DISEÑO E IMPLEMENTACIÓN DE UN PROTOTIPO PARA LA COMPARACIÓN DE SECUENCIAS 2D DE PROTEINAS

Jenny Sofía Gómez Guerrero

Humberto Ruiz Roa

ESCUELA DE INGENIERÍA DE SISTEMAS E INFORMÁTICA

UNIVERSIDAD INDUSTRIAL DE SANTANDER

Bucaramanga – Julio de 2008

DISEÑO E IMPLEMENTACIÓN DE UN PROTOTIPO PARA LA COMPARACIÓN DE SECUENCIAS 2D DE PROTEINAS

Jenny Sofía Gómez Guerrero

Humberto Ruiz Roa

Trabajo de Grado para optar al título de

Ingeniero de Sistemas

Director

MPe. Henry Arguello Fuentes

Codirector

Ph.D. Cristian Blanco Tirado

ESCUELA DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
UNIVERSIDAD INDUSTRIAL DE SANTANDER
Bucaramanga – Julio de 2008

Al Espíritu Santo, por ser mi luz

A mi madre Beatriz por ser mi mayor compañía y mi gran fortaleza

A mi padre Luis por su esfuerzo y constante apoyo

A Willy por darme ánimos y por haber llegado a mi vida

Jenny

A Dios por la existencia,
A mis padres, por todo lo que me han dado en
la vida, especialmente por sus sabios consejos
y por estar a mi lado en los momentos difíciles.
A mi hermano por estar en las buenas y en las malas.
A mis abuelitos Antonio y Leonilde por que me ayudaron a dar este
importante paso en mi vida y estoy seguro que están orgullosos de mi.
A mi familia, especialmente a mi tía Blanca, Gloria, Carmen por estar
siempre dispuestos a ayudarme.

Humberto

AGRADECIMIENTOS

Este trabajo de grado fue posible gracias a la colaboración, apoyo y conocimientos brindados por los profesores Cristian Blanco, Henry Arguello, Jorge Hernández.

Agradecemos particularmente:

A Dios, por todo

A nuestros padres, por ser motivación y colaboración para salir adelante a lo largo de nuestras vidas.

A Alfonso, por el apoyo con sus conocimientos en biología.

Al grupo GIFTEX, por el compañerismo.

A nuestros amigos sinceros, por los momentos compartidos y por creer en nosotros.

RESUMEN

TITULO: DISEÑO E IMPLEMENTACIÓN DE UN PROTOTIPO PARA LA COMPARACIÓN DE SECUENCIAS 2D DE PROTEINAS[1]

AUTORES: Jenny Sofía Gómez Guerrero y Humberto Ruiz Roa[2]

PALABRAS CLAVES: Análisis de agregados hidrofóbicos, HCA, comparación, elucidación, cluster, secuencia, proteína, identidad.

DESCRIPCIÓN:

El diseño y desarrollo de un prototipo que permita y apoye la comparación y elucidación de secuencias 2D de proteínas aplicando el método de Análisis de Clusters Hidrofóbicos (HCA) ha sido el eje central y el resultado del presente proyecto, el cual busca ser un aporte a la tecnología de identificación de secuencias.

El análisis de agregados hidrofóbicos, es un método de comparación que facilita la caracterización de secuencias de proteínas localizadas en entornos de baja identidad, el cual, ha hecho posible predecir conformaciones estructurales, y características funcionales difícilmente detectables por los métodos de comparación 1D.

Ésta metodología, depende de la habilidad ojo-cerebro humanos para reconocer, descifrar y asociar correctamente las similitudes entre imágenes complejas con desigual información biológica; por tanto, constituye un proceso dispendioso que frecuentemente exige realizar múltiples alineamientos hasta encontrar el de mejor grado de identidad.

La realización del presente proyecto, ofrece como resultado una alternativa computacional para superar dichos inconvenientes en la comparación por HCA, ya que apropia los criterios inherentes al método para disminuir el grado de subjetividad y de dificultad en la comparación habitual.

En las pruebas realizadas, se obtuvieron resultados similares a los obtenidos en comparaciones echas por expertos usando el método visual, adicionalmente, se instaló el sistema de comparación sobre una aplicación soportada en ambiente Web, la cual se probó en línea sobre un servidor bajo plataforma Linux en el grupo GIFTEX adscrito a la escuela de Química de la Universidad Industrial de Santander.

[1] Proyecto de Grado

[2] Facultad de Ingenierías Físicomecánicas. Escuela de Ingeniería de Sistemas e Informática. Director: Henry Arguello Fuentes. Codirector: Cristian Blanco Tirado.

ABSTRACT

TITLE: DESIGN AND IMPLEMENTATION OF A PROTOTYPE FOR THE COMPARISON OF TWO-DIMENSIONAL (2D) PROTEINS SEQUENCES[3]

AUTHORS: Jenny Sofía Gómez Guerrero and Humberto Ruiz Roa[4]

KEY WORDS: Hydrophobic Cluster Analysis, HCA, comparison, elucidation, cluster, sequences, protein, identity.

DESCRIPTION:

The design and development of a prototype that allows and supports the comparison and elucidation of 2D protein sequences applying the method of Hydrofobic Cluster Analysis (HCA) has been the backbone and the result of this project, which seeks to be a contribution to the technology of sequences identification.

The Hydrofobic Cluster Analysis is a method of comparison that facilitates the characterization of proteins sequences located in low identity environments, which, has made possible to predict difficultly detectable structural conformations and functional features by the 'linear' (1D) methods of comparison.

This methodology depends on the human eye – brain ability to recognize, decipher and associate correctly the similarities between complex images with unequal biological information; therefore, constitutes a time expensive process that frequently requires to make multiple alignments up to finding that of better degree of identity.

The realization of the present project, offers as result a computacional alternative to overcome the above mentioned disadvantages in the comparison through HCA, since it adapts the criteria inherent to the method to decrease the degree of subjectivity and difficulty in the habitual comparison.

In the realized tests, results similar to the obtained ones were obtained in comparisons made by experts using the visual method, additionally, the system of comparison was installed on an application supported Web environment, which was proved on line over a Linux server platform at the GIFTEX group joined to the Chemistry School of the Santander Industrial University

[3] Degree Work

[4] Faculty of Physical-Mechanical Engineering. Computer Science and Systems Engineering School. Director: Henry Arguello Fuentes Codirector: Cristian Blanco Tirado.

TABLA DE CONTENIDO

Capitulo 1 18
Marco Teórico 18
1.1 Las Proteínas 18
1.1.1 Aminoácidos 19
1.1.1.1 Aminoácidos Hidrofóbicos 20
1.1.1.2 Aminoácidos Hidrofílicos 20
1.1.2 Aspectos Generales de la estructura de las proteinas 20
1.1.2.1 La α- Hélice 22
1.1.2.2 La β –Plegada o laminar 23
1.1.2 Clasificación de las Proteínas 24
1.2 Comparación de Secuencias 26
1.2.1 Comparación básica por identidades 27
1.2.2 Comparación por semejanza 28
1.2.3 Alineamiento de secuencias biológicas 29
1.4 Hydrophobic Cluster Análysis – HCA 31
1.4.1 ¿Qué es HCA? 31
1.4.2 Zona Twiligth 31
1.4.3 Descripción del Método HCA 31
1.4.3.1 Código P 32
1.4.3.2 Resumen de la nomenclatura de clusters 33
1.4.3.3 Representación bidimensional 33

Capítulo 2 38
Construcción del Modelo Computacional 38
2.1 Metodología 38
2.1.1 Diagrama de casos de uso 39
2.1.2 Descripción de casos de uso 40
2.2 Generación de gráficas bidimensionales en PostScript 46
2.2.1 Gráfica HCA 47
2.3 Desarrollo de los algoritmos de comparación 49
2.3.1 Comparación rápida con HCA 50

2.3.2 Propuesta: Algoritmo de Comparación basado en el Análisis Hidrofóbico de Clusters 55
2.3.2.1 Metodología 55
2.3.2.2 Descripción del algoritmo 58
2.3.2.3 Criterios de comparación 58
2.3.2.4 Análisis de resultados 63

Capítulo 3 70
Conclusiones y Recomendaciones 70
3.1 Conclusiones 70
3.2 Recomendaciones 71

Bibliografía 72
Anexos 77

LISTA DE TABLAS

Tabla 1 Resumen de resultados método Comparación rápida con HCA 54
Tabla 2 Resumen de resultados método Algoritmo de Comparación Basado en el Análisis Hidrofóbico de Clusters 67

LISTA DE FIGURAS

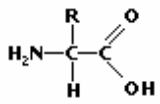
Figura 1. Estructura de un aminoácido con el grupo amino y carboxilo [Joachim, 2007]...19
Figura 2. Estructura de las proteinas…… 21
Figura 3. Estructuras secundarias: Hélice alfa. Curtis 2000…….22
Figura 4. Estructuras secundarias de las proteínas: Hoja plegada β. Curtis 2000……23
Figura 5. Estructuras Alfa y Beta, dentro de la estructura terciaria de una proteína……24
Figura 6. Fosfolipasa D - Aminoácidos entre las estructuras secundarias……..24
Figura 7. Fosfolipasa D Estructuras alfa y beta……24
Figura 8. Comparación básica por identidades…….27
Figura 9. Alineamiento óptimo por identidades. …….28
Figura 10. Intervalo Twiligth Zone……...31
Figura 11. Secuencia lineal y código binario……32
Figura 12. Análisis de Clústers Hidrofóbicos……...35
Figura 13 Representación de clusters en HCA…….36
Figura 14 Ejemplo de comparación de 2 secuencias usando el análisis de clusters hidrofóbicos. …… 37
Figura 15 Diagrama de casos de uso…… 39
Figura 16 Modelo de Gráfica HCA……...47
Figura 17 Estrategia de comparación – Algoritmo……...49
Figura 18 Proceso de generación del código único para cada proteína……51
Figura 19 Alineamiento matricial de secuencias……...56
Figura 20 Patrón de distancia constante entre aminoácidos Hidrofóbicos coincidentes…...61
Figura 21 Presencia de saltos en el alineamiento en los hallazgos de aminoácidos coincidentes e idénticos…….62
Figura 22 Zona de similitud……..63
Figura 23 Rompimiento por Prolina…… 63
Figura 24 Comparación ORFX vs DUF614……..68
Figura 25 Comparación 4Q21 Vs 1WKQ…….69

LISTA DE ANEXOS

Anexo A Listado Aminoácidos Esenciales……...77

GLOSARIO

AMINOÁCIDO: Importante clase de compuestos orgánicos que contienen un grupo amino ($-NH_2$) y un grupo carboxilo (-COOH). Veinte de estos compuestos son los constituyentes de las proteínas (Ver anexo). Todos ellos responden a la siguiente fórmula general:



COMPARACIÓN: El objetivo de comparar dos secuencias es encontrar la posición relativa de ambas en las que se produzca mayor número de coincidencias entre sus componentes. El valor "número de coincidencias o identidades" representaría la valoración de su parecido y podría darse en valores absolutos o porcentuales (normalizado) dividiéndolo entre el máximo valor posible (la longitud de la secuencia más corta).

CLUSTER: Grupo de elementos pertenecientes a una población o conjunto que los contiene, los cuales poseen las mismas propiedades.

DOMINIOS: Porción de una proteína con estructura terciaria definida (40-350 aminoácidos). En general asociados a una función particular.

ENLACE PEPTÍDICO: Al reaccionar un ácido carboxílico (COOH) con una amina (NH2), el enlace resultante se denomina una amida. El enlace amida formado entre dos alfa-aminoácidos recibe el nombre de enlace peptídico. A los aminoácidos unidos por enlaces peptídico ya no se les llama aminoácidos, sino residuos de aminoácidos o péptido.

ESTRUCTURA PRIMARIA: Corresponde a la secuencia lineal de los aac[5] que conforman la proteína. Esta secuencia contiene toda la información que define la forma tridimensional y función de la proteína

ESTRUCTURA SECUNDARIA: La conformación que adoptan aminoácidos adyacentes en la cadena polipeptídica se denomina estructura secundaria. Se distinguen dos tipos de estructura secundaria que se repiten en diferentes proteínas: α-hélice y β-plegada. (α =alfa y β=beta)

ESTRUCTURA TERCIARIA: El arreglo tridimensional de todos los átomos de una proteína es lo que se denomina estructura terciaria.

ESTRUCTURA CUATERNARIA: Unidades en estructura terciaria unidas entre si por interacciones no covalentes.

FOLD: Plegamiento de la proteína. También se utiliza para referirse a la arquitectura global de la proteína.

HIDROFOBICIDAD: Condición de repelencia al agua

MOTIVOS (MOTIF): Pequeñas regiones conservadas en proteínas de la misma familia

PÉPTIDO: Un péptido es un polímero o molécula de aminoácidos. La cadena formada por varios aminoácidos unidos recibe el nombre de polipéptido

PROTEÍNAS: Cadenas de aminoácidos que forman secuencias lineales especificas

PDB: Banco de datos de proteínas.

[5] Aac Aminoácidos

PROTEÓMICA: Ciencia que estudia todo lo relacionado con las proteínas.

RESIDUOS: Aminoácidos unidos por enlace peptídico.

SECUENCIA: Es la representación lineal de una proteína. Cadena finita y ordenada de símbolos pertenecientes a un alfabeto. El número de símbolos de la cadena representa su longitud

SECUENCIAS ORTÓLOGAS: Secuencias similares en 2 organismos diferentes que aparecen a causa de un evento que marque la aparición de una nueva especie (mutación). La funcionalidad se conserva.

SECUENCIAS PARÁLOGAS: Secuencias similares en un mismo organismo que aparecen a causa de un evento de duplicación de genes.

SECUENCIAS XENÓLOGAS: Secuencias similares que aparecen a causa de eventos de transferencia horizontal (simbiosis, virus, etc.)

Capitulo 1
Marco Teórico

1.1 Las Proteínas

Las Proteínas son compuestos orgánicos constituidos por una o más cadenas lineales de aminoácidos plegadas en forma fibrosa o globular.

Son moléculas que pueden llegar a tener un enorme grado de complejidad estructural. De esta complejidad deriva la ilimitada variedad de funciones que las proteínas desempeñan. Entre otras, intervienen en diversas funciones vitales esenciales, como el metabolismo, la contracción muscular o la respuesta inmunológica.

