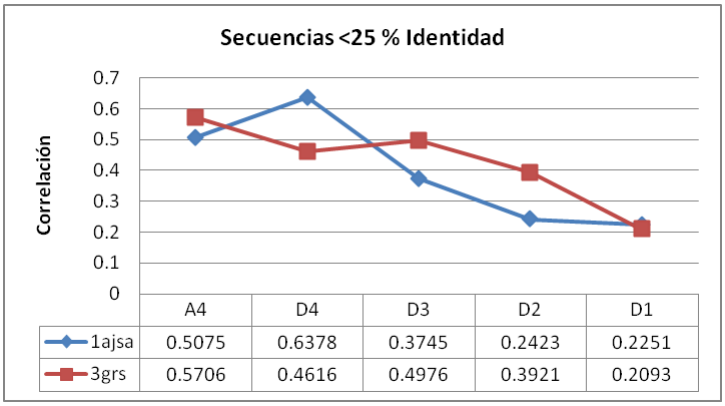

Figura 28. Coeficientes de correlación para secuencias con porcentaje de identidad inferior al 25% de acuerdo a la información entregada en el BaliBase

El gráfico HCA de las secuencias con identidad inferior al 25% con mayor nivel de correlación, en este caso se toma como ejemplo las secuencias del grupo 1ajsa (P45523 y P0A9L3) permitirá al experto en HCA un análisis profundo para detectar similaridad en aquella secuencia que muestre algún grado de correlación representativo, o descartar información no relevante en las conformaciones hidrofóbicas.

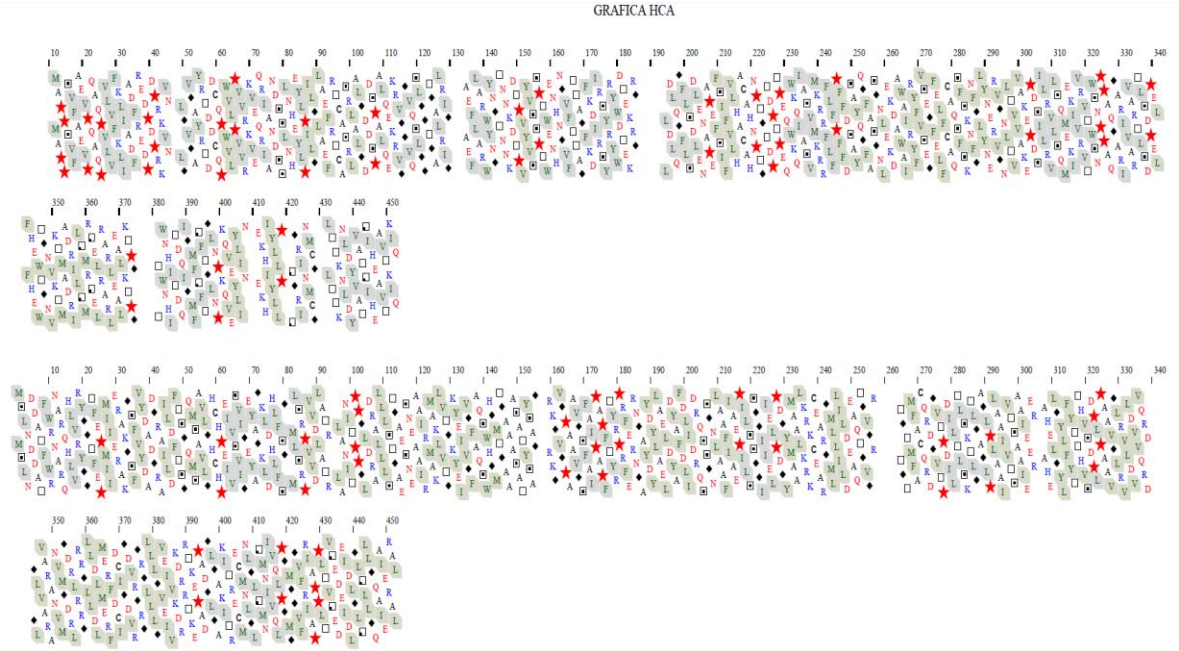

Figura 29. Representación HCA de las secuencias de proteínas pertenecientes al grupo 1ajsa (P00503 y P16932). Imagen obtenida mediante adaptación de software DrawHCA [63]**.**

Para realizar unas pruebas con los resultados obtenidos, se tomó como referencia las secuencias de comparación en la pre-selección obtenida con la red SOM. Se realizan pruebas a las 5 secuencias que se obtuvieron mejores resultados tabla 2., los coeficientes de correlación se observan en la tabla 30.

Secuencia	CoefA_4	Coef D_4	Coef D_3	Coef D_2	Coef D_1
3ia3	1	1	1	1	1
1o1o	1	1	1	1	1
1rvw	1	1	1	1	1
1y0c	1	1	1	1	1
4mqc	1	1	1	1	1

Figura 30. Resultados Correlación para las secuencias de la tabla 2. expuesta para la selección de secuencias utilizando una red SOM.

