APLICACIÓN DE MINERÍA DE DATOS PARA LA AGILIZACIÓN DEL PROCESO DE ANOTACIÓN DE PROTEÍNAS MEDIANTE LA SELECCIÓN DE SECUENCIAS EN BASES DE DATOS DENSAMENTE POBLADAS

DIANA MARCELA GRANADOS JIMÉNEZ
PEDRO ARTURO RINCÓN TELLO

ESCUELA DE INGENIERÍA DE SISTEMAS E INFORMÁTICA

FACULTAD DE INGENIERIAS FISICOMECANICAS

UNIVERSIDAD INDUSTRIAL DE SANTANDER

BUCARAMANGA

2013

APLICACIÓN DE MINERÍA DE DATOS PARA LA AGILIZACIÓN DEL PROCESO DE ANOTACIÓN DE PROTEÍNAS MEDIANTE LA SELECCIÓN DE SECUENCIAS EN BASES DE DATOS DENSAMENTE POBLADAS

DIANA MARCELA GRANADOS JIMÉNEZ
PEDRO ARTURO RINCÓN TELLO

Trabajo de Grado para optar al título de
Ingeniería de Sistemas

Director
MsC. Lola Xiomara Bautista

Codirector(es)

Ph.D. Cristian Blanco Tirado
MsC.Darío José Delgado Quintero

ESCUELA DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
FACULTAD DE INGENIERIAS FISICOMECANICAS
UNIVERSIDAD INDUSTRIAL DE SANTANDER
BUCARAMANGA
2013

DEDICATORIA

Este libro está dedicado a Dios por ser la luz de mi camino y permitirme alcanzar esta meta propuesta.

A mis padres Luis y Ana por sus sabios concejos, por su gran apoyo en momentos de dificultad y por creer en mí.

A mis hermanos Alejandra y Alexander por darme en cada momento una chispa de alegría.

A mi familia especialmente a mis tíos Abel y Mercedes por inspirarme a seguir adelante acompañado de sus oraciones.

A mi compañero Arturo del cual aprendí mucho.

A Darío José delgado Quintero y cindy Dayana Solano por guiarme en este proceso y compartir concejos los cuales me han permitido crecer como persona y profesional.

A mis amigos Mónica, Viviana, Roció, Mateo, Carlos, Cristian y Jorge porque a pesar de estar a grandes distancias siempre me motivaron para hacer de este proyecto una realidad.

Diana

DEDICATORIA

Dedico este trabajo, fruto de un gran esfuerzo, a Alejandra, por brindarme siempre su afecto, su compañía y los mejores consejos. Sus palabras de aliento hicieron esto posible.

A mis padres Stella y Jaime, por su paciencia y apoyo constante.

A Diana por su perseverancia.

A Darío y Cindy, por ofrecerme su conocimiento, experiencia y por ser los mejores guías.

Al profesor Christian, por inspirarme.

Arturo

AGRADECIMIENTOS

Los Autores del proyecto queremos agradecer

A Dios por permitirnos realizar este proyecto.

A la profesora Lola Xiomara Bautista, Directora de nuestro proyecto por sus concejos, aportes al proyecto y por creer en nosotros.

Al Profesor Cristian Blanco tirado Codirector de nuestro proyecto por confiar en nosotros y dedicarnos parte de su tiempo además de la idea de desarrollar este proyecto.

Al Ingeniero Darío José Delgado Quintero Codirector de nuestro proyecto por compartir sus conocimientos y por orientarnos en la realización del mismo.

A la Ingeniera Cindy Dayana Solano por su dedicación y sus aportes al proyecto.

A nuestra familia por apoyarnos y creer en nosotros.

A nuestros amigos y compañeros del GIIB por acompañarnos en este proceso

Gracias a la Universidad Industrial de Santander por darnos las bases para hacer de nosotros unos profesionales.

Al centro nacional de Investigaciones del café por los seguimientos en el trascurso del proyecto.

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

Pág.

ÍNDICE DE TABLAS

ÍNDICE DE ANEXOS

RESUMEN

TITULO: APLICACIÓN DE MINERÍA DE DATOS PARA LA AGILIZACIÓN DEL PROCESO DE ANOTACIÓN DE PROTEÍNAS MEDIANTE LA SELECCIÓN DESECUENCIAS EN BASES DE DATOS DENSAMENTE POBLADAS.[1]

AUTOR(ES): Pedro Arturo Rincón Tello **-** Diana Marcela Granados Jiménez[2]

PALABRAS CLAVE: Análisis de Clusters Hidrofobicos, Bases de Datos Densamente Pobladas, Minería de Datos.

DESCRIPCIÓN:

En este documento se indica las fases a seguir para dar lugar al desarrollo y validación de un algoritmo computacional, que agilizará el proceso de comparación de secuencias de proteínas del café codificadas a través de la metodología **HCA** (Hydrophobic Cluster Analysis), la cual ofrece buenos resultados, pero es incapaz de procesar automáticamente grandes cantidades de secuencias ni trabajar en conjunto con una base de datos.