El término proteína deriva del griego *proteios*, que significa primero. [Starr, 2004]. Se descubrieron en 1838 y hoy se sabe que son los componentes principales de las células y que suponen más del 50% del peso seco de los animales. Se estima que el ser humano tiene unas 30.000 proteínas distintas, de las que sólo un 2% se ha descrito con detalle.

Toda proteína, (salvo raras excepciones de dos nuevos aminoácidos descubiertos la Selenocisteína y la Pirrolisina[6]) desde las humanas hasta las que forman las bacterias unicelulares, son el resultado de las distintas combinaciones entre 20 aminoácidos esenciales distintos (ver Anexo A), compuestos a su vez por carbono, hidrógeno, oxígeno, nitrógeno y, a veces, azufre.

[6] http://nar.oxfordjournals.org/cgi/content/abstract/35/15/4952

A la representación lineal de una proteína se le conoce con el nombre de ***secuencia.***

Por lineal entendemos que los aminoácidos están unidos a lo largo de una sola línea, con un principio (que convencionalmente se asigna al N-término y un final que es el C-término).[7] La cadena no presenta ramificaciones, y la unión entre aminoácidos se hace mediante enlaces peptídicos[8] , que son enlaces entre grupos amino (NH_2) y carboxilo (COOH).

Actualmente, la comparación de proteínas es un proceso clave en la identificación de características estructurales y funcionales comunes. Al comprender la estructura y funcionamiento general de las proteínas se logra entender las abundantes expresiones normales y anormales de los seres vivos.

1.1.1 Aminoácidos

Un aminoácido es cualquier molécula que contiene un grupo funcional ácido (COOH) y un grupo amino (NH_2). Existen 20 aminoácidos diferentes de cuyas combinaciones se forman las proteínas.

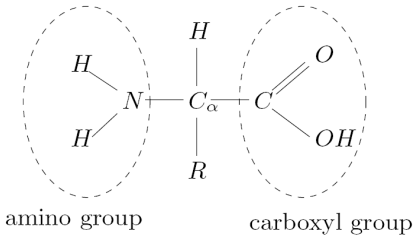


Fig. 1 Estructura de un aminoácido con el grupo amino y carboxilo [Joachim, 2007]

[7] Modelos Moleculares, 5: Proteínas – Dpto. Bioquímica y Biología Molecular –U Salamanca España
[8] Def. Enlace Peptídico –Glosario

Una clasificación

Según el grado de hidrofobicidad de los aminoácidos; éstos se pueden clasificar en Hidrofóbicos e Hidrofílicos. Una proteína de membrana, por ejemplo, debe tener una parte hidrofóbica para interactuar con la membrana plasmática y por lo menos, otra sección que interactúe con el citoplasma, la hidrofílica.

1.1.1.1 Aminoácidos Hidrofóbicos

Cadenas no polares
Cadenas laterales con enlaces C-C y C-H
No afines al agua
Ej. Alanina, Valina, Leucina, Isoleucina, Metionina, Tirosina, Fenilanina, Triptófano

1.1.1.2 Aminoácidos Hidrofílicos

Cadenas laterales polares
Cadenas laterales con átomos electronegativos, como el O y el N.
Afines al agua.

1.1.2 Aspectos Generales de la estructura de las proteinas

El nivel más básico de estructura proteica, llamado estructura primaria (fig. 2a), es la secuencia lineal de aminoácidos que está determinada, a su vez, por el orden de los nucleótidos en el ADN o en el ARN. Las diferentes secuencias de aminoácidos a lo largo de la cadena afectan de distintas formas a la estructura de la proteína. Fuerzas como los enlaces de hidrógeno, los puentes disulfuro, la atracción entre cargas positivas y negativas, y los *enlaces hidrófobicos* (repelentes del agua) e *hidrófilicos* (afines al agua) hacen que la

molécula se enrrolle o pliegue y adopte una estructura secundaria (fig. 2b), la cual, dada su importancia, será objeto de estudio mas adelante.

Cuando las fuerzas provocan que la molécula se vuelva todavía más compacta, como ocurre en las proteínas globulares, se constituye una estructura terciaria (fig. 2c) donde la secuencia de aminoácidos adquiere una conformación tridimensional. Se dice que la molécula tiene estructura cuaternaria cuando está formada por más de una cadena polipeptídica (fig. 2d), como ocurre en la hemoglobina y en algunas enzimas:

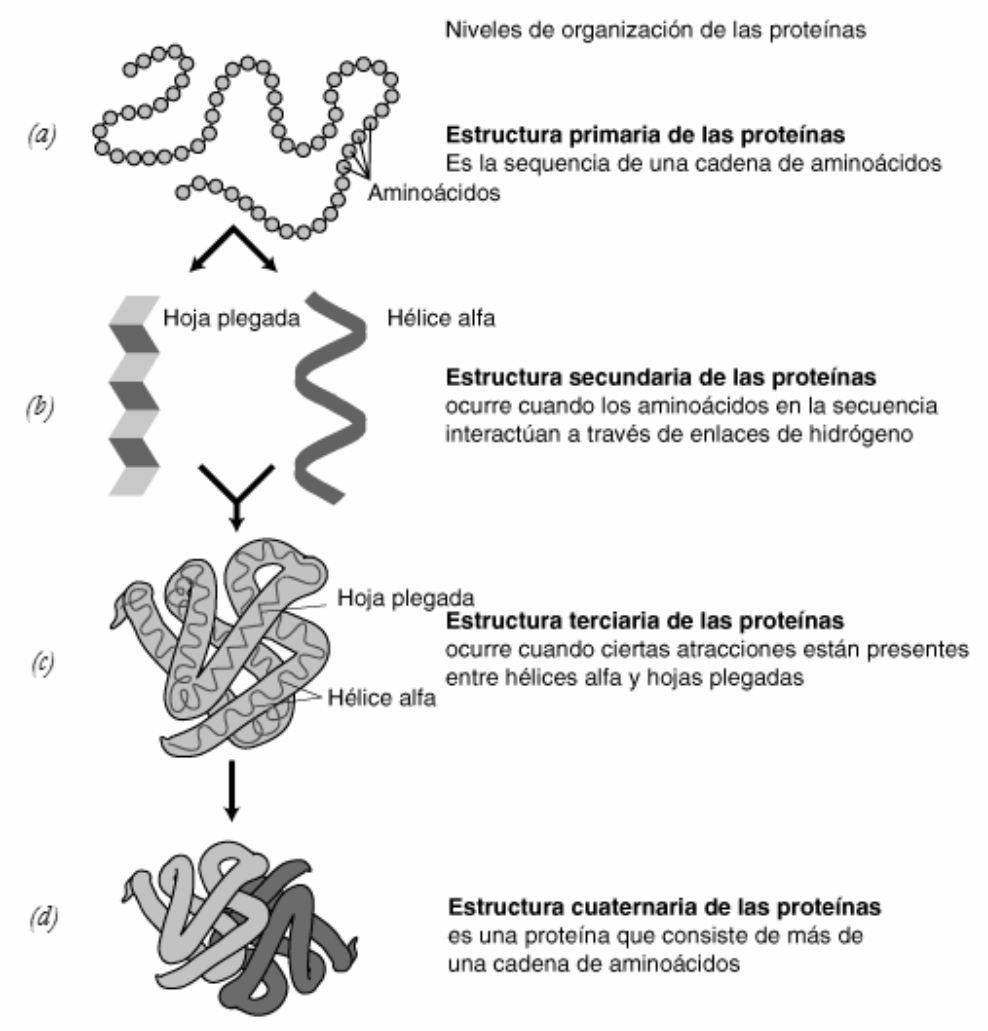


Fig. 2 Estructura de las proteinas. (a) Estructura primaria, cadena lineal (b) Las cadenas lineales se pueden organizar en estructuras con forma definida entre ellas las hélices alfa y las hojas beta plegadas. (c) Estructura terciaria, plegamiento en una subunidad (d) Estructura cuaternaria, varias subunidades.

Estructuras secundarias helicoidales y laminares

Una misma cadena polipeptídica puede adquirir diferentes estructuras secundarias en diferentes segmentos de la misma (fig. 5) según los ángulos que forman entre sí los planos peptídicos consecutivos; y esto depende del tipo de aminoácidos que están unidos, es decir de la estructura primaria de ese segmento [Garrido 2002]. Existen varios tipos de estructura secundaría periódica, entre ellos los más frecuentes son: Hélice alfa, Hoja plegada o estructura beta.

1.1.2.1 La α- Hélice

La formación de enlaces hidrógeno entre los grupos de la unión peptídica, da lugar a una estructura helicoidal dextrógira (Fig. 3).

La estructura que se obtiene es la α-hélice, en la cual los aminoácidos se disponen según un helicoide dextrógiro regular con las cadenas laterales (representadas por la letra R) dirigidas hacia el exterior de la hélice

Recibió su nombre a partir de los estudios de Pauling y Corey, que propusieron esta estructura para las α-queratinas, proteínas fibrosas que constituyen el principal contingente del pelo, de las uñas y de otros derivados dérmicos. Otras proteínas fibrosas cuya estructura es en gran parte α-helicoidal son la miosina del músculo y el fibrinógeno del plasma sanguíneo [Battaner 2001].

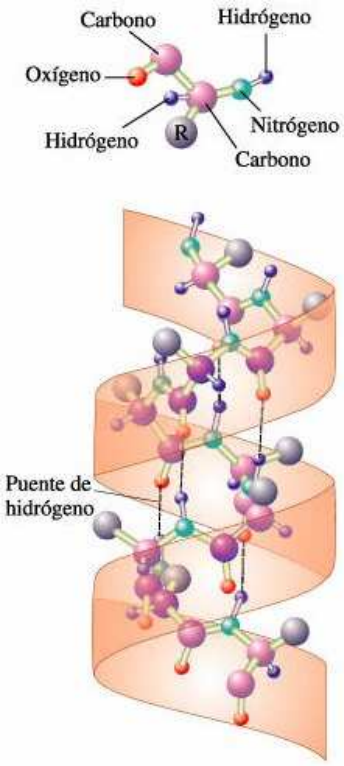


Fig. 3 Estructuras secundarias: Hélice alfa. Curtis 2000

1.1.2.2 La β –Plegada o laminar

Los estudios primitivos sobre cristalografía de rayos X de las α-queratinas llevados a cabo por **Astbury** y su grupo mostraron que el estiramiento en un ambiente húmedo de éstas hacía variar su estructura, por lo que dieron a esta nueva conformación el nombre de **Estructura** β (fig. 4), y a la proteína así modificada, el nombre de **β-queratina**.

El estudio de muchas otras proteínas ha mostrado que se trata de estructuras muy frecuentes no sólo en las proteínas fibrosas, sino también en las globulares.

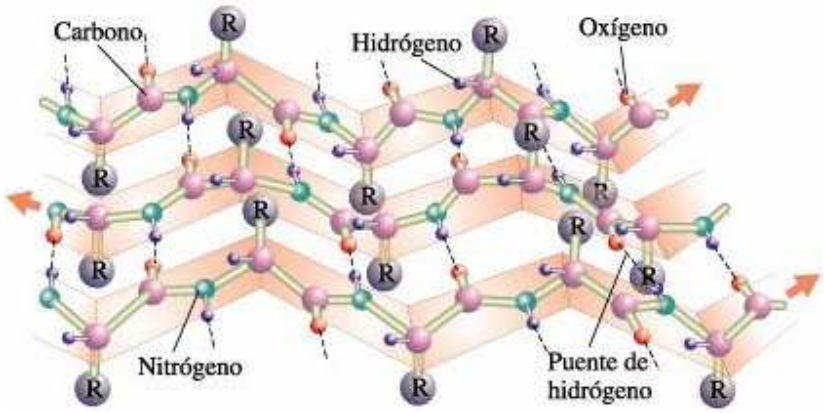


Fig. 4 Estructuras secundarias de las proteínas: Hoja plegada β. Curtis 2000

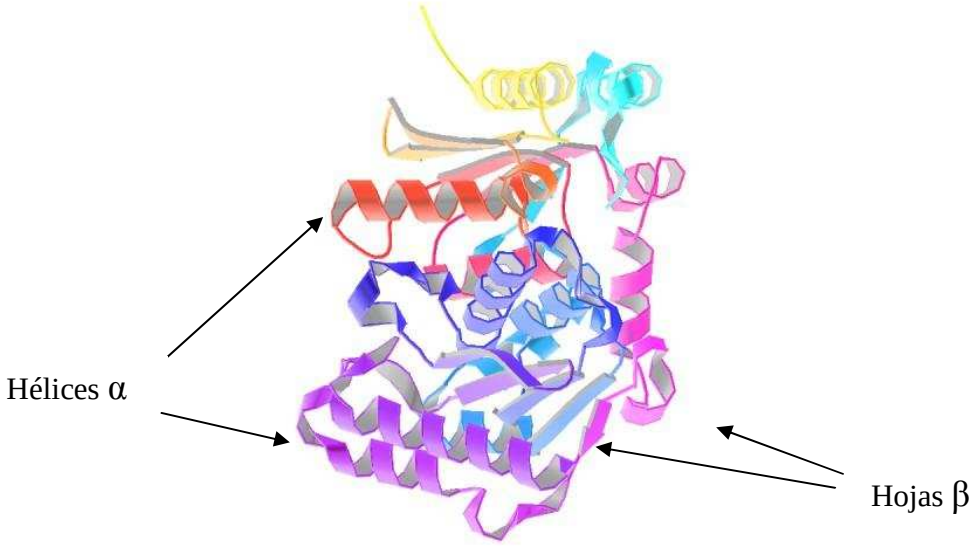


Fig. 5 Estructuras Alfa y Beta, dentro de la estructura terciaria de una proteína

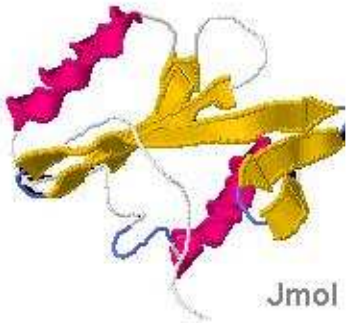


Fig. 7 Fosfolipasa D Estructuras alfa y beta

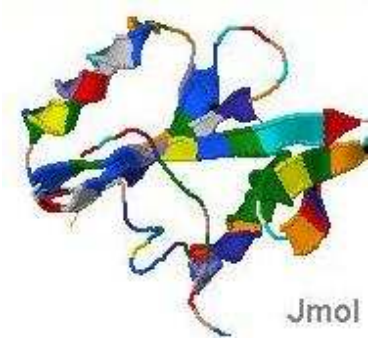


Fig. 6 Fosfolipasa D - Aminoácidos entre las estructuras secundarias

1.1.2 Clasificación de las Proteínas

Las proteínas se agrupan por comparación de estructuras secundarias regulares en: todo-α, todo-β, α / β, α + β y proteínas poco estructuradas [Levit & Chothia, 1976; Richarson, 1981]. Se estableció la siguiente división:

Proteínas todo-α: Proteínas que presentan 3 o mas hélices o bien mas de un 60% hélice α y menos de un 5% de estructuras β.