- Representaciones HCA:

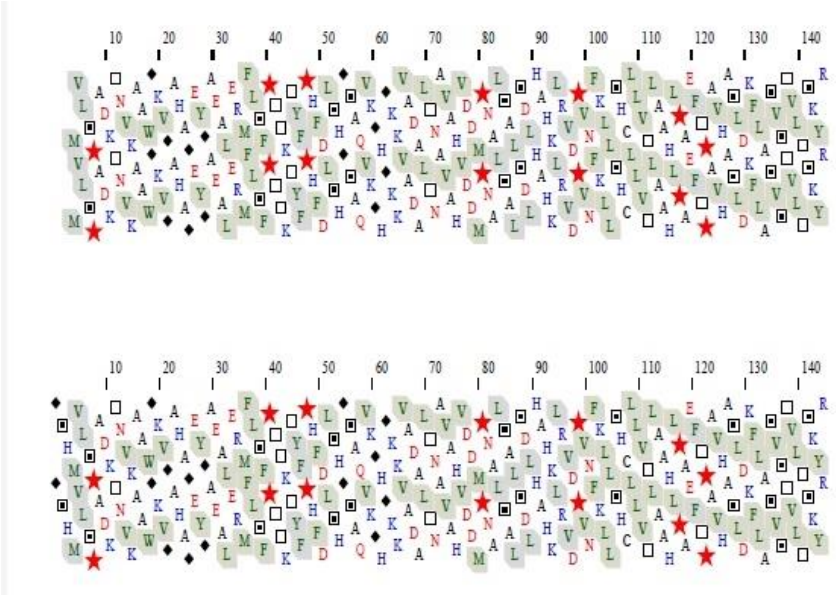

Figura 31. Representación HCA para las secuencias Hemoglobina y 3ia3 perteneciente al grupo 300 de la agrupación del PDB propuesta en la pre-selección de secuencias. Imagen obtenida mediante adaptación de software DrawHCA [56].

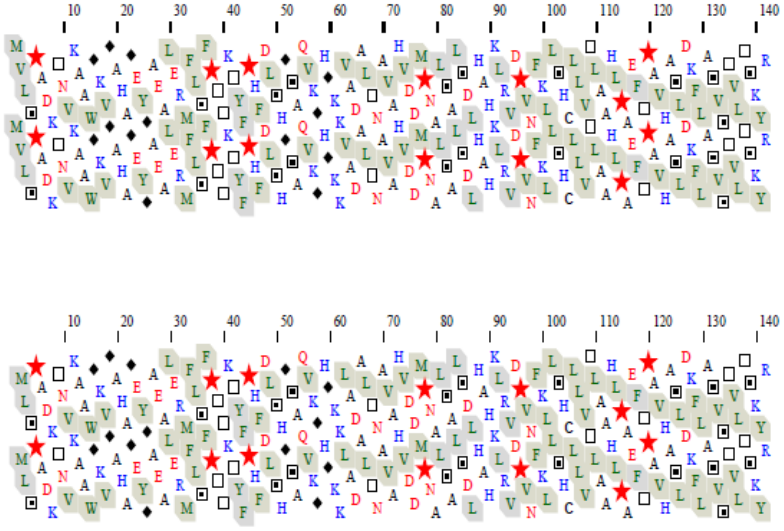

Figura 32. Representación HCA para las secuencias Hemoglobina y 1o1o perteneciente al grupo 300 de la agrupación del PDB propuesta en la pre-selección de secuencias. Imagen obtenida mediante adaptación de software DrawHCA [56].

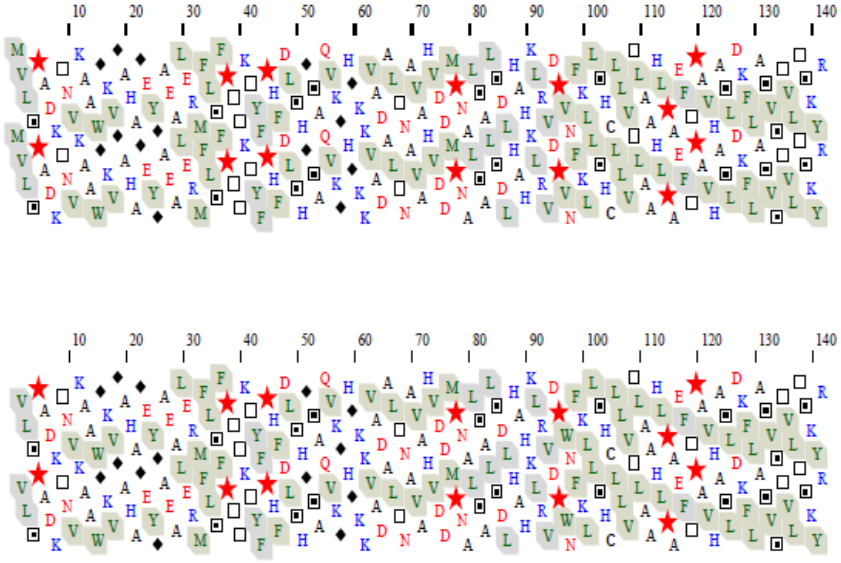

Figura 33. Representación HCA para las secuencias Hemoglobina y 1rvw perteneciente al grupo 300 de la agrupación del PDB propuesta en la pre-selección de secuencias. Imagen obtenida mediante adaptación de software DrawHCA [56]

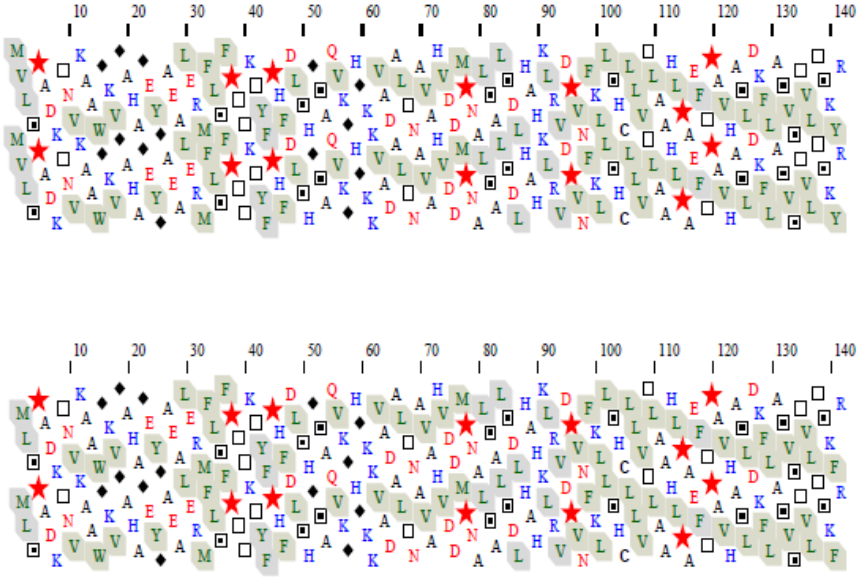

Figura 34. Representación HCA para las secuencias Hemoglobina y 1y0c perteneciente al grupo 300 de la agrupación del PDB propuesta en la pre-selección de secuencias. Imagen obtenida mediante adaptación de software DrawHCA [56]