Este proyecto se basa en la aplicación de la técnica de Minería de Datos **SOM** (Self Organizing Maps), que junto con el algoritmo **VCM** (Vector Composición de Momento), permite extraer, recodificar, clasificar y agrupar de acuerdo a su contenido estructural un conjunto de secuencias de proteínas almacenadas en la base de datos conocida como **PDB** (Protein Data Bank), logrando una significativa reducción de los tiempos de búsqueda y le selección de secuencias con un alto porcentaje de similaridad.

La realización del presente proyecto ofrece como resultado, una base de datos reestructurada y una función de búsqueda, que en conjunto permiten la agilización del proceso de anotación de proteínas y propone la posibilidad de encontrar datos que en algún momento fueron ignorados y que puedan aportar información valiosa sobre la similaridad entre secuencias, y por lo tanto para el proceso de anotación.

---

[1] Trabajo de grado

[2] Facultad de Ingenierías Físico-Mecánicas Escuela de Ingeniería de Sistemas e Informática Director PhD(C).Lola Xiomara Bautista Codirector(es): Cristian Blanco Tirado, Darío José Delgado Quintero

ABSTRACT

TITLE: DATA MINING APPLICATION TO SPEED UP THE PROTEIN ANNOTATION PROCESS THROUGH SELECTION OF SEQUENCES IN DENSELY POPULATED DATABASES[3]

AUTHORS: Pedro Arturo Rincón Tello – Diana Marcela Granados Jiménez[4]

KEYWORDS: Hydrophobic cluster analysis, densely populated databases, Data Mining.

DESCRIPTION:

This paper indicates the steps to result in the development and validation of a computational algorithm, which will speed up the comparison process of coffee protein sequences encoded through the **HCA** (Hydrophobic Cluster Analysis) methodology, which already offers good results, but needs human expert interaction and therefore it's results may be ambiguous and it's also unable to automatically process large amounts of sequences nor to work together with a database.

This project is based on the application of the **SOM** (Self-Organizing Maps) a Clustering Data Mining technique which along with the **VCM** (Composition Moment Vector), allows to extract, recode, classify and group, according to its structural –and therefore functional- content, a set of protein sequences stored in the international data base known as **PDB** (Protein Data Bank), thus achieving a significant search time reduction and the selection of sequences with a high similarity percentage.

The realization of this project provides as a result, a restructured database and a search function, which together allow the speeding up of the protein annotation process, the reduction o ambiguity and offering the possibility of finding data that were previously ignored and that can provide valuable information about the similarity between sequences and therefore for the annotation process.

[3] Degree Project

[4] Physical-Mechanical Engineering Faculty Systems and Informatics Engineering School
Director: PhD(C). Lola Xiomara Bautista Co-Director(s): Cristian Blanco Tirado, Darío José Delgado Quintero

INTRODUCCIÓN

La aplicación de técnicas computacionales a la gestión y posterior análisis de datos biológicos fue el inicio de lo que conocemos hoy en día como bioinformática y biología computacional, disciplinas que continuamente requieren e incentivan la creación de nuevas metodologías que faciliten tanto el estudio de la información biológica como la mejora en el análisis de sus resultados. **“**Los grandes volúmenes de datos**”** se suelen citar como una de las características más relevantes de la bioinformática debido a sus tasas exponenciales de crecimiento [Abascal, 2003]. Este crecimiento se debe en gran medida a los estudios que han permitido pasar de **“**algunos pocos genes a genomas completos**”** [GenBank], lo cual se traduce en un aumento exponencial de información.

Actualmente, la proteómica ha adquirido significativa importancia y gran expectativa en cuanto a los beneficios que sus descubrimientos puedan traer en el futuro. Lo cual está ligado a la información disponible y a la capacidad de procesamiento de información que tienen los equipos de cómputo actuales.

Procesar información de origen biológico no es tarea fácil más aun cuando de por medio se trabaja con bases de datos en las cuales se manejan grandes volúmenes de información correspondientes a secuencias de ADN o proteínas. Un caso particular a referenciar es el Protein data bank [PDB] una base de datos que alberga cerca de 79265estructurascorrespondientes a proteínas, caracterizada por pasar de tan sólo 7 estructuras inicialmente con año de registro 1976 a87681estructurasrelativas a proteínas, ADN, ARN y otras como capsides virales; experimentando un crecimiento exponencial [PDB]. Esta información permite realizar estudios comparativos entre secuencias de proteínas en busca de

información con relación a su homología y su funcionalidad, y en donde el análisis de dicha información de forma sistemática puede tomar mucho tiempo y esfuerzo.

Una de las herramientas utilizadas por los biólogos para el análisis de información biológica es la comparación de secuencias. La cual busca encontrar zonas de similitud significativa en las secuencias, ubicando de esta manera características de interés común o diferencial entre varias cadenas ya sean de proteínas o ADN. Algunas de las herramientas más conocidas para esta tarea son BLAST [TATUSOVA & MADDEN, 2011] y FASTA [PEARSON & LIPMAN, 1988], entre otras, las cuales se encargan de comparar secuencias de proteínas entre sí en busca de similaridad. Si bien, estas herramientas son en general eficientes para procesar los datos, algunas presentan dificultades en su desempeño debido a falencias