- **Bundle:** Hélices α colocadas casi paralelas o antiparalelas las unas con las otras.
- **Non-bundle:** Grupos de hélices α que no pueden ser clasificados como bundle. Estas hélices a suelen presentar una gran variación en ángulos y tamaños.
- **Poca estructura secundaria:** Proteínas pequeñas, con poca estructura secundaria regular, compactas y con una o dos hélices α.

Proteínas todo-β: Proteínas ricas en hojas β ya sean paralelas, antiparalelas o mixtas (<8% en forma de hélice α y mas del 45% en estructura β. Los dos grupos mayoritarios son sándwiches (dos hojas β retorcidas y empaquetadas) y barriles (una hoja β que va girando).
Proteínas con hélices α y hojas β: Las dos clasificaciones definen distintamente las proteínas que presentan estructuras α y β, pero siempre con mas del 30% de su estructura en hélice α y mas del 30% en hoja β.

α / β (α y β alternantes): Proteínas que presentan hélices α y hojas β de forma alternante.

α + β (α y β en disposición aleatoria): Proteínas que presentan hélices α y hojas β consecutivamente colocadas.

Proteínas con poca estructura secundaria regular o coll: son proteínas que tienen una pequeña proporción de estructuras secundarias regulares o que no pueden ser asignadas a otras clases. Suelen ser proteínas de pequeño tamaño

1.2 Comparación de Secuencias

Una secuencia es una cadena lineal, finita y ordenada de símbolos pertenecientes a un alfabeto. El número de símbolos de la cadena representa su longitud.

Un alfabeto (A) es un conjunto de símbolos diferentes usados para representar secuencias.

- ADN A = {a, c, g, t/u}
- Proteínas A = {a, c, d, e, f, g, h, i, k, l, m, n, p, q, r, s, t, v, w, y}

El objetivo (inicial y algorítmico) de comparar dos secuencias es encontrar la posición relativa de ambas en las que se produzca mayor número de coincidencias entre sus componentes. El valor "número de coincidencias o identidades" representaría la valoración de su parecido y podría darse en valores absolutos o porcentuales (normalizado) dividiéndolo entre el máximo valor posible (la longitud de la secuencia más corta).

La comparación de secuencias consiste en buscar todas las zonas de similitud significativa entre dos o más secuencias para localizar características de interés común o diferencial entre varias secuencias.

Comparar exhaustivamente dos secuencias implica comprobar cada posición de una de ellas contra cada posición de la otra. La información derivada puede relacionarse con las funciones, la estructura o evolución de células u organismos. Las relaciones entre secuencias pueden ser [CECS, 2003]:

- Homologas: secuencias similares en 2 organismos diferentes derivadas de una secuencia ancestro común.
- Ortólogas: secuencias similares en 2 organismos diferentes que aparecen a causa de un evento que marque la aparición de una nueva especie (mutación). La funcionalidad se conserva.

- Parálogas: secuencias similares en un mismo organismo que aparecen a causa de un evento de duplicación de genes.
- Xenólogas: secuencias similares que aparecen a causa de eventos de transferencia horizontal (simbiosis, virus, etc.)

1.2.1 Comparación básica por identidades

"Desplazar las secuencias" equivale a colocar una de ellas en vertical, la otra en horizontal y recorrer cada una de las diagonales de la matriz así formada, acumulando el número de coincidencias que se produzcan (Figura 8). La diagonal en la que se produzca mayor número de Identidades representará el desplazamiento relativo que mejor alinea las secuencias. El número de identidades que se contabilicen en dicha diagonal valorará el parecido (por identidades) entre las secuencias. Si representamos las coincidencias por un punto (Figura 8-b.) obtendremos lo que se denomina un DotPlot o matriz de puntos [Maizel & Lenk, 1981], en el que los fragmentos diagonales que se forman ofrecen información visual acerca de las relaciones entre las secuencias en comparación. En este caso el número de operaciones será proporcional al tamaño de la matriz formada por las dos secuencias (NxM siendo N y M las longitudes de las secuencias). Se escribe $O(N^2)$ si se asume que N y M son en promedio iguales (cuando se realizan muchas comparaciones).

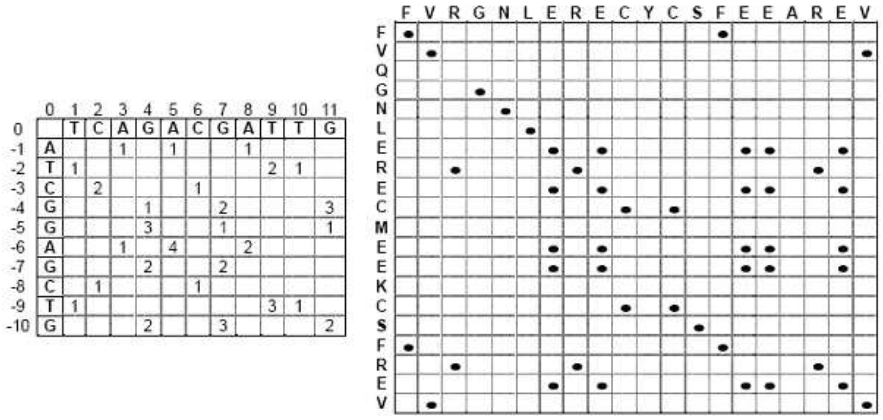


Fig. 8 Comparación básica por identidades

El mejor alineamiento por identidades y sin interrupciones para el ejemplo la figura 8a se muestra a continuación:

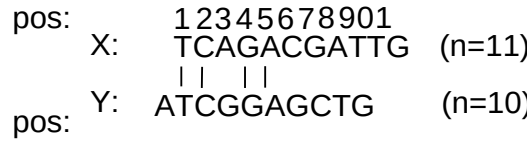


Fig. 9 Alineamiento óptimo por identidades. Caso de la figura 8a

1.2.2 Comparación por semejanza

El uso de identidades, con ser el punto de partida, es muy limitado y se muestra insuficiente para establecer la relación entre secuencias [Doolittle, 1981], entre otras cosas porque:

- La sustitución de un aminoácido por otro de propiedades similares (tamaño, carga, propiedades químicas, etc.), puede no tener gran influencia sobre la función global de la proteína.
- Es necesario considerar la inserción y la pérdida (deleción) de residuos.
- Es necesario considerar la cantidad de información que acarrea cada símbolo del alfabeto (asociada a la frecuencia de aparición del símbolo).

Por ello, las ideas actuales para comparar secuencias biológicas se basan en determinar una cierta función distancia entre las dos secuencias, que pretende informar acerca de la cercanía entre dos secuencias (en dependencia de los criterios de evaluación que se utilicen, esta será una distancia evolutiva, estructural, etc.). Por ejemplo, la distancia [Sellers, 1974] puede representar el mínimo coste de transformar la secuencia X en la Y por medio de la aplicación de una serie de transformaciones (sustituciones, inserciones, delecciones), cada una de las cuales tiene asociada un coste.

1.2.3 Alineamiento de secuencias biológicas

Para realizar una inspección detallada de la similitud de secuencias es necesario alinear las regiones comunes a ambas secuencias.
Sin embargo, esto tiene un costo: al alinear las secuencias colocándolas una junto a otra, solo se puede alinear un segmento de una secuencia con un segmento de la otra y deben mantenerse en el orden en que aparecen.

Los alineamientos sirven para [Garciarrubio et al, 2003]:

- Encontrar patrones de conservación.
- Descubrir homólogos.
- Inferir los eventos del proceso evolutivo (Mutaciones).

Algunos métodos para alineamiento de 2 secuencias son [Brungger, 2003] [Molb, 2003]:

- Visuales: si el carácter de la columna y el renglón coinciden se llena la celda. Se considera a las diagonales como regiones de similaridad.
- Fuerza bruta: Determinar todas las posibles subsecuencias para X y Y, sin embargo consume gran cantidad de tiempo.
- Programación dinámica: La idea de la programación dinámica es plantear la solución de un problema en términos de un caso más sencillo, y este en términos de otro más sencillo, y así sucesivamente (enfoque recursivo). En el caso de un alineamiento, problema para dos secuencias se puede plantear recursivamente en términos de subsecuencias de cada vez menor tamaño. El problema se resuelve en N * M cálculos. La programación dinámica evita evaluar todas las trayectorias. Se usa cuando hay muchas soluciones posibles, y se necesita encontrar una solución óptima. Dentro de los algoritmos de programación dinámica se encuentra:
- Alineamiento global (Needleman-Wunsch)
- Alineamiento local (Smith-Waterman)

- Matrices PAM o Dayhoff
- Matrices BLOSUM
- Algoritmos heuristicos. Una heurística es un procedimiento que permite hallar una solución aproximada a un determinado problema, sin asegurar que el resultado es óptimo, pero ofreciendo a cambio mayor velocidad de calculo (y en ocasiones, la única posibilidad de hallar una solución). A continuación se mencionan algunos métodos utilizados:
- Basados en palabras (k tupla)
- FASTA
- BLAST (Basic Local Alignment Search Tool)

Es necesario establecer la diferencia entre comparar y alinear en el contexto de la bioinformática. La comparación implica que las secuencias se encuentren en una posición dada, mientras que la alineación implica encontrar la posición óptima para conseguir un mejor resultado en la comparación.

1.4 Hydrophobic Cluster Análysis – HCA

1.4.1 ¿Qué es HCA?

Hydrophobic Cluster Analysis (HCA) es una forma eficiente de comparar secuencias altamente divergentes a través de la información de la estructura secundaria implícita derivada de los clusters hidrofóbicos. [Eudes, Le Tuan, Delettré, Mornon, Callebaut 2007][9]

Este método científico, se usa principalmente para los casos donde los métodos lineales son limitados: En la zona de baja similaridad entre proteínas, conocida como ***twligth-zone***.

1.4.2 Zona Twiligth

Se denomina ***Zona Twiligth*** al rango de proteínas que al ser comparadas con los métodos lineales tales como BLAST o FASTA presentan un porcentaje de similitud entre 25 a 30%

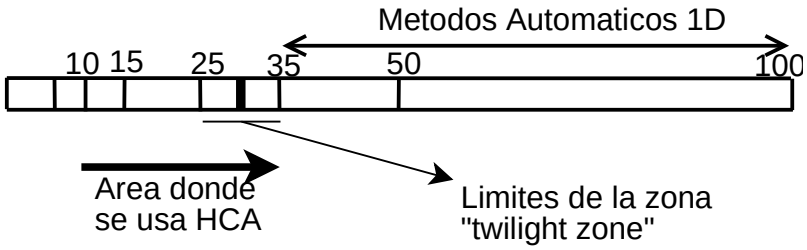


Fig. 10 Intervalo Twiligth Zone

1.4.3 Descripción del Método HCA

[9] *A generalized analysis of hydrophobic and loop clusters within globular protein sequences*

Siete residuos hidrofobicos (V, I, L, F, M, Y, W) integran el alfabeto hidrofóbico de HCA. Los aminoácidos hidrofóbicos (V, I, L, F, M, Y, W) se codifican como 1, mientras los demás se codifican con 0. De esta forma, cualquier clúster hidrofóbico comienza y finaliza en 1, además hay separación de clúster si existe 4 o mas 0`s consecutivos o si aparece alguna prolina (P) (fig. 11).

Fig. 11 Secuencia lineal y código binario.

De esta forma, una secuencia se puede representar como un conjunto de ceros y unos según el tipo de aminoácido que corresponda (cero para lo hidrofilicos y 1 para los Hidrofóbicos).

1.4.3.1 Código P

Además del código binario existe otro código denominado ‘Peitsch code’ o ‘código P’ definido como la suma de las potencias de 2, con el índice de acuerdo a la posición de cada número en el código binario (la última posición corresponde a cero), cada potencia es multiplicada por el código binario:

$$\mathbf{110101} => \mathbf{1} \times 2^5 + \mathbf{1} \times 2^4 + \mathbf{0} \times 2^3 + \mathbf{1} \times 2^2 + \mathbf{0} \times 0^1 + \mathbf{1} \times 2^0 = 53)$$

1 Código P de un cluster

El código P, permite una alternativa sencilla de descripción de los clúster, por lo tanto es muy utilizado en términos de almacenamiento y clasificación computacional, especialmente para clúster muy largos. [Eudes, 2007],

1.4.3.2 Resumen de la nomenclatura de clusters

Dicha codificación binaria, que es aplicable a la ***graficación*** de matrices **HCA** se traducen en ciertas reglas básicas de nomenclatura las cuales han sido estudiadas y probadas desde sus inicios en el año 1987 aproximadamente. Estas reglas son:

- Todo cluster inicia y finaliza en un aminoácido Hidrofóbico.
- Un cluster se reconoce por la separación existente entre aminoácidos hidrofóbicos la cual debe ser de al menos 4 aminoácidos no Hidrofóbicos conocidos como Aminoácidos Hidrofílicos.
- La presencia del aminoácido prolina (P) incide directamente en la ruptura de un cluster en dos clusters diferentes.

1.4.3.3 Representación bidimensional

El principio básico para la grafica helicoidal para HCA fue descrito por Gaboliaud et al. (1987). Una secuencia de aminoácidos sobre una cinta, la cual es enrollada helicoidalmente sobre un cilindro con un promedio de 3.6 aminoácidos por vuelta para un α-hélix (fig. 12a), en el turno 5, esto es, para el aminoácido 19 se observa que él se encuentra ubicado paralelamente al primer aminoácido (fig. 12b).

La matriz inclinada que se obtiene, se replica sobre sí misma, en orden a restituir el entorno total que cada aminoácido tiene sobre la representación α –hélice. (fig. 12c). Los aminoácidos hidrofóbicos tienden a conformar clusters o agrupaciones. Estos clusters corresponden a las caras internas de estructuras regulares secundarias en las proteínas (fig. 12d).