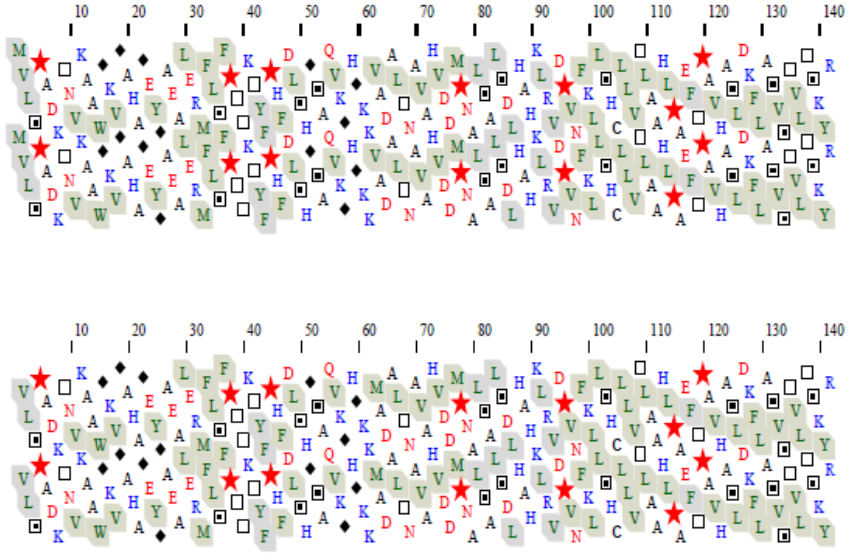

Figura 35. Representación HCA para las secuencias Hemoglobina y 4mqc perteneciente al grupo 300 de la agrupación del PDB propuesta en la pre-selección de secuencias. Imagen obtenida mediante adaptación de software DrawHCA [56].

7. Conclusiones y recomendaciones

7.1. Conclusiones

En este trabajo se desarrolló un método automático de comparación de secuencias de proteínas con una nueva propuesta de codificación que permite encontrar similitud en secuencia de acuerdo a su contenido estructural. Utilizando la DWT y análisis de correlación es posible encontrar similaridad de secuencias que de otra manera no serían comparables utilizando las metodologías convencionales.

El algoritmo desarrollado permite reducir el campo de búsqueda en el estudio de funcionalidad en secuencia o contenido estructural con HCA, ya que se pueden analizar múltiples secuencias y seleccionar aquellas que indiquen un grado de similitud significativa para que el experto pueda realizar un análisis profundo y poder así dar un valor identidad entre aquellas que presenten correlación relevante en al menos alguno de sus coeficientes.

Utilizando la DWT es posible en términos del contenido hidrofóbico de una proteína, cuando es convertida en una señal digital, realizar la descomposición en los diferentes coeficientes donde se obtiene información de las estructuras locales que las componen de acuerdo con sus agrupaciones hidrofóbicas, lo que permite tener de una forma más efectiva un grado de similitud en secuencias de acuerdo al orden en el que se encuentran las agrupaciones en la cadena.

El sistema implementado permite crear una automatización del método HCA con la extracción de la información de las agrupaciones hidrofóbicas, lo que puede generar una aproximación a un método de análisis para las secuencias que se encuentran en la zona de penumbras, lo que sería un aporte para rescatar aquellas secuencias que no han sido posibles de identificar, para este sistema utilizando la *Wavelet* madre Bior3.1 a nivel de descomposición 4, es posible detectar similitud con la codificación propuesta.

Varios factores influyen directamente en todo el proceso de comparación. El alineamiento de secuencias sigue siendo un componente de prioridad para en el proceso de identificación de grupos hidrofóbicos. Al igual es de observar que independiente de la longitud original de una secuencia esta puede cambiar representativamente de acuerdo al contenido de las cadenas hidrofóbicas identificadas con HCA, la codificación permite un análisis centrado solo en aquella información que puede ser asociada a una forma estructuras en secuencia.

Aquellas secuencias que presentan porcentaje de identidad superior al 35%, poseen correlación fuerte, es decir superior al 0.7; para secuencias que se encuentran en con identidad entre el 20% y 40% al menos dos de sus escalas se encuentran con correlación superior a 0.5. Y aquellas en donde se ha encontrado identidad del 20%, no poseen correlación o una sola escala superior a 0.5, en donde

se requiere un análisis enfocado para descartar información no relevante en secuencia.

7.2. Recomendaciones y trabajos futuros

Para la extensión y mejora del trabajo realizado, se plantean algunas recomendaciones en aspectos relacionados con la verificación del modelo y la implantación del mismo, estas son:

- Desde el punto de vista de verificación del modelo se recomienda realizar un análisis de las implicaciones biológicas de los resultados obtenidos de este trabajo, esto debido a que las pruebas realizadas se centran en la valides numérica de los resultados más que en la connotación biológica de los mismos. Además, se recomienda la construcción de una base de datos con secuencias identificadas con el método HCA, para tener un marco de referencia en las pruebas de modelos con un enfoque similar, ya que no se encuentra uno en la literatura.