Como compuesto orgánico de una o más cadenas polipeptídicas, una proteína puede presentar diferentes plegamientos, lo cual hace de la proteína un ente tridimensional.

Como se ve a continuación, segmentos de la proteína pueden llevarse a representaciones 2D por medio de diagramas (fig. 13)

Actualmente, el método HCA está adaptado para que el reconocimiento se haga basado en la habilidad humana del sistema ojo-cerebro y requiere de gran entrenamiento del experto el reconocer, descifrar y asociar correctamente las similitudes entre complejas imágenes con desigual información biológica. [C. Gaboriaud et al, 1987; I. Callebaut et al, 1997] (fig. 14). A partir de la comparación visual, se puede obtener el porcentaje de identidad entre las proteínas, entendido como la relación entre el número de aminoácidos conservados en ambas secuencias y la longitud en la proteína mas larga.

$$\% \text{ Identidad} = \frac{\text{Número Aminoácidos Conservados}}{\text{Longitud proteína más larga}}$$

Ec. 1 Fórmula para el cálculo del porcentaje de identidad entre secuencias

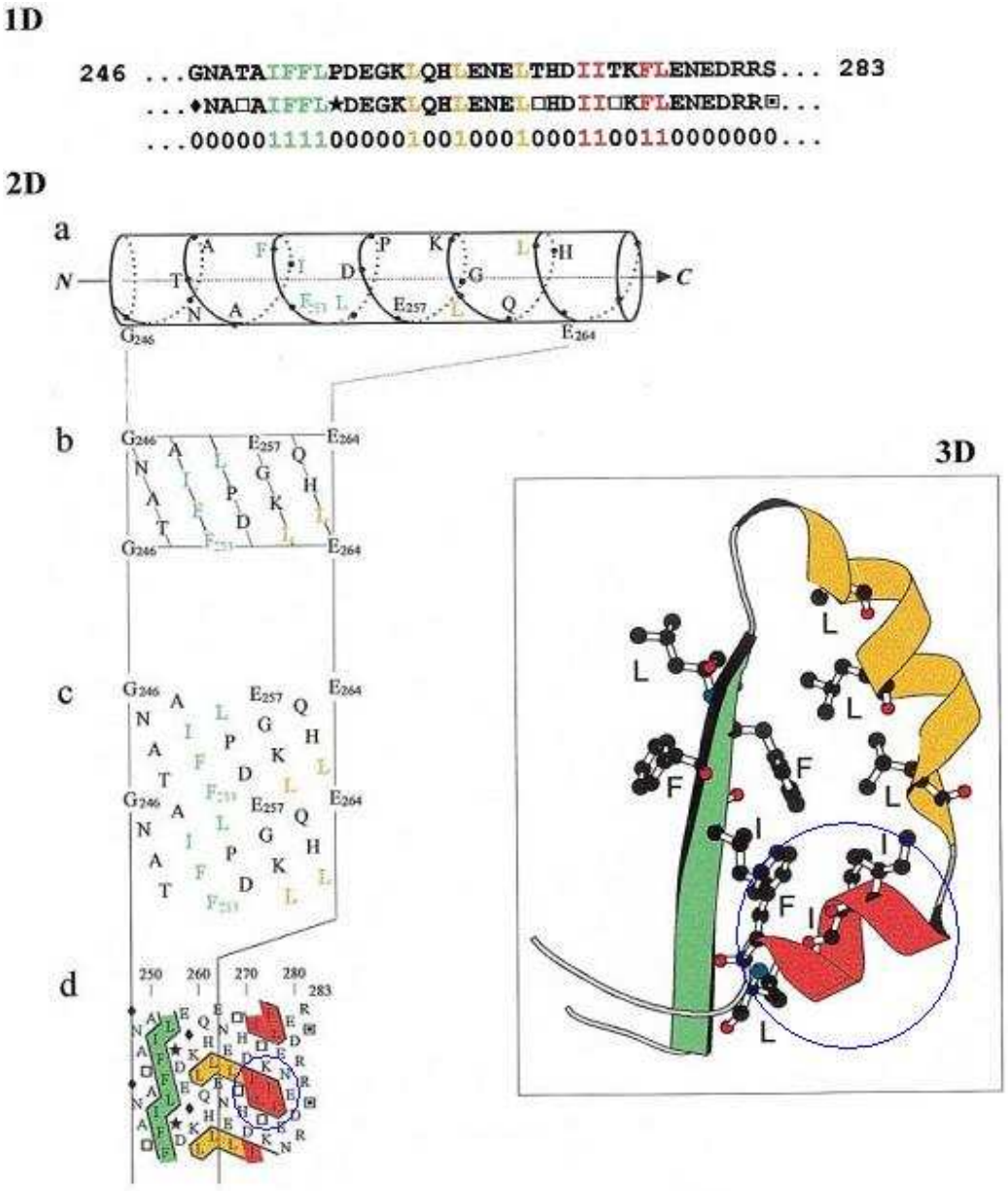


Fig. 12 Análisis de Clústers Hidrofóbicos: Entre la secuencia lineal y la estructura. Principios del diagrama 2D. [C. Gaboriaud et al, 1997]. (a) Enrollamiento de un segmento lineal de aminoácidos de la proteína humana α 1-antitripsina formando una hélice. (b) Construcción de la matriz. Los aminoácidos 246 y 264 quedan alineados paralelamente. (c) Replicación de la matriz (d) Esta conformación bidimensional permite identificar estructuras tridimensionales agrupadas en clusters de aminoácidos hidrofóbicos, como se observa para la hélice α y para la estructura β (verde).

(a)

(b)

(c)

(d)

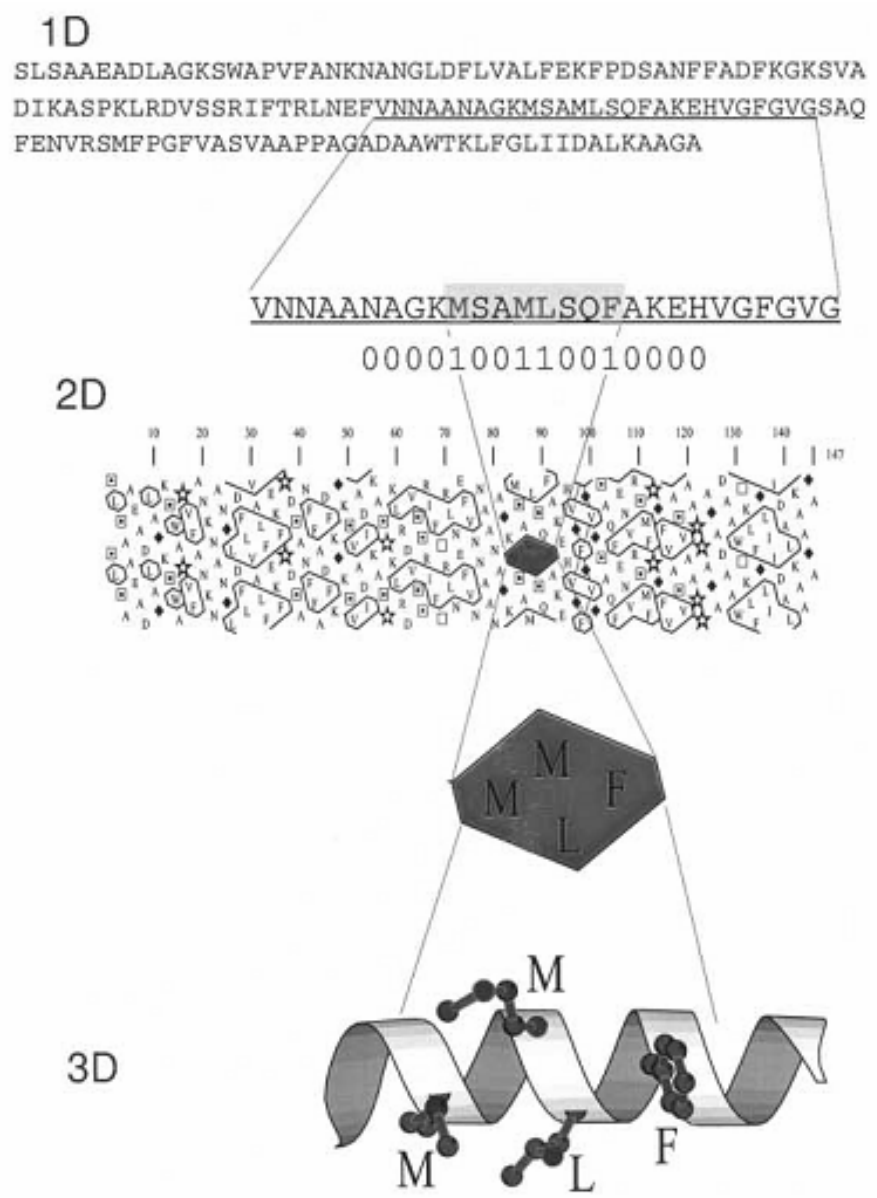


Fig. 13 Representación de clusters en HCA. CMLS 53 (1997), Birkhäuser Verlag, CH-4010 Basel/Switzerland. (a) Secuencia lineal 1D. (b) Clusters hidrofóbicos en código binario. (c) Representación 2D de la proteína. (d) Representación 3D. Hélice α.

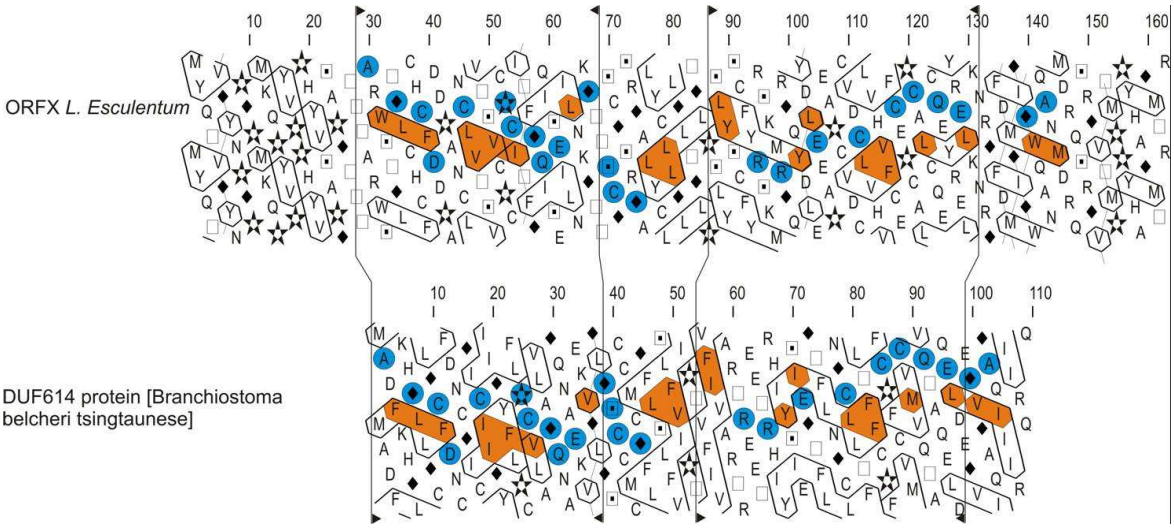


Fig. 14 Ejemplo de comparación de 2 secuencias usando el análisis de clusters hidrofóbicos.
Identidad = 31/ 115 = 27%

Capítulo 2
Construcción del Modelo Computacional

2.1 Metodología

Para el modelado inicial de la herramienta se utilizó UML (Lenguaje Unificado de Modelado). Un proceso de desarrollo que reúne el conjunto de actividades necesarias para transformar los requisitos en un conjunto de artefactos que representan el producto final. El Lenguaje de Modelado Unificado UML es un lenguaje para visualizar, especificar, construir y documentar los artefactos de un sistema que involucra una gran cantidad de software [Jacobson et al., 2000]. El Proceso unificado esta dirigido por casos de uso, centrado en la arquitectura, y es iterativo e incremental. Se conoce como casos de uso un conjunto de secuencias de acciones, que conducen a un resultado observable de interés para un usuario.

Uno de los pasos importantes dentro del proceso unificado es definir claramente los requisitos del usuario, para este proyecto ha sido planteado como la construcción de una herramienta computacional que permita comparar secuencias de proteínas utilizando la técnica HCA. Para construir una versión del producto siguiendo esta metodología se realizaron cuatro fases, de la siguiente manera:

1. Inicio: Establecer la planificación del proyecto.
2. Elaboración: Establecer un plan para el proyecto y una arquitectura correcta.

3. Construcción: Desarrollar el sistema.
4. Transición: Proporcionar el sistema a sus usuarios finales

Se utilizó el proceso de desarrollo unificado de software como metodología para llevar a cabo la construcción del sistema [Jacobson et al., 2000].

A continuación se muestra el diagrama de casos de uso (Figura 15), posteriormente se describe el comportamiento del sistema mediante el flujo normal y el alterno.

2.1.1 Diagrama de casos de uso

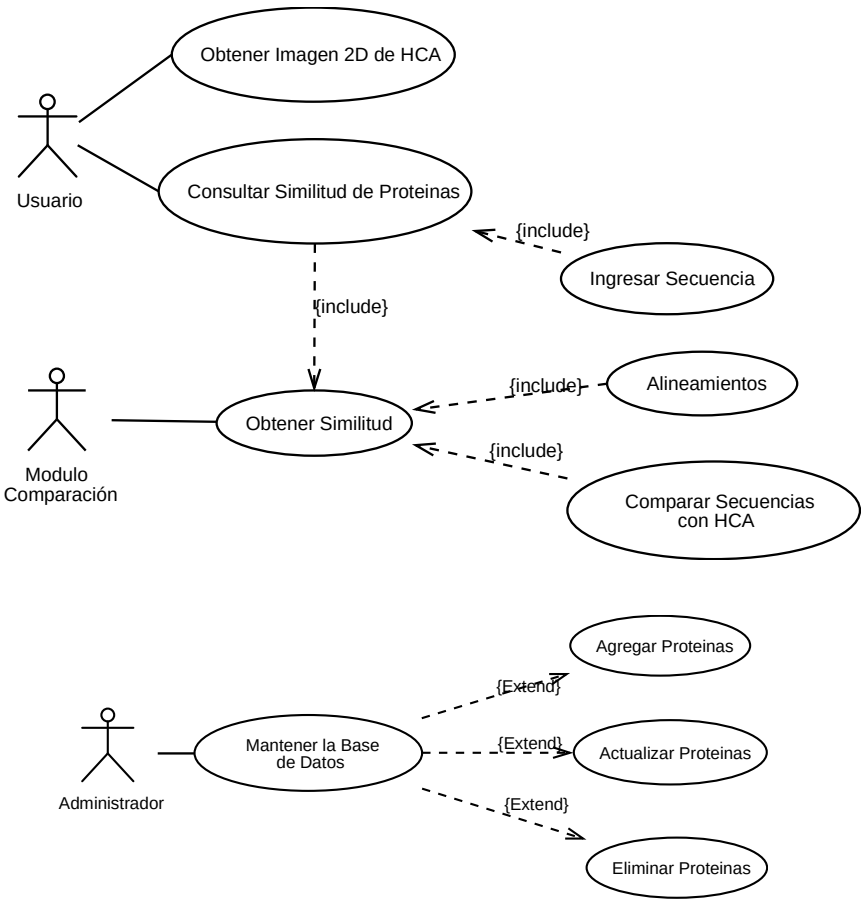


Fig. 15 Diagrama de casos de uso

2.1.2 Descripción de casos de uso

- **Caso de Uso: Mantener la base de datos**

Descripción: Este caso de uso es iniciado por el Administrador y permite hacerle mantenimiento al sistema agregando proteínas, actualizando proteínas o eliminando proteínas.
Actores: Administrador
Prioridad: Alta
Riesgo: Alto
Precondiciones: Disponible conexión al sistema.

Flujo normal de eventos:

Actores	**Sistema**
El Administrador ingresa identificador al sistema.	
	El sistema verifica que este identificador sea válido
	El sistema muestra las actividades que el Administrador puede realizar: agregar proteínas, actualizar datos de proteínas o eliminar proteínas.
El Administrador selecciona la actividad y la desarrolla	
	Se muestran los resultados obtenidos e indica que el proceso ha terminado

Flujos alternos: Si el identificador es inválido, el sistema no le permite el acceso.

Postcondiciones: La información se almacena en la base de datos.

Requerimientos no funcionales: La interfaz presentada debe ser agradable

Descripción de los casos de uso especializados, derivados del caso de uso Mantener la Base de Datos.

- **Caso de uso: Agregar proteínas.**

Descripción: Este caso de uso permite agregar nuevas proteínas al sistema.

Actores: Administrador

Prioridad: Alta

Riesgo: Alto

Precondiciones: Haber iniciado el caso de uso Mantener la Base de Datos

Flujo normal de eventos

Actores	Sistema
El Administrador ingresa la proteína nueva que quiere que este en la base de datos	
El Administrador envía la información	
	El sistema valida: Verifica que los datos ingresados sean coherentes.
	Agrega nueva proteína a la base de datos

Postcondiciones: Regresar al menú principal de mantener la Base de Datos

- **Caso de uso: Actualizar proteínas del sistema**

Descripción: Este caso de uso permite modificar y actualizar información de las proteínas que estén en el sistema.

Actores: Administrador

Prioridad: Alta

Riesgo: Alto

Precondiciones: Haber iniciado el caso de uso Mantener la Base de Datos.

Flujo normal de eventos

Actores	**Sistema**
El Administrador selecciona proteína para modificar o para actualizar.	
	El sistema muestra los campos que se pueden modificar.
El Administrador hace las modificaciones pertinentes y actualiza la información.	
	El sistema muestra los cambios realizados
	El sistema guarda los cambios realizados en la base de datos.

Flujos alternos: El Administrador puede cancelar la operación antes de guardar los cambios y de esta forma no afecta la base de datos.

Postcondiciones: Regresar al menú principal de Mantener la Base de Datos.

- **Caso de uso: Eliminar proteínas del sistema**

Descripción: Este caso de uso permite eliminar las proteínas de la base de datos que el Administrador considere no sean necesarias para el estudio.

Actores: Administrador

Prioridad: Alta

Riesgo: Alto

Precondiciones: Haber iniciado el caso de uso Mantener la Base de Datos.

Flujo normal de eventos

Actores	**Sistema**
El Administrador indica cual proteína va a eliminar	
	El sistema elimina la proteína seleccionada.

Postcondiciones: regresar al menú principal de Mantener la Base de Datos.

- **Caso de Uso: Consultar similitud**

Descripción: Este caso de uso es iniciado por el usuario y permite obtener el grado de similitud con las proteínas de la base de datos.

Actores: Usuario

Prioridad: Alta

Riesgo: Bajo

Precondiciones: Disponible conexión al sistema, haber realizado el caso de uso Ingresar Proteína.

Flujo normal de eventos:

Actores	**Sistema**
El usuario elige consultar similitud para la proteína que quiera estudiar.	

	El sistema muestra la lista de las proteínas con el grado de similitud respecto a la proteína de referencia.

Requerimientos no funcionales: La interfaz presentada debe ser agradable.

Descripción de los casos de uso especializados derivados del caso de uso Consultar Similitud.

- **Caso de uso: Ingresar secuencia de proteína**

Descripción: Este caso de uso permite ingresar la secuencia de la proteína que se piensa caracterizar y obtener similitud.
Actores: Usuario
Prioridad: Alta
Riesgo: Medio
Precondiciones: Haber accedido al sistema.

Flujo normal de eventos

Actores	Sistema
El usuario ingresa la secuencia de la proteína a comparar.	
	El sistema analiza la secuencia.
	El sistema muestra un mensaje de confirmación.

Flujos alternos: Si la secuencia de la proteína no es coherente, el sistema muestra un mensaje al usuario para que introduzca una secuencia válida.
Postcondiciones: Ir al menú principal de Consultar Similitud.

- **Caso de uso: Obtener similitud**

Descripción: Este caso de uso permite obtener el grado de similitud entre la proteína ingresada y las proteínas de la base de datos

Actores: Modulo de Comparación, Usuario.

Prioridad: Alta

Riesgo: Medio

Precondiciones: haber iniciado el caso de uso Consultar Similitud.

Flujo normal de eventos

Actores	**Sistema**
	El sistema presenta las dos secuencias a comparar.
El Modulo de comparación coge las dos secuencias y da el grado de similitud entre las proteínas.	

Poscondiciones: Regresar a consultar similitud.

2.2 Generación de gráficas bidimensionales en PostScript

Para la representación de los resultados del algoritmo se ha decidido presentarlos en un documento con formato postscript y pdf, el cual es generado a partir del primero por medio de la instrucción ps2pdf, propia de linux.

Formato de Salida de Datos

Postscript es un lenguaje de descripción de página, lo que significa que por medio de él se puede detallar el contenido de un documento particionado en páginas. Este lenguaje de descripción es empleado comúnmente por impresoras y visto como gráfico, es inherentemente de naturaleza vectorial. Esto significa que a diferencia de una imagen pixelada (la cual es descrita por una matriz de muestras equi-espaciadas de intensidad de luz), ésta contiene explícitamente la información que permite reconstruir los elementos que la componen como líneas, textos, formas, colores, etc.

El lenguaje *postscript* es de uso común en aplicaciones que requieren este tipo de recursos y su estandarización está liderada por *Adobe*. Actualmente se encuentra en su versión 3. Los archivos postscript pueden ser creados con formato de texto plano, lo cual permite visualizar su código fuente en cualquier editor de texto. Para la creación del archivo desde el programa en C, se ha usado la librería estándar de C *fstream* que permite escribir y añadir progresivamente texto a un archivo de texto.

Estructura de un archivo PostScript

Los archivos *postscript* guardan una estructura común, que consiste en una cabecera, donde se hacen algunas definiciones de procedimientos y redefiniciones de nombres de algunos comandos, con el fin de facilitar el desarrollo de la segunda parte que corresponde a la información propiamente dicha de la imagen o el documento.

2.2.1 Gráfica HCA

Para generar la imagen y presentar los resultados es necesario conocer detalladamente la organización que tienen las componentes de esta.

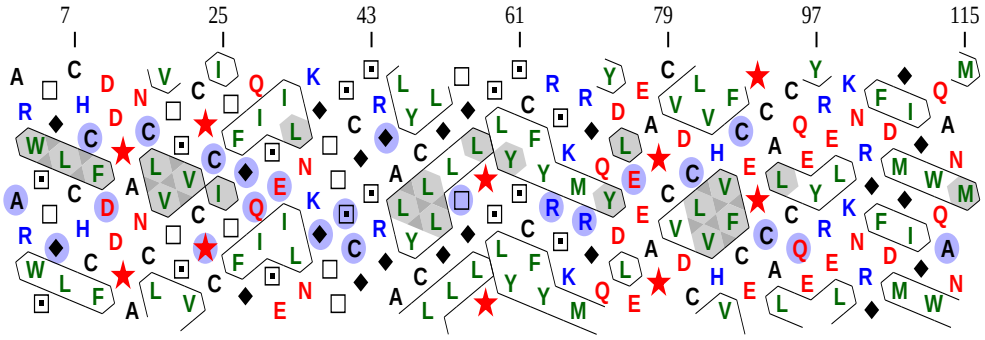


Fig. 16 Modelo de Gráfica HCA

Una imagen HCA es un arreglo bidimensional de la secuencia de aminoácidos de la proteína (sección 1.3.3), la cual se organiza en patrones que agrupan 4, 3, 4, 4 y 3 elementos respectivamente, para formar un patrón geométrico que se repite cada 18 aminoácidos. Este patrón es producto del arreglo generado al desplegar la secuencia lineal alrededor de una hélice o cilindro y tomando un corte vertical. Así mismo para facilitar la visualización de los aminoácidos vecinos, se replica el arreglo bidimensional en la parte inferior de la misma, dando como resultado una imagen como la mostrada en la figura 16.

Es importante notar que si se expresa la posición de cada aminoácido en sus componentes *x* y *y*, la distancia horizontal entre dos aminoácidos consecutivos de la cadena es constante, mientras la componente *y* de la posición de los aminoácidos varía cíclicamente según la secuencia 0 – 5 – 10 – 15 – 2 – 7 – 12 – 17 – 4 – 9 – 14 – 1 – 6 – 11 – 16 – 3 – 8 –13 – 0, lo cual equivale a *5k(modulo 18)* con *k=0,1,2,…,17*. De esta forma se obtiene el arreglo mostrado en la figura 16 donde se observa la vecindad de un aminoácido con 6 elementos de la secuencia.

Partiendo de la ubicación de los aminoácidos, se puede calcular como los puntos medios, los vértices que forman las figuras de las líneas y las sombras.

Generación del Archivo

Para la generación del archivo es importante tener en cuenta el orden en el que se describen los componentes de la imagen. En el lenguaje *postscript*, las figuras que se imprimen primero, pasan a formar parte del fondo y cualquier objeto nuevo puede sobrescribir el antiguo. Por eso para este archivo, se ha impreso inicialmente las sombras de la imagen, posteriormente las líneas y finalmente las letras, que deben ser visibles ante cualquier circunstancia.

2.3 Desarrollo de los algoritmos de comparación

A continuación se muestra la estrategia seguida para la búsqueda en la base de datos utilizando HCA.

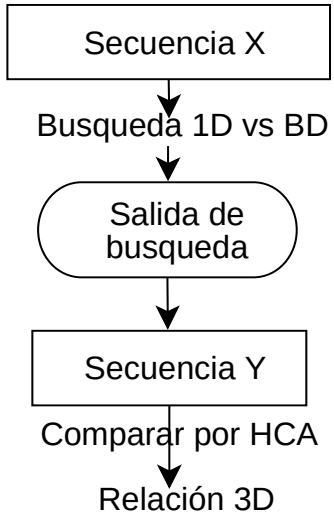


Fig. 17 Estrategia de comparación - Algoritmo

Se parte de la secuencia X desconocida, la cual se compara con una serie de proteínas almacenadas en la base de datos obteniendo como resultado una lista donde se presenta las proteínas con su grado de similitud respecto a la proteína desconocida.

Posteriormente, el científico en base a dicha lista sugerida escoge aquellas proteínas que desea comparar con el Algoritmo de Comparación Basado en HCA, el cual genera la gráfica con el alineamiento basado en HCA y el porcentaje de identidad aproximado entre las dos secuencias alineadas.

2.3.1 Comparación rápida con HCA

El objetivo de este algoritmo es poder de forma rápida hacer el proceso de comparación de secuencias de proteínas, partiendo de un concepto fundamental que tiene en cuenta hca en su análisis: el código P.
Básicamente se analiza el comportamiento que tiene el código P a lo largo de la Proteína. Se genera un código único para cada proteína y posteriormente se compara sus códigos utilizando el concepto de la comparación básica por identidades (Descrito en el capitulo 2), de esta forma se puede hacer una jerarquización por similitudes a partir de los resultados de este método, para posteriormente aplicar un método de comparación mas detallado y que tiene en cuenta algunos elementos importantes a tener en cuenta.

Para el primer algoritmo donde se hace un sondeo rápido de comparación de la proteína desconocida con la base de datos, se considera el comportamiento entre clústers hidrofóbicos a lo largo de la proteína (figura 18-b). A cada clúster hidrofóbico le corresponde un código P, y por tanto a cada proteína le corresponde una lista de códigos p únicos para dicha proteína (figura 18-c).

Teniendo esta lista de códigos P para cada proteína, se obtiene una firma o código único binario para cada secuencia, en donde se analiza la relación de tamaño entre los clústers hidrofóbicos que conforman la proteína de la siguiente forma:
Si el código P del clúster es mayor que el siguiente, se tiene 1, si es falso, 0 (figura 18-d).
De esta forma se cuenta para cada secuencia de proteína con un código único que presenta el comportamiento y variación de los clústers hidrofóbicos a lo largo de la secuencia de la proteína.