- Desde el punto de vista de implementación se cree que este modelo permitiría realizar un proceso de anotación de secuencias en bases de datos de forma automática, para ello se trabajó en un modelo de búsqueda de bases de datos basado en información estructural que se presenta en los antecedentes del proyecto, que en conjunto con el modelo desarrollado permitirán realizar anotaciones, donde realizando un pre-procesamiento de posibles secuencias con similitud se reduciría significativamente el campo de búsqueda de secuencias antes de una verificación detalla entre pares de secuencias.

Referencias Bibliográficas

[1] B. Rost, “Twilight zone of protein sequence alignments,” *Protein Engineering*, vol. 12, no. 2, pp. 85–94, Febrero 1999.

[2] C. Gaboriaud, V. Bissery, T. Benchetrit, y J. P. Mornon, “Hydrophobic cluster analysis: an efficient new way to compare and analyse amino acid sequences.,” *FEBS Lett.*, vol. 224, no. 1, pp. 149–55, Noviembre 1987.

[3] I. Callebaut, G. Labesse, P. Durand, a Poupon, L. Canard, J. Chomilier, B. Henrissat, y J. P. Mornon, “Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives.,” *Cellular an Molecular Life Science*, vol. 53, no. 8, pp. 621–45, Agosto 1997.

[4] Faure, Guilhem, and Isabelle Callebaut. "Identification of hidden relationships from the coupling of Hydrophobic Cluster Analysis and Domain Architecture information." Bioinformatics 29.14 1726-1733, 2013.

[5] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, y P. E. Bourne, “The Protein Data Bank,” *Nucleic Acids Research,* vol. 28, no. 1, pp. 235–242, Oct 2000.

[6] K. D. Pruitt, T. Tatusova, y D. R. Maglott, “NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins,” *Nucleic Acids Research*, vol. 35, suppl 1, pp. D61–65, Enero 2007.

[7] UNIPROT CONSORTIUM, et al. "Reorganizing the protein space at the Universal Protein Resource (UniProt)". Nucleic acids research, vol 40, p. 1-5 gkr981, 2011.

[8] A. G. Murzin, S. E. Brenner, T. Hubbard,y C. Chothia, “SCOP: A Structural Classification of Proteins Database for the Investigation of Sequences and Structures”, *Journal of molecular biology*, vol. 247, no 4, pp. 536–540, 1995.

[9] W. R. Pearson and D. J. Lipman, “Improved tools for biological sequence comparison,” *Proceedings of the National Academy of Sciences*, vol. 85, no. 8, pp. 2444–8, Abril 1988.

[10] J. D.Thompson, D. G. Higgins; T. J. GIBSON, "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice". Nucleic acids research, vol. 22, no 22, p. 4673-4680, 1994.

[11] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, “Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.,” Nucleic Acids Res., vol. 25, no. 17, pp. 3389–402, Septiembre 1997.

[12] C. H. de Trad, Q. Fang, and I. Cosic, “Protein sequence comparison based on the wavelet transform approach.,” *Protein Eng.*, vol. 15, no. 3, pp. 193–203, Mar. 2002.

[13] . Böckenhauer D. Bongartz, "Basics of Molecular Biology" en Algorithmic Aspects of Bioinformatics, Ed. Springer-Verlag New York, Inc, pp. 7-8, 2007.

[14] W.K. SUNG," Introduction to molecular biology", en *Algorithms in bioinformatics: A practical introduction*. CRC Press, pp. 1-4, 2009.

[15] M. Ruiz Villareal, y R. Bailey ,[online] Protein Structure: [cited 9 Apr. 2013] About Education, , Available from: <url: http://biology.about.com/od/molecularbiology/ss/protein-structure.htm>
[16] D. J. Selkoe. "Alzheimer's disease: genes, proteins, and therapy," Physiological reviews, vol. 81, no 2, p. 741-766,2001.
[17] W. Pirovano, K. A. Feenstra, J. Heringa ."The meaning of alignment: lessons from structural diversity". BMC bioinformatics, vol. 9, no 1, p. 1-7, 2008.
[18] R. Chenna, et al. "Multiple sequence alignment with the Clustal series of programs". Nucleic acids research, vol. 31, no 13, p. 3497-3500, 2003.
[19] T. A.Tatusova , y T. L. Madden, “BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences,” *FEMS Microbiology Letters*, vol. 174, no. 2, pp. 247–50, Mayo 1999.
[20] Z. Aung y K. L. Tan, “Rapid retrieval of protein structures from databases,” *Drug Discovery Today*, vol. 12, no. 17–18, pp. 732–739, Septiembre 2007.
[21] R. D. Finn, J. Clements, y S. R. Eddy, “HMMER web server: interactive sequence similarity searching,” *Nucleic Acids Research.*, vol. 39, no. Web Server issue, pp. W29–W37, Julio 2011.
[22] R. D. Finn, A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E. L. L. Sonnhammer, J. Tate, y M. Punta, “Pfam: the protein families database,” *Nucleic Acids Research*, vol. 42, no. Database issue gktl1223, pp. D222–D230, Enero 2014.
[23] Z. Wen, K. Wang, M. Li, F. Nie, y Y. Yang, “Analyzing functional similarity of protein sequences with discrete wavelet transform,” vol. 29, no.3 ,pp. 220–228, Abril 2005.
[24] R. C. Edgar, “MUSCLE: multiple sequence alignment with high accuracy and high throughput.,” *Nucleic Acids Research*, vol. 32, no. 5, pp. 1792–7, Enero 2004.
[25] I. Callebaut, et al. "Hydrophobic cluster analysis and modeling of the human Rh protein three-dimensional structures". Transfusion clinique et biologique, vol. 13, no 1, p. 70-84, 2006.
[26] R. Eudes, K. L. Tuan, J. Delettré, J. P. Mornon, y I. Callebaut, “A generalized analysis of hydrophobic and loop clusters within globular protein sequences,” *BMC Structural Biology*, vol. 7,no.1, pp.1- 22, Enero 2007.
[27] P. J. Silva, “Assessing the reliability of sequence similarities detected through hydrophobic cluster analysis,” Proteins: Structure, Function, and Bioinformatics, vol 70, no 4. pp. 1588–1594, Abril 2007.
[28] K. Katoh, K. Misawa, K. Kuma, y T. Miyata, “MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform.,” *Nucleic Acids Research*, vol. 30, no. 14, pp. 3059–66, Julio 2002.
[29] K. B. Murray, D. Gorse, y J. M. Thornton, “Wavelet transforms for the characterization and detection of repeating motifs.,” *J. Mol. Biol.*, vol. 316, no. 2, pp. 341–63, Febrero 2002.
[30] J. Qiu, R. Liang, X. Zou, y J. Mo, “Prediction of protein secondary structure based on continuous wavelet transform.,” *Talanta*, vol. 61, no. 3, pp. 285–93, Noviembre 2003.