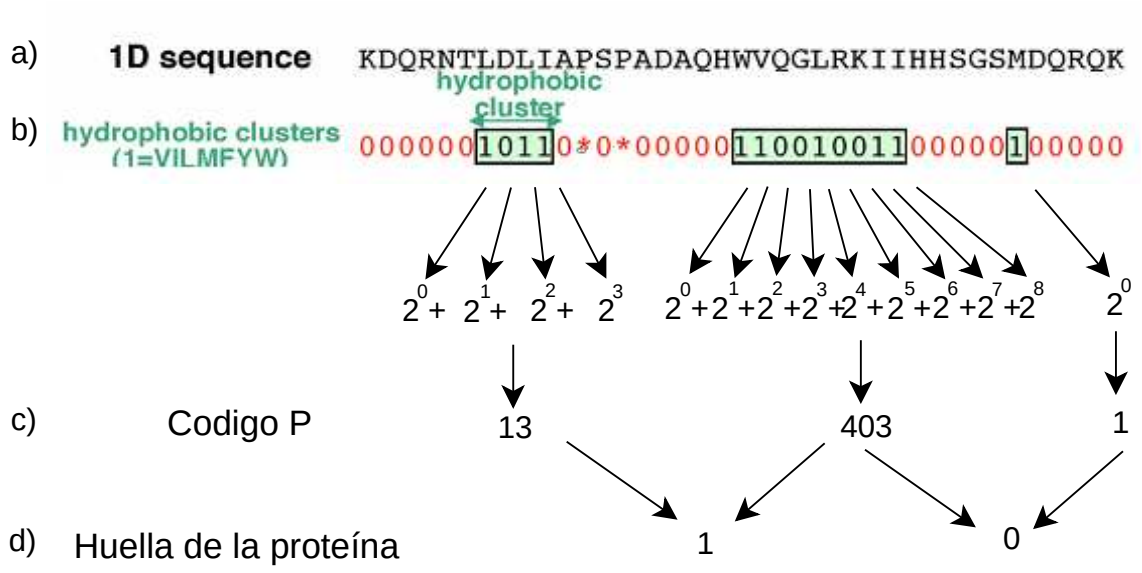


Fig. 18 Proceso de generación del código único para cada proteína. a) secuencia lineal de la proteína. b) código binario de la secuencia identificando los respectivos clusters hidrofóbicos, c) Código P para cada cluster hidrofóbico. d) Huella de la proteína (Se compara los códigos P de los clusters hidrofóbicos; si el código p del siguiente cluster es mayor, el valor es 1, si no, 0).

Posteriormente se realiza una comparación básica por identidades (corrimiento de secuencias) entre estos códigos únicos de las secuencias, obteniendo el mejor alineamiento tal como se muestra a continuación:

Huella proteína 1: **11010101**
Huella proteína 2: 11**1101011**111

Para este caso el mejor alineamiento se obtuvo con los elementos en negrilla, finalmente, el porcentaje de similitud se obtiene al dividir este número de identidades (elementos en negrilla) entre la longitud de la huella mas corta, para este caso se obtiene 87.5% de similitud, lo cual indica que es muy probable que sean proteínas homologas.

Se realizó una serie de comparaciones para analizar el comportamiento del algoritmo (Comparación rápida con hca) planteado, siguiendo la siguiente metodología:

- Comparación de una proteína consigo misma – 100%
- Comparación de una proteína con cualquiera de sus mitades – 50%
- Comparación de una proteína con parte de ella
- Comparación de 2 proteínas diferentes, teniendo de antemano el porcentaje de similitud dado por el análisis del experto.

Para los 3 primeros casos el algoritmo de comparación rápida con hca obtiene un acierto del 100%. A continuación se muestra un ejemplo para la ORFX L. Esculentum comparándola consigo mismo.

- **Comparación de una proteína consigo misma – 100%**

MYQTVGYNPGPMKQPYVPPHYVSAPGTTTARWSTGLCHCFDDPANCLVTSVCPCI
TFGQISEILNKGTTSCGSRGALYCLLGLTGLPSLYSCFYRSKMRGQYDLEEAPCVCL
VHVFCEPCALCQEYRELKNRGFDMGIGWQANMDRQSRGVTMPPYHAGMTR

MYQTVGYNPGPMKQPYVPPHYVSAPGTTTARWSTGLCHCFDDPANCLVTSVCPCI
TFGQISEILNKGTTSCGSRGALYCLLGLTGLPSLYSCFYRSKMRGQYDLEEAPCVCL
VHVFCEPCALCQEYRELKNRGFDMGIGWQANMDRQSRGVTMPPYHAGMTR

Comparación rápida hca: 100% similitud

Para el último caso se encuentra que con los ejemplos realizados los porcentajes de similitud están muy cercanos a los obtenidos por el experto, esto teniendo en cuenta que este algoritmo realiza un sondeo rápido y por tanto no considera toda la información que se pudiera utilizar. Es necesario aclarar que las proteínas deben tener un tamaño lo suficiente para que haya una gran cantidad de clusters y se pueda contar con mayor información del comportamiento de los cluster a lo largo de la secuencia de la proteína.

A continuación se muestra 2 ejemplos donde se analiza la comparación de 2 proteínas diferentes:

ORFX L. Esculentum

MYQTVGYNPGPMKQPYVPPHYVSAPGTTTARWSTGLCHCFDDPANCLVTSVCPCI
TFGQISEILNKGTTSCGSRGALYCLLGLTGLPSLYSCFYRSKMRGQYDLEEAPCVDC
LVHVFCEPCALCQEYRELKNRGFDMGIGWQANMDRQSRGVTMPPYHAGMTR

DUF614 protein [Branchiostoma belcheri tsingtaunese]

MADFKHGLLGCFDNCGICIIGYFLPCVLAGQNAEKVGLGSCCMCGFLSLFVIPTVFI
VARTREETRHIYSIEGTFLNGCLLTFFCPFCVMVQTAQELDEGVGAQIIIRQ

Método visual hca: 27% similitud
Comparación rápida hca: 20% similitud

4Q21 c-Ha-ras1 p21 protein [Homo Sapiens]

MTEYKLVVVGAGGVGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVIDGETCLLDIL
DTAGQEEYSAMRDQYMRTGEGFLCVFAINNTKSFEDIHQYREQIKRVKDSDDVPM
VLVGNKCDLAARTVESRQAQDLARSYGIPYIETSAKTRQGVEDAFYTLVREIRQHK
LRKLNPPDESGPGCMSCKCVLS

1WKQ Chain B, Crystal Structure Of Bacillus Subtilis Guanine Deaminase

MHHHHHHAMNHETFLKRAVTLACEGVNAGIGGPFGAVIVKDGAIIAEGQNNVTTS
NDPTAHAEVTAIRKACKVLGAYQLDDCILYTSCEPCPMCLGAIYWARPKAVFYAA
EHTDAAEAGFDDSFIYKEIDKPAEERTIPFYQVTLTEHLSPFQAWRNFANKKEY

Método visual hca: 15% similitud
Comparación rápida hca: 20% similitud

Resumen de resultados método Comparación rápida con HCA

Tabla 1 Resumen de resultados método Comparación rápida con HCA

Proteína conocida	**Proteína desconocida**	**Método visual HCA**	**Comparación Rapida HCA**
ORFX L. Esculentum	ORFX L. Esculentum	100%	100%
ORFX L. Esculentum	DUF614 protein [Branchiostoma belcheri tsingtaunese]	27% similitud	20% similitud
4Q21 c-Ha-ras1 p21 protein [Homo Sapiens]	1WKQ Chain B, Crystal Structure Of Bacillus Subtilis Guanine Deaminase	15% similitud	20% similitud

2.3.2 Propuesta: Algoritmo de Comparación basado en el Análisis Hidrofóbico de Clusters

2.3.2.1 Metodología

Al iniciar el diseño de un algoritmo que permitiera comparar secuencias de proteínas, llegando a resultados similares a los proporcionados por la comparación que actualmente se realiza de forma manual, usando el análisis hidrofóbico de clusters; se partió de la base que los algoritmos lineales preexistentes; son métodos que se basan en ciertas determinaciones para el hallazgo de las similitudes, las cuales, en su mayoría, no están relacionadas con el método HCA. Como se observa en el capítulo correspondiente a algoritmos de alineamiento.

El Análisis Hidrofóbico de Clusters, como tal, busca encontrar similitudes en proteínas que según los métodos de comparación convencionales no presentan un alto grado de identidad. Aquí es donde radica su complejidad.

En pruebas realizadas de la mano con algunas personas diestras en la comparación visual por medio del método HCA, se encontró que una misma persona puede realizar para cada par de proteinas un número variable de alineamientos según la habilidad del experto. Lo anterior, demuestra que hay un componente de subjetividad que no es fácilmente predecible.

Sin embargo, se encontró que los criterios utilizados para dichas comparaciones, los cuales se emplean por los expertos de una manera intuitiva producto de la praxis, podrían deducirse y fueron ellos los que dieron la clave para generar un algoritmo con un alto grado de acierto.

A la hora de implementar dichos criterios, se plantearon alternativas que partiendo de un análisis lineal permitieran llegar al algoritmo de comparación final, se pasó por varias fases. En alguna de ellas, se planteó la posibilidad de analizar las proteinas alineándolas a lo largo de una matriz de tamaño n y considerando las coincidencias a lo largo de su diagonal:
En la siguiente figura:

X	Corresponde a emparejamientos entre aminoácidos hidrofílicos idénticos
1	Corresponde a emparejamientos entre aminoácidos hidrofóbicos idénticos
	Corresponde a emparejamientos entre aminoácidos hidrofóbicos no idénticos

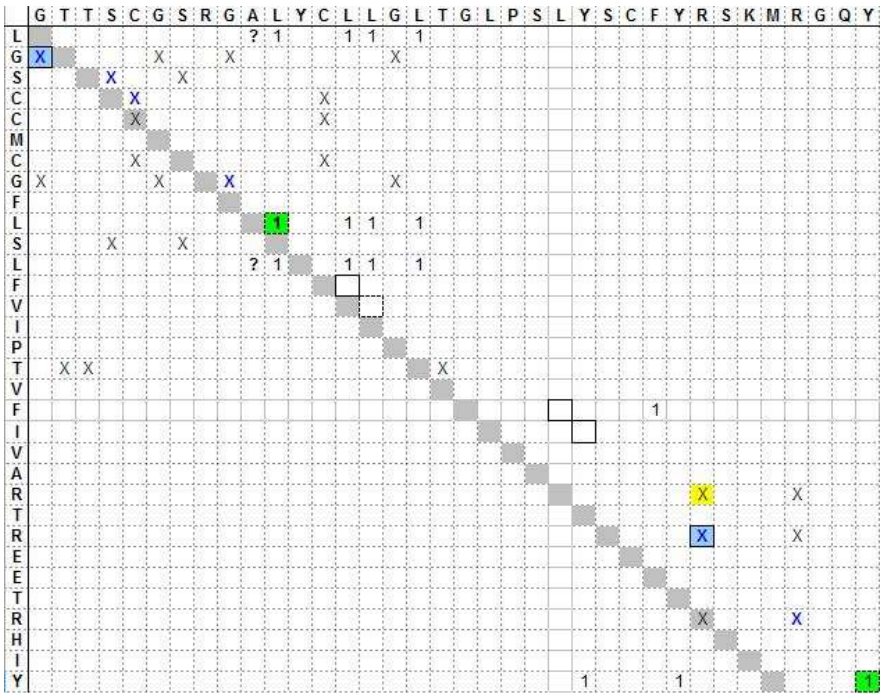


Fig. 19 Alineamiento matricial de secuencias

Si se toma la diagonal central de la matriz como el punto donde cada secuencia se corresponde con la otra en la misma posición, es decir una alineación 0-0; se observa que hay corrimientos entre dichas coincidencias lo cual se traduce en una matriz con varias diagonales con aciertos.
Para este caso, si la diagonal principal es la "0" las coincidencias correspondieron a las diagonales -1, +1 y +4.

Éste sistema de alineamiento, deducido previamente al conocimiento de la existencia de un método muy similar conocido como DOT PLOT, es una manera útil de localizar rápidamente las estructuras comunes en dos secuencias, incluyendo inserciones, deleciones, etc. Y fue útil para inferir las coincidencias que conservaran la misma distancia entre ellas, en una misma diagonal, pero no permitió identificar la 'regla general' que determina los saltos en la diagonal.

El desarrollo del algoritmo de comparación apropia, integra y asume parte de las reglas de ***nomenclatura de clusters***[10] para la graficación HCA y ajusta aquellas que no coinciden con la estrategia de comparación basada en HCA usada por los expertos.

Este método que actualmente se trabaja por los científicos de forma manual, se basa en la experticia del sistema ojo-cerebro del científico o persona altamente entrenada para encontrar patrones repetitivos y definir agrupaciones que puedan representar similitudes entre las secuencias y que por tanto revelen nuevos horizontes en cuanto al descubrimiento de nuevas relaciones estructurales y evolutivas entre proteínas que anteriormente se consideraban lejanas.

Es por esto, que para el diseño del algoritmo, se planteó como requerimiento importante el utilizar y preservar las reglas estipuladas para el cálculo del porcentaje de Identidad entre secuencias de proteínas que están descritas en los documentos sobre el método HCA[6,]

[10] *Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives* - L.Callebaut, G Labesse, P. Durand , A. Poupon , L. Canard , J. Chomilier , B. Henrissat and J. P. Mornon * 1997

2.3.2.2 Descripción del algoritmo

El algoritmo de comparación se basa en un modelo lineal ya que computacionalmente, un análisis lineal de las se**c**uencias proteicas representa beneficios en cuanto a tiempo y velocidad de procesamiento.

Para entender la lógica del algoritmo; se usará como terminología propia dos tipos de alineamientos:

- ***Alineamiento total:*** Referente al emparejamiento final que resulte de la aplicación del algoritmo de comparación de las secuencias.

- ***Alineamientos parciales o simplemente subalineamientos***: Estos alineamientos que son parte del alineamiento total, permiten escoger el mejor subalineamiento del amplio conjunto de posibles emparejamientos entre las subsecuencias de las proteinas.

2.3.2.3 Criterios de comparación

Los criterios utilizados para determinar la escogencia del mejor alineamiento global son los siguientes:

Criterios Generales:

Nacen de la observación del método HCA. Son inherentes al método.

- Todo cluster de aminoácidos hidrofóbicos seleccionado debe estar respaldado por al menos un aminoácido hidrofílico en ambas proteínas.

- Pueden existir saltos en el alineamiento. Los saltos o GAPS están determinados por la variación en la distancia relativa entre clusters seleccionados en la proteína conocida respecto a la desconocida.

- Los clusters seleccionados en ambas proteínas, deben ser idénticos respecto a su conformación. Es decir deben conservar sus distancias relativas internas iguales en ambas subsecuencias.

- El mejor alineamiento será aquel que proporcione mayor valor al cálculo del porcentaje de identidad entre las proteínas[11]

Criterios Particulares:

Los define el programador, según la forma del algoritmo.