[31] E. Pirogova, et al. "Development of new computational amino acid parameters for protein structure/function analysis within the resonant recognition model," Engineering in Medicine and Biology Society. Proceedings of the 23rd Annual International Conference of the IEEE. IEEE, 2001. p. 2890-2893, 2001.
[32] P. Morozov, T. Sitnikova, y G. Churchill, “A New Method for Characterizing Replacement Rate Variation in Molecular Sequences : Application of the Fourier and Wavelet Models to Drosophila and Mammalian Proteins,” Genetics, vol. 154, no 1, pp. 381-395 2000.
[33] I. Cosic, J. Fang- "Evaluation of different wavelet constructions (designs) for analysis of protein sequences.".International Conference on Digital Signal Processing Proceedings. IEEE, 2002.
[34] National Center for Biotechnology Information. VAST: Vector Alignment Search Tool-Non-redundant PDB chain set [cited 5 May. 2015]. [Online]. Available :<url: http://structure.ncbi.nlm.nih.gov/Structure/VAST/nrpdb.html>.
[35] J. Ruan , K. Wang, J. Yang, L. A. Kurgan, K. J. Cios. “Highly accurate and consistent method for prediction af helix and strand content from primary protein sequences”, Artificial Intelligence, vol 35, no.1, pp. 19-35, Febrero 2005.
[36] S. a. Mingoti and J. O. Lima, “Comparing SOM neural network with Fuzzy c-means, K-means and traditional hierarchical clustering algorithms,” European Journal of Operational Research ,vol. 174, no.3, pp. 1742–1759, 2006.
[37] P. Mangiameli, S. K. Chen, and D. West, “A comparison of SOM neural network and hierarchical clustering methods,” European Journal of Operational Research, vol. 93, no 2, p. 402-417, 1996.
[38] B. Segerman , et al. "Bioinformatic tools for using whole genome sequencing as a rapid high resolution diagnostic typing tool when tracing bioterror organisms in the food and feed chain,".International journal of food microbiology, vol. 145, p. S167-S176, 2011.
[39] J. Söding, A. Biegert, A.N. Lupas. "The HHpred interactive server for protein homology detection and structure prediction". Nucleic acids research, vol. 33, no suppl 2, p. W244-W248, 2005.
[40] C. Gaboriaud , et al. "Hydrophobie cluster analysis reveals duplication in the external structure of human α-interferon receptor and homology with γ-interferon receptor external domain," FEBS letters, vol. 269, no 1, p. 1-3, 1990.
[41] I. Callebaut, et al. "Hydrophobic Cluster Analysis Reveals a Third Chromodomain in theTetrahymenaPdd1p Protein of the Chromo Superfamily," Biochemical and biophysical research communications, vol. 235, no 1, p. 103-107, 1997.
[42] J. D. Thompson, T. J. Gibson, y D. G. Higgins, " Multiple sequence alignment using ClustalW and ClustalX,” en Current protocols in bioinformatics, Wiley Online Library, pp. 2.3.1–2.3.22, 2002.
[43] F. Naznin, R. Sarker,D. Essam. "Progressive alignment method using genetic algorithm for multiple sequence alignment," Evolutionary Computation, IEEE Transactions on, vol. 16, no 5, p. 615-631, 2012.
[44] D. J. Lipman and W. R. Pearson, “Rapid and Sensitive Protein Similarity