- Al iniciar la comparación, el experto suministra un aminoácido de referencia para cada proteína; para dicha escogencia se le presenta la gráfica HCA. Estos aminoácidos se asumen como los primeros hallazgos en la búsqueda, partiendo de la presunción que el algoritmo toma como sugerencia del científico dichos puntos de la comparación.

- Para el cálculo de la separación entre clusters, no se tuvo en cuenta la presencia del aminoácido Prolina. La razón se justifica mas adelante en la sección: *Análisis previo para la definición de los criterios de comparación.*

- Se realizan ***subalineamientos*** tomando como base secciones de la proteína.

[11] Porcentaje de identidad, sección 1.3.3

- Al calcular el mejor subalineamiento entre dos subcadenas se tienen en cuenta las posiciones extremas de los aminoácidos hidrofóbicos dentro de dichas subcadenas. De esta forma al iniciar la comparación se hace el cálculo de la posición de partida y llegada para la búsqueda.

- Estas secciones son analizadas de manera lineal buscando encontrar entre todos los posibles alineamientos, el conjunto de aminoácidos que representa el mayor número de coincidencias en aminos hidrofóbicos e hidrofílicos idénticos, para cada subsecuencia.
- Esta determinación se basa en encontrar aquél subalineamiento que aporta un mayor puntaje o *score* para el hallazgo del **porcentaje final de identidad** entre las secuencias.

El algoritmo inicia la búsqueda dentro del cluster de partida en cada proteína y alinea el cluster buscando las relaciones entre aminoácidos hidrofóbicos e hidrofílicos circundantes. A esta búsqueda se le llama: ***Búsqueda Inicial.*** Cuando el programa encuentra un punto de separación entre clusters, inicia la fase de ***Búsqueda por subalineamientos.*** Al final del procedimiento se obtiene el ***Alineamiento Total*** de las proteínas.

Búsqueda Inicial:

La búsqueda inicial pretende encontrar similitudes en el radio cercano a los aminoácidos de partida. Para esto es importante localizar y resaltar en las secuencias los aminoácidos hidrofóbicos idénticos importantes para el cálculo de la similitud, además los aminoácidos hidrofóbicos no idénticos, que no cuentan para el cálculo de la similitud, pero permiten caracterizar la conformación de clusters los cuales son importantes para la conformación de la estructura secundaria de la proteína.

En el proceso de búsqueda también se resaltan aquellos aminoácidos hidrofílicos exactamente iguales que en HCA corresponden a los círculos y que son de gran importancia al igual que los primeros para el cálculo del porcentaje de identidad entre las proteínas.

Se determinan los ***Patrones de distancia*** es decir, el calculo de la distancia relativa al aminoácido de referencia en la proteína 1 o proteína conocida.

Análisis base para la definición de los criterios de comparación:

En la primera fase, se encontraron ciertos patrones de distancia entre los aminoácidos importantes para el cálculo de la identidad (fig. 20), sin embargo, repetidamente una vez que se hallaba un patrón para la comparación, se presentaba un salto o GAP que no cumplía con el patrón de búsqueda inicial (fig. 21).

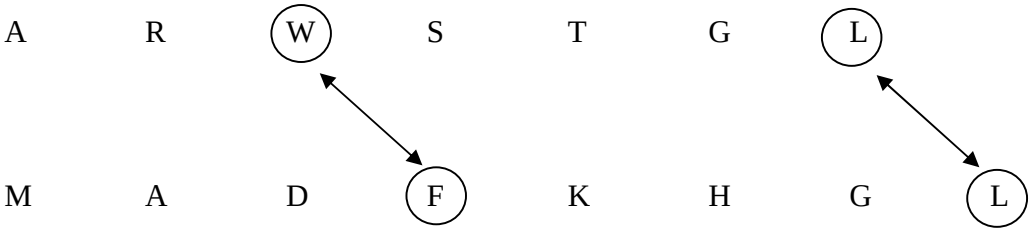


Fig. 20 Patrón de distancia constante entre aminoácidos Hidrofóbicos coincidentes

Patrón de Distancia: +1

En la figura 20 el patrón de distancia para la ocurrencia de aminoácidos Hidrofóbicos en ambas secuencias se mantiene constante. Se dice que el patrón es +1 porque hay un corrimiento de un aminoácido en la proteína dos, respecto a su homólogo en la proteína 1.

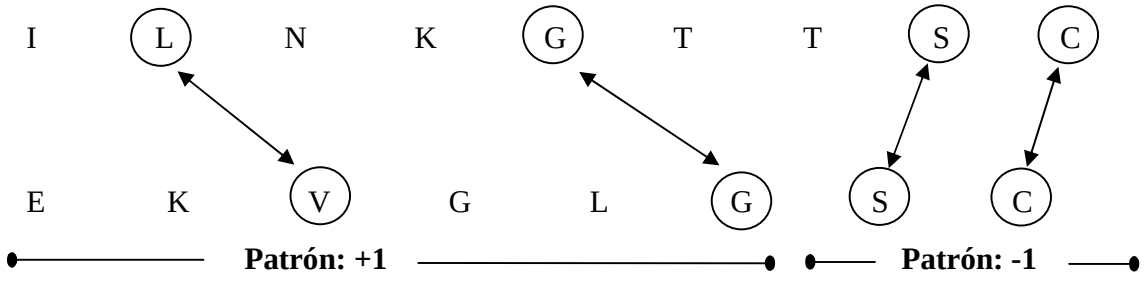


Fig. 21 Presencia de saltos en el alineamiento en los hallazgos de aminoácidos coincidentes e idénticos

En este caso se observa que hay una variación en el patrón de distancia relativa. En la figura 21 se observa que la variación se da entre aminoácidos hidrofílicos (G, S). El patrón de distancia cambia de +1 a -1. De esta forma, se reconoce un salto o ***gap*** y esto conlleva a que el algoritmo deba ajustarse al nuevo patrón.

Es importante resaltar que no existe alguna medida que indique de cuánto será la variación en el patrón de distancia entre un grupo de aminoácidos con igual distancia relativa y el siguiente. Tampoco se conoce de ante mano en qué momento(s) se dará dicho cambio.

En busca de la solución apropiada, se hizo una adición al método de comparación; tratando de utilizar una mayor información que permitiera inferir o determinar de alguna forma la lógica de dichos saltos y lograr por ende, hacerlos predecibles mediante el algoritmo. Se probaron las siguientes estrategias:

- Identificación de los clusters presentes (o agrupaciones de aminoácidos hidrofóbicos) en la cadena, por medio de la regla del código P. Según la cual, la separación ente clusters está dada por:

- El número de aminoácidos hidrofílicos que separan los clusters debe ser mayor a 4
- El rompimiento de las cadenas causado por el aminoácido Prolina: **P** (o estrella en la representación HCA)

2.3.2.4 Análisis de resultados

Luego de dicha prueba con estos ingredientes se concluyó:

El código P aporta información, sin embargo ésta no fue determinante para la solución del anterior problema de comparación, ya que como se especificará mas adelante, se tuvo que reinterpretar la forma como se estaba calculando cada uno de los clusters.

El código P de una proteína está implícito en el método de graficación de HCA. Cuando se encuentra con una prolina, la gráfica HCA indica inmediatamente la separación de dos clusters[12]. Sin embargo, dicha incidencia de la prolina no es tenida en cuenta por los expertos al momento de la comparación visual. Por ejemplo:

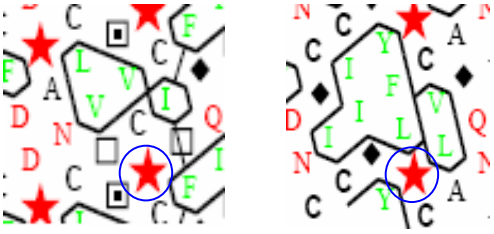


Fig. 23 Rompimiento por Prolina

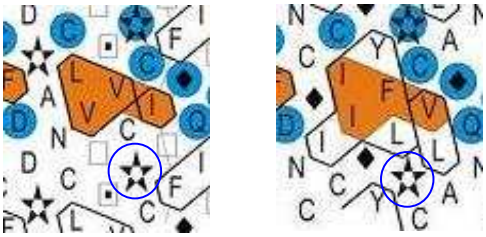


Fig. 22 Zona de similitud

La *figura 23* nos muestra partes de dos gráficas HCA sin comparación entre ellas, en las cuales se observa la incidencia de la prolina (la estrella inferior) en la ruptura de cada cluster.

[12] *Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives* - L.Callebaut, G Labesse, P. Durand , A. Poupon , L. Canard , J. Chomilier , B. Henrissat and J. P. Mornon * 1997

La fig*ura 22* por otro lado, es tomada del análisis de ***comparación HCA*** realizado por un biólogo[13] sobre dichas secuencias; y nos presenta que aún habiendo rompimiento, se pueden tomar zonas que abarquen dichos clusters como una subunidad.

Por lo anterior, en el diseño y desarrollo del algoritmo basado en HCA, se usó el criterio de comparación HCA y no el de graficación, por tanto, se obtiene que la excepción de rompimiento para la prolina no es aplicable en la comparación.

Búsqueda por subalineamientos:

Dentro de la búsqueda por subalineamientos se ha definido como el mejor alineamiento aquél que:

- Tenga ocurrencia de aminos hidrofílicos y al menos un amino hidrofóbico idéntico[14] o coincidente[15].

- El punto de ruptura de una subcadena esté determinado por las siguientes condiciones:

    - Al menos 4 aminoácidos hidrofílicos después de un hidrofóbico
    - Cuando alguna de las dos secuencias llega a su final.

Se muestran a continuación algunos ejemplos de los resultados obtenidos. Se tiene en cuenta que el punto de referencia para el análisis de los resultados es tomado con base en alineamientos hechos por estudiantes y docentes de la Universidad, quienes nos facilitaron algunos ejemplos de comparación manual.

[13] Alfonso Pineda , 10º semestre Biología UIS.
[14] Aminoácidos Idénticos: Aquellos que coincide en Denominación y posición relativa dentro de un cluster, en ambas secuencias
[15] Aminoácidos Coincidentes: Aquellos que siendo distintos, coinciden en posición relativa dentro de un cluster en ambas secuencias.

Se calculó el porcentaje de error como:

$$\%Error = (Valor\ Teórico - Valor\ Experimental / Valor\ Teórico) * 100$$

Ec. 2 Cálculo del porcentaje de error

Valor Teórico: Es el valor base del porcentaje obtenido en comparaciones verificadas por expertos según el método tradicional.

Valor Experimental: Es el porcentaje de Identidad generado para la misma comparación por el Algoritmo de Comparación Basada en HCA, y observable en el archivo PostScript que genera el programa.

Ejemplo Alineamiento 1:
(Fig. 24)

ORFX L. Esculentum

ARWSTGLCHCFDDPANCLVTSVCPCITFGQISEILNKGTTSCGSRGALYCLLGLTGLPSLYSCFYRSKMRGQY
DLEEAPCVDCLVHVFCEPCALCQEYRELKNRGFDMGIGWQANM

DUF614 protein [Branchiostoma belcheri tsingtaunese]

MADFKHGLLGCFDNCGICIIGYFLPCVLAGQNAEKVGLGSCCMCGFLSLFVIPTVFIVARTREETRHIYSIEG
TFLNGCLLTFFCPFCVMVQTAQELDEGVGAQIIIRQ

% Identidad: 26.6%
% Teórico : 27%

% Error: 1.48 %

Ejemplo Alineamiento 2

(fig. 25)

4Q21 c-Ha-ras1 p21 protein [Homo Sapiens]

MTEYKLVVVGAGGVGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVIDGETCLLDILDTAGQEEYSAMRDQ
YMRTGEGFLCVFAINNTKSFEDIHQYREQIKRVKDSDDVPMVLVGNKCDLAARTVESRQAQDLARSYGIP
YIETSAKTRQGVEDAFYTLVREIRQHKLRKLNPPDESGPGCMSCKCVLS

1WKQ Chain B, Crystal Structure Of Bacillus Subtilis Guanine Deaminase

MHHHHHHAMNHETFLKRAVTLACEGVNAGIGGPFGAVIVKDGAIIAEGQNNVTTSNDPTAHAEVTAIRKA
CKVLGAYQLDDCILYTSCEPCPMCLGAIYWARPKAVFYAAEHTDAAEAGFDDSFIYKEIDKPAEERTIPF
YQVTLTEHLSPFQAWRNFANKKEY

% Identidad: 14.02%

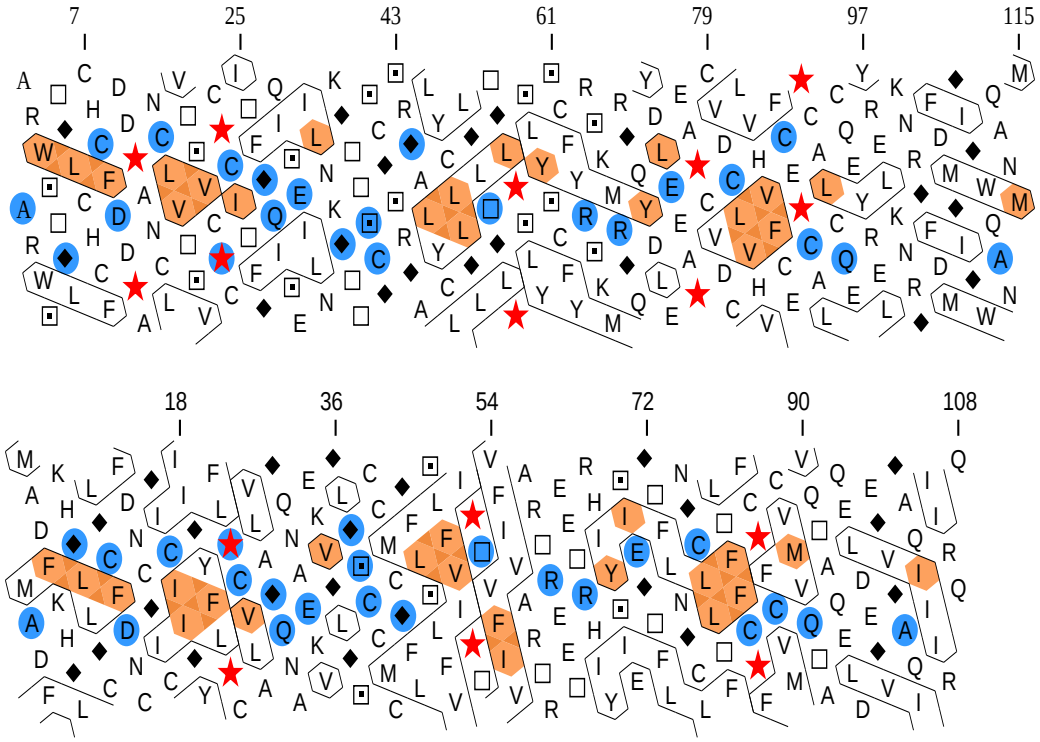
% Teorico: 15%

% Error: 6.5 %

Tabla 2 Resumen de resultados método Algoritmo de Comparación Basado en el Análisis Hidrofóbico de Clusters