Searches," *Science*, vol. 227, no 4693, p. 1435-1441, Mazo 1985.
[45] G. Mapping, D. Bergsma, F. A. Mcmorris, R. Creagan, F. Rucciuti, F. H. Ruddle, J. A. Tischfield, R. P. Creagan, F. Ricciuti, V. G. Dev, P. A. Miller, P. W. Allderdice, J. Miller, C. B. Laurell, R. Weitkamp, D. L. Rucknagel, M. L. Petras, T. R. Chen, E. A. Nichols, J. Tischfield, J. R. Debro, S. N. Foundation, and A. Bernhard, "Amino Acid Difference Formula to Help Explain Protein Evolution," *Science*, vol. 185, no 4154, Marzo 1974.
[46] K. M. Bloch, G. R. Arce, "Chapter 9 Analyzing Protein Sequences Using Signal Analysis Techniques" En Computational and Statistical Approaches to Genomics. Springer US,. p. 137-161, 2006.
[47] PHAM, Tuan D. Spectral distortion measures for biological sequence comparisons and database searching. Pattern Recognition, 2007, vol. 40, no 2, p. 516-529.
[48] j. Su, y J. Bao, "A Wavelet Transform Based Protein Sequence Similarity Model," *Applied Mathematics & Information Sciences,* vol. 1110, no. 3, pp. 1103–1110, Febrero 2013.
[49] K. D. Rao,y S. Swamy, "Analysis of Genomics and Proteomics Using DSP Techniques," *IEEE Transactions On Circuits And Systems,* vol. 55, no. 1, pp. 370–378, Febrero 2008.
[50] M. Yan, Z. S. Lin, and C. T. Zhang, "A new fourier transform approach for protein coding measure based on the format of the Z curve.," *Bioinformatics*, vol. 14, no. 8, pp. 685–90, Enero 1998.
[51] R.-P. Liang, S.-Y. Huang, S.-P. Shi, X.-Y. Sun, S.-B. Suo, and J.-D. Qiu, "A novel algorithm combining support vector machine with the discrete wavelet transform for the prediction of protein subcellular localization.," *Computers in Biology and Medicine*, vol. 42, no. 2, pp. 180–7, Febrero 2012.
[52] M. Chen, B. Liu, W. Yan, y B. Shen, "Wavelet Transform Based Protein Decoy Discrimination," *En Bioinformatics and Biomedical Engineering, ICBBE 2009. 3rd International Conference on. IEEE. pp. 1-4*, Junio 2009.
[53] P. Lio, "Wavelets in bioinformatics and computational biology: state of art and perspectives", *Bioinformatics*, vol. 19, no. 1, pp. 2–9, Julio 2003.
[54] H. Hirakawa, y S. Kuhara, "Prediction of Hydrophobic Wavelet Cores Analysis of Proteins Using Wavelet Analysis," *Genome Informatics*, vol.8, pp. 61–70, 1996.
[55] J.-D. Qiu, X.-Y. Sun, J.-H. Huang, and R.-P. Liang, "Prediction of the types of membrane proteins based on discrete wavelet transform and support vector machines," *Protein J.*, vol. 29, no. 2, pp. 114–9, Febrero 2010.
[56] A. Krishnan, K.-B. Li, y P. Issac, "Rapid detection of conserved regions in protein sequences using wavelets.," *In Silico Biology*, vol. 4, no. 2, pp. 133–48, Enero 2004.
[57] J.-D. Qiu, J.-H. Huang, R.-P. Liang, and X.-Q. Lu, "Prediction of G-protein-coupled receptor classes based on the concept of Chou's pseudo amino acid composition: an approach from discrete wavelet transform," *Analytical Biochemistry*, vol. 390, no. 1, pp. 68–73, Julio 2009.
[58] S. . Shi, J. D. Qiu, X.-Y. Sun, J. H. Huang, S.Y. Huang, S.-B. Suo, R.-P. Liang,

and L. Zhang, “Identify submitochondria and subchloroplast locations with pseudo amino acid composition: approach from the strategy of discrete wavelet transform feature extraction.,” *Biochimica et Biophysica Acta*, vol. 1813, no. 3, pp. 424–30, Marzo 2011.

[59] & L. Oyster, C. K., Hanten, W. P., *Introduction to research: A guide for the health science professional. Lippincott.* 1987.

[60] A. Bahr, J. D. Thompson, J. Thierry, y O. Poch, “BAliBASE (Benchmark Alignment dataBASE): enhancements for repeats, transmembrane sequences and circular permutations,” *Nucleic Acids Research*, vol. 29, no. 1, pp. 323–326, Octubre 2001.

[61] L. Pasti, B. Walczak, D. L. Massart, P. Reschiglian, “Optimization of signal denoising in discrete wavelet transform,” *Chemometrics and Intelligent Laboratory System*, vol. 48, no. 1, pp. 21–34, Junio 1990.

[62] S. F. ALTSCHUL , et al. "Basic local alignment search tool. Journal of molecular biology", vol. 215, no 3, p. 403-410, 1990.

[63] Grupo de invetigación Giftex, Universidad Industrial de Santander. Códigos editados para la invetigación con base a DRAWHCA. Director Cristian Blanco Tirado . 2016.

Bibliografía

ALTSCHUL, Stephen F., et al. Basic local alignment search tool. Journal of molecular biology, 1990, vol. 215, no 3, p. 403-410.

ALTSCHUL, Stephen F., et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.En: Nucleic acids research, 1997, vol. 25, no 17, p. 3389-3402.

AUNG, Zeyar; TAN, Kian-Lee. Rapid retrieval of protein structures from databases. En: Drug discovery today, 2007. vol. 12, no 17, p. 732-739.

BAHR, Anne, et al. BAliBASE (Benchmark Alignment dataBASE): enhancements for repeats, transmembrane sequences and circular permutations. Nucleic Acids Research, 2001, vol. 29, no 1, p. 323-326.

BLOCH, Karen M.; ARCE, Gonzalo R. Analyzing protein sequences using signal analysis techniques. En Computational and Statistical Approaches to Genomics. Springer US, 2006. p. 137-161.

BÖCKENHAUER, Hans-J. y BONGARTZ, Dirk. Basics of molecular Biology. En: Algorithmic Aspects of Bioinformatics. 1 ed. Springer-Verlag Berlin Heidelberg. 2007. p 7-10.

CALLEBAUT, I., et al. Hydrophobic cluster analysis and modeling of the human Rh protein three-dimensional structures. En: Transfusion clinique et biologique, 2006, vol. 13, no 1, p. 70-84.

CALLEBAUT, Isabelle, et al. Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives. Cellular and Molecular Life Sciences CMLS, 1997, vol. 53, no 8, p. 621-645.

CALLEBAUT, Isabelle, et al. Hydrophobic Cluster Analysis Reveals a Third Chromodomain in theTetrahymenaPdd1p Protein of the Chromo Superfamily. Biochemical and biophysical research communications, 1997, vol. 235, no 1, p. 103-107.

CHEN, Minxin, et al. Wavelet transform based protein decoy discrimination. En Bioinformatics and Biomedical Engineering, 2009. ICBBE 2009. 3rd International Conference on. IEEE, 2009. p. 1-4.