Proteína 1	**Proteína 2**	**% Teórico (método visual)**	**% Identidad Algoritmo Comparación Basado en HCA**	**% Error**
ORFX L. Esculentum	DUF614 protein [Branchiostoma belcheri tsingtaunese]	27%	26.6%	1.48 %
4Q21 c-Ha-ras1 p21 protein [Homo Sapiens]	1WKQ Chain B, Crystal Structure Of Bacillus Subtilis Guanine Deaminase	15%	14.02%	6.5 %

Alineamiento 1



Porcentaje de Identidad: 26.6 %

Fig. 24 Comparación ORFX vs DUF614

Alineamiento 2:

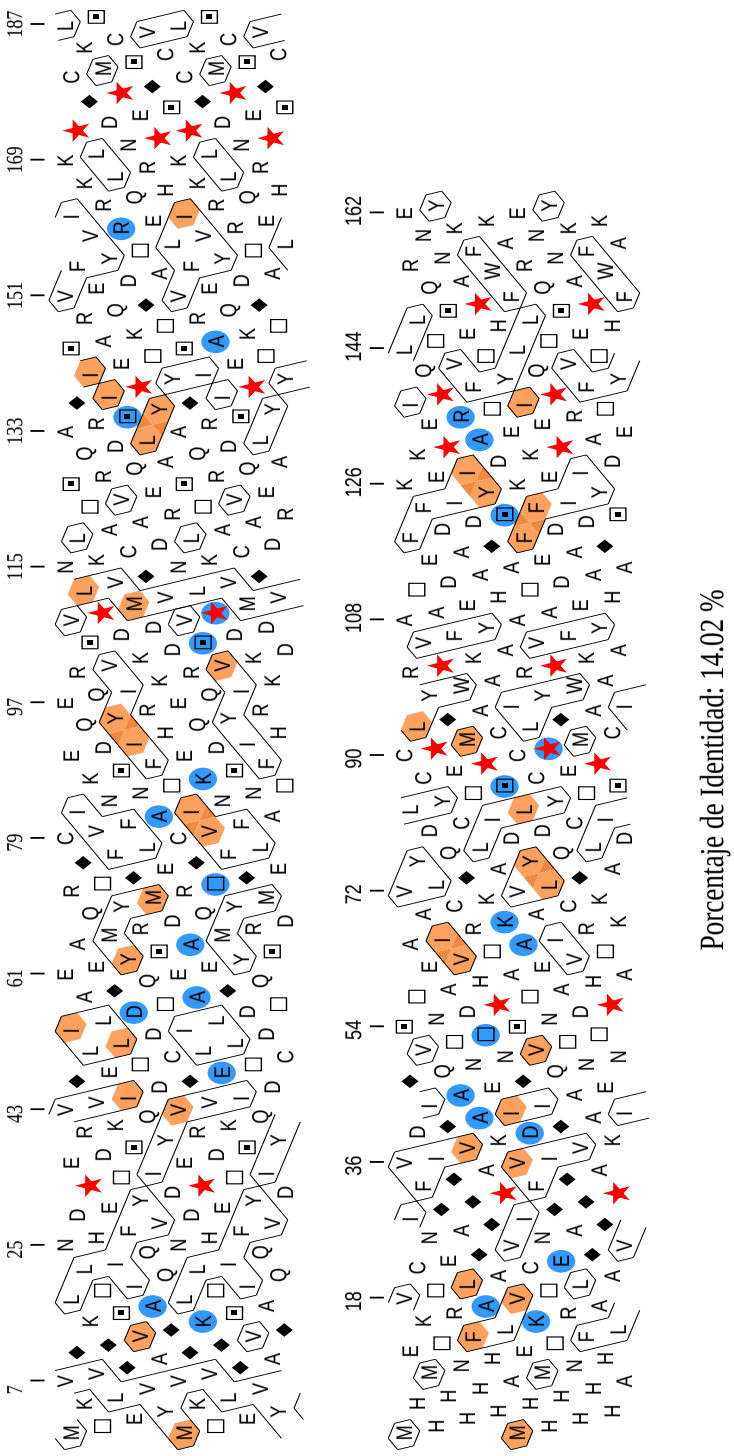


Fig. 25 Comparación 4Q21 Vs 1WKQ

Capítulo 3

Conclusiones y Recomendaciones

3.1 Conclusiones

- Se construyó una herramienta computacional que permite la comparación de secuencias de proteínas utilizando los conceptos del análisis de clusters hidrofóbicos (HCA), para apoyar la labor de los expertos en dicho procedimiento.

- Se establecieron los criterios correspondientes de comparación de secuencias de proteínas, utilizando los conceptos empleados en el Análisis de Clusters Hidrofóbicos HCA.

- Se encontró que para los alineamientos usando el Algoritmo de Comparación Basado en HCA (sección 2.3.2), el grado de error es variable, según la complejidad de las proteínas a comparar y el aminoácido de partida que se indique al inicio de la comparación.

- Mediante la utilización de la herramienta computacional, se logra la comparación de secuencias de proteínas con buenos resultados para las secuencias prueba.

3.2 Recomendaciones

- Ampliar la base de datos de proteínas ingresando nuevos casos de estudio de manera que se facilite al científico el acceso a un amplio repositorio de secuencias.

- Considerar otros criterios que puedan mejorar el método de comparación cuando se presentan ciertos casos de clusters adyacentes.

- Si bien se alcanzaron resultados satisfactorios, la continuidad de la investigación puede complementarse con el estudio con otras técnicas como el tratamiento digital de imágenes, considerando que el proceso que se realiza en la comparación es netamente visual.

Bibliografía

[Alberts, 1996] Alberts. Molecular biology of the cell. Barcelona: Ediciones Omega, 3ª ed., 1996.

[Altschul, et al. 1991], Altschul, S.F. (1991) “Amino acid substitution matrices from an information theoretic perspective". J.Mol.Biol, (219), 555-565

[Baldi, 2001] Pierre Baldi and Soren Brunak, (2001). Bioinformatics, the machine learning approach. Second edition. MIT.

[Battaner 2001] Modelos Moleculares, Departamento de Bioquímica y Biología Molecular. U Salamanca. http://www.usal.es/~dbbm//modmol/index.html .Consultado Enero 2008

[Baxevanis, et al, 2001], Andreas D. Baxevanis, B. F. Francis Ouellette. (2001), Bioinformatics, A practical Guide to the Analysis of Genes and Proteins. Second edition. Wiley-interscience.

[Bernstein, et al., 1977], Bernstein FC, Koetzle TF, Williams GJB, Meyer EF, Brice MD, Rodgers JR, Kennard O, Shimanovichi T, Tasumi M. (1997). The protein data bank: a computer-based archival file for macromolecular structures. J. Mol. Biol. 112:535-542.

[Bork et al., 1994], Bork P, Ouzounis C, Sander C. (1994). From genome sequences to protein function. Curr. Opin. Struct. Biol. 4:393-403.

[Clote, 2000], Peter Clote, Rolf Backofen, (2000). Computational Molecular Biology an Introduction. Wiley.

[C. Gaboriaud et al, 1987] C. Gaboriaud, V. Bissery, T. Benchetrit and J.-P. Mornon, Hydrophobic cluster analysis; an efficient new way to compare and analyse amino acid sequences. Febs Letters, 224:149-155, 1987.

[Curtis, 2000] Biología. Ed. Panamericana, Buenos Aires. Sexta edición, 2000.

[Doolittle, 1981], Doolittle, R.F. (1981), "Similar amino acid sequences: chance or common ancestry?" Science, (214), 149-159

[DrawHCA] [http://bioserv.impmc.jussieu.fr/hca-form.html] consulta: junio 2008

[Eudes, 2007], Richard Eudes, Khanh Le Tuan, Jean Delettré, Jean-Paul Mornon, Isabelle Callebaut (2007). A generalized analysis of hydrophobic and loop cluster within globular protein sequences, BMC structural Biology.

[Garrido, 2002] Proteínas.
http://w3.cnice.mec.es/eos/MaterialesEducativos/mem2002/proteinas/tema/index.htm
Consultado Junio 2008.

[Guyton et al, 1996] Guyton, Arthur C y Hall, John E. Tratado de Fisiología Médica. Madrid: McGraw-Hill - Interamericana de España, 9ª ed., 1996.

[Henikoff et. al 1991], Henikoff, S. and Henikoff, J.G. (1991) "Automated assembly of protein blocks for database searching". NAR (19), 6565-6572

[Henikoff et.al, 1992], Henikoff, S. and Henikoff, J.G., (1992) "Amino acid substitutions matrices from protein blocks", Proc.Natl.Acad.Sci. USA, (89), 10915-10919

[Holm L. & Sander C., 1996], Holm, L. & Sander, C. (1996). The FSSP database: fold classification based on structure-structure alignment of proteins. Nucleic Acid Res. 24: 206-209.

[I. Callebaut et al, 1997] I. Callebaut, G. Labesse, P. Durand, A. Poupon, L. Canard, J. Chomilier, B. Henrissat and J. P. Mornon, Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives, 1997

[Jacobson et al., 2000], Jacobson, L., Booch, G., and Rumbaugh, J. (2000). El Proceso Unificado de Desarrollo de Software. Pearson Education. S:A:, Madrid.

[Joachim, 2007], Hans Joachim Bockenhauer, Dirk Bongartz, (2007), Algorithmic Aspects of Bioinformatics, Natural Computing series, Springer.

[Lehninger, 1993] Lehninger, Albert L. Principios de bioquímica. Barcelona: Ediciones Omega, 2ª ed., 1993.

[Lesk, 2002], Arthur M, Lesk (2002), Introduction to bioinformatics. Oxford.

[Levit & Chothia, 1976], Levitt, M. & Chothia. C. (1981). Structural patterns in globular proteins. Nature, 261:552-558.

[Lipman & Pearson, 1985], Lipman D.J. and Pearson W.R. (1985), "Rapid and sensitive protein similarity search", Science, (227), 1435-1444

[Madej et al., 1995] Madej, T., (1995). Threading analysis suggests that the obese gene product may be a helical cytokine. FEBS lett. 373:356-359.

[Maizel & Lenk, 1981], Maizel, J.V. and Lenk, R.P, (1981), "Enhanced graphic matrix analysis of nucleic acid and protein sequences", Proc.Natl.Acad.Sci.USA 78(12) 7665-7669

[Mas Benavente, 2000], José Manuel Mas Benavente. (2000). Comparación computacional de estructuras de proteínas, aplicación al estudio de un inhibidor de carboxipeptidasa como agente antitumoral. Tesis doctoral. Universidad Autónoma de Barcelona.

[Mas et al, 1998], Mas JM, Aloy P, Marti-Renom MA, Oliva B, Blanco-Aparicio C, Molina MA, de LLorens R, Querol E, Aviles FX (1998). Protein similarities beyond disulphide bridge topology. J Mol. Biol. 284:541-548.

[Murzin AG., et al. 1995], Murzin AG, Brenner SE, Hubbard T, Chothia C. (1995). SCOP a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 247:536-540.

[Netto, 2001] Fisicanet.
http://www.fisicanet.com.ar/biologia/introduccion_biologia/ap12_aminoacidos_y_proteinas.php. Consultado Junio de 2008

[Orengo C., et al. 1996], Orengo, C, A. and Taylor, W. R. (1996). SSAP: Sequential structure alignment program for protein structure comparison. In: Computer methods for macromolecular sequence analysis. Edited by R. F. Doolittle. Orlando, USA, Academic Press, 266:617-635.

[Pearson & Lipman, 1988], Pearson W.R. and Lipman D.J.; (1988) "Improved tools for biological sequence comparison", Proc.Natl,Acad.Sci. USA (85), 2444-2448

[Posfait et.al 1989], Posfai J., Bhagwat, A.S, Posfai G., and Roberts, R.J. (1989) "Predictive motifs derived from cytosine methyltransferases". NAR (17), 2421-2435

[Richarson, 1981], Richardson, J. S. (1981). The anatomy and taxonomy of protein structure. Advan. Prot. Chem. 34:167-339.

[Rodríguez et al, 2008] La Estructura de las Proteínas - Departamento de Bioquímica- Facultad de Química UNAM Mexico. http://depa.pquim.unam.mx/proteinas/estructura/ Consultado: Mayo 2008

[Sellers, 1974], Sellers, P.H. (1974) "On the theory and computation of evolutionary distances SIAM", J.Appl.Maths, (26), 787- 793

[Starr, 2004] Starr- Taggart Biología, La Unidad y Diversidad de la Vida, Thompson 10º Ed. 2004

[Stryer, 1996] Stryer, L. Bioquímica. 2 vols. Barcelona. Editorial Reverté, 4ª ed., 1996.

[Tejedor 2008] Proteinas Globulares Material Bioquímica– U Alcala, España 2007

[Trelles, 2000], Oswaldo Trelles, (2000). Comparación de secuencias biológicas, algoritmia. Doctorado en bioinformática. UAM.

Anexos

ANEXO A

Listado Aminoácidos Esenciales

NOMBRE	ABRV	SÍMBOLO
Alanina	Ala	A
Arginina	Arg	R
Asparagina	Asn	N
Aspártico	Asp	D
Cisteina	Cys	C
Fenilalanina	Phe	F
Glicina	Gly	G
Glutámico	Glu	E
Glutamina	Gln	Q
Histidina	His	H
Isoleucina	Ile	I
Leucina	Leu	L
Lisina	Lys	K
Metionina	Met	M
Prolina	Pro	P
Tirosina	Tyr	Y
Treonina	Thr	T
Triptófano	Trp	W
Serina	Ser	S
Valina	Val	V