CHENNA, Ramu, et al. Multiple sequence alignment with the Clustal series of programs. Nucleic acids research, 2003, vol. 31, no 13, p. 3497-3500.

COSIC, I.; FANG, J. Evaluation of different wavelet constructions (designs) for analysis of protein sequences. En International Conference on Digital Signal Processing Proceedings. IEEE, 2002.

DE TRAD, Chafia Hejase; FANG, Qiang; COSIC, Irena. Protein sequence comparison based on the wavelet transform approach. Protein engineering, 2002, vol. 15, no 3, p. 193-203.

EDGAR, Robert C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. En: Nucleic acids research, 2004, vol. 32, no 5, p. 1792-1797.

EUDES, Richard, et al. A generalized analysis of hydrophobic and loop clusters within globular protein sequences. En: BMC structural biology, 2007, vol. 7, no 1, p. 1-22.

FAURE, Guilhem; CALLEBAUT, Isabelle. Identification of hidden relationships from the coupling of Hydrophobic Cluster Analysis and Domain Architecture information. En: Bioinformatics, 2013, vol. 29, no 14, p. 1726-1733.

FINN, Robert D., et al. Pfam: the protein families database. En: Nucleic acids research, 2013, vol 42, p. D222-D230, gkt1223.

FINN, Robert D.; CLEMENTS, Jody; EDDY, Sean R. HMMER web server: interactive sequence similarity searching. Nucleic acids research, 2011. vol 39 p. W29-W37, gkr367.

GABORIAUD, Christine, et al. Hydrophobic cluster analysis: an efficient new way to compare and analyse amino acid sequences. En: FEBS letters, 1987, vol. 224, no 1, p. 149-155.

GABORIAUD, Christine, et al. Hydrophobie cluster analysis reveals duplication in the external structure of human α - interferon receptor and homology with γ - interferon receptor external domain. FEBS letters, 1990, vol. 269, no 1, p. 1-3.

GRANTHAM, R. Amino acid difference formula to help explain protein evolution. Science, 1974, vol. 185, no 4154, p. 862-864.

Grupo de invetigación Giftex, Universidad Industrial de Santander. Códigos editados para la invetigación con base a DRAWHCA. Director Cristian Blanco Tirado . 2016.

HIRAKAWA, Hideki; KUHARA, Satoru. Prediction of hydrophobic cores of proteins using wavelet analysis. En: Genome Informatics, 1997, vol. 8, p. 61-70.

KATOH, Kazutaka, et al. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic acids research, 2002, vol. 30, no 14, p. 3059-3066.

KRISHNAN, Arun; LI, Kuo-Bin; ISSAC, Praveen. Rapid detection of conserved regions in protein sequences using wavelets. In silico biology, 2004, vol. 4, no 2, p. 133-148.

LIANG, Ru-Ping, et al. A novel algorithm combining support vector machine with the discrete wavelet transform for the prediction of protein subcellular localization. Computers in biology and medicine, 2012, vol. 42, no 2, p. 180-187.

LIO, Pietro. Wavelets in bioinformatics and computational biology: state of art and perspectives. Bioinformatics, 2003, vol. 19, no 1, p. 2-9.

LIPMAN, David J.; PEARSON, William R. Rapid and sensitive protein similarity searches. Science, 1985, vol. 227, no 4693, p. 1435-1441.

M. Ruiz Villareal, y R. Bailey ,[online] Protein Structure: [cited 9 Apr. 2013] About Education, , Available from: <url: http://biology.about.com/od/molecularbiology/ss/protein-structure.htm>

MANGIAMELI, Paul; CHEN, Shaw K.; WEST, David. A comparison of SOM neural network and hierarchical clustering methods. En: European Journal of Operational Research, 1996, vol. 93, no 2, p. 402-417.

MINGOTI, Sueli A.; LIMA, Joab O. Comparing SOM neural network with Fuzzy c-means, K-means and traditional hierarchical clustering algorithms. En: European Journal of Operational Research, 2006, vol. 174, no 3, p. 1742-1759.

MOROZOV, Pavel, et al. A new method for characterizing replacement rate variation in molecular sequences: application of the Fourier and wavelet models to Drosophila and mammalian proteins. Genetics, 2000, vol. 154, no 1, p. 381-395.

MURRAY, Kevin B.; GORSE, Denise; THORNTON, Janet M. Wavelet transforms for the characterization and detection of repeating motifs. Journal of molecular biology, 2002, vol. 316, no 2, p. 341-363.

MURZIN, Alexey G., et al. SCOP: a structural classification of proteins database for the investigation of sequences and structures. En: Journal of molecular biology, 1995, vol. 247, no 4, p. 536-540.

National Center for Biotechnology Information. VAST: Vector Alignment Search Tool-Non-redundant PDB chain set [cited 5 May. 2015]. [Online]. Available :<url: http://structure.ncbi.nlm.nih.gov/Structure/VAST/nrpdb.html>.

NAZNIN, Farhana; SARKER, Ruhul; ESSAM, Daryl. Progressive alignment method using genetic algorithm for multiple sequence alignment. Evolutionary Computation, IEEE Transactions on, 2012, vol. 16, no 5, p. 615-631.

OYSTER, Carol K.; HANTEN, William P.; LLORENS, Lela A. Introduction to research: A guide for the health science professional. Lippincott Williams & Wilkins, 1987.

PASTI, L., et al. Optimization of signal denoising in discrete wavelet transform. Chemometrics and intelligent laboratory systems, 1999, vol. 48, no 1, p. 21-34.

PEARSON, William R.; LIPMAN, David J. Improved tools for biological sequence comparison. En: Proceedings of the National Academy of Sciences, 1988, vol. 85, no 8, p. 2444-2448.

PHAM, Tuan D. Spectral distortion measures for biological sequence comparisons and database searching. Pattern Recognition, 2007, vol. 40, no 2, p. 516-529.

PIROGOVA, Elena, et al. Development of new computational amino acid parameters for protein structure/function analysis within the resonant recognition model. En Engineering in Medicine and Biology Society, 2001. Proceedings of the 23rd Annual International Conference of the IEEE. IEEE, 2001. p. 2890-2893.

PIROVANO, Walter; FEENSTRA, K. Anton; HERINGA, Jaap. The meaning of alignment: lessons from structural diversity. BMC bioinformatics, 2008, vol. 9, no 1, p. 1-7.

PRUITT, Kim D.; TATUSOVA, Tatiana; MAGLOTT, Donna R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. En: Nucleic acids research, 2007, vol. 35, p. D61-D65.

QIU, Jian-Ding, et al. Prediction of G-protein-coupled receptor classes based on the concept of Chou's pseudo amino acid composition: an approach from discrete wavelet transform. Analytical Biochemistry, 2009, vol. 390, no 1, p. 68-73.

QIU, Jianding, et al. Prediction of protein secondary structure based on continuous wavelet transform. Talanta, 2003, vol. 61, no 3, p. 285-293.

QIU, Jian-Ding, et al. Prediction of the types of membrane proteins based on discrete wavelet transform and support vector machines. The protein journal, 2010, vol. 29, no 2, p. 114-119.

RAO, K. Deergha; SWAMY, M. N. S. Analysis of genomics and proteomics using DSP techniques. Circuits and Systems I: Regular Papers, IEEE Transactions on, 2008, vol. 55, no 1, p. 370-378.

ROST, Burkhard. Twilight zone of protein sequence alignments. En: Protein engineering, 1999, vol. 12, no 2, p. 85-94.

RUAN, Jishou, et al. Highly accurate and consistent method for prediction of helix and strand content from primary protein sequences. Artificial Intelligence in Medicine, 2005, vol. 35, no 1, p. 19-35.

SEGERMAN, Bo, et al. Bioinformatic tools for using whole genome sequencing as a rapid high resolution diagnostic typing tool when tracing bioterror organisms in the food and feed chain. En: International journal of food microbiology, 2011, vol. 145, p. S167-S176.

SELKOE, Dennis J. Alzheimer's disease: genes, proteins, and therapy. En: Physiological reviews, 2001, vol. 81, no 2, p. 741-766.

SHI, Shao-Ping, et al. Identify submitochondria and subchloroplast locations with pseudo amino acid composition: approach from the strategy of discrete wavelet transform feature extraction. Biochimica et Biophysica Acta (BBA)-Molecular Cell Research, 2011, vol. 1813, no 3, p. 424-430.

SILVA, Pedro J. Assessing the reliability of sequence similarities detected through hydrophobic cluster analysis. Proteins: Structure, Function, and Bioinformatics, 2008, vol. 70, no 4, p. 1588-1594.

SÖDING, Johannes; BIEGERT, Andreas; LUPAS, Andrei N. The HHpred interactive server for protein homology detection and structure prediction. Nucleic acids research, 2005, vol. 33, no suppl 2, p. W244-W248.

SU, Jie; BAO, Junpeng. A wavelet transform based protein sequence similarity model. Applied Mathematics & Information Sciences, 2013, vol. 7, no 3, p. 1103-1010.

SUNG, Wing-Kin. Introduction to Molecular Biology. En: Algorithms in bioinformatics: A practical introduction. CRC Press.2009. p 1-9.

TATUSOVA, Tatiana A.; MADDEN, Thomas L. BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. FEMS microbiology letters, 1999, vol. 174, no 2, p. 247-250.

THOMPSON, Julie D., et al. Multiple sequence alignment using ClustalW and ClustalX. En :Current protocols in bioinformatics, 2002, p. 2.3. 1-2.3. 22.

THOMPSON, Julie D.; HIGGINS, Desmond G.; GIBSON, Toby J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.En: Nucleic acids research, 1994, vol. 22, no 22, p. 4673-4680.

UNIPROT CONSORTIUM, et al. Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic acids research, 2011, vol 40, p. 1-5 gkr981.

WEN, Zhi-ning, et al. Analyzing functional similarity of protein sequences with discrete wavelet transform. Computational biology and chemistry, 2005, vol. 29, no 3, p. 220-228.

WESTBROOK, John, et al. The protein data bank and structural genomics. En: Nucleic acids research, 2003, vol. 31, no 1, p. 489-491.

YAN, Ming; LIN, Zhe Suai; ZHANG, Chun Ting. A new fourier transform approach for protein coding measure based on the format of the Z curve. Bioinformatics, 1998, vol. 14, no 8, p. 685-690.

Anexos

A. Cálculo de los valores EIIP normalizados

Ecuación de normalización:

$$X' = \frac{X - \mu}{\sigma}$$

Media EIIP: 0.049

Desviación estándar: 0.0402341

Aminoácido	EIIP	Valor EIIP Normalizado
Leu	0.0000	-1,217872664
Ile	0.0000	-1,217872664
Asn	0.0036	-1,128396305
Gly	0.0050	-1,093599943
Val	0.0057	-1,076201762
Glu	0.0058	-1,073716308
Pro	0.0198	-0,72575269
His	0.0242	-0,616392695
Lys	0.0371	-0,295769076
Ala	0.0373	-0,290798167
Tyr	0.0516	0,064621815
Trp	0.0548	0,144156356
Gln	0.0761	0,673558147
Met	0.0823	0,827656321
Ser	0.0829	0,842569047
Cys	0.0829	0,842569047
Thr	0.0941	1,120939942
Phe	0.0946	1,133367214
Arg	0.0959	1,165678121
Asp	0.1263	1,921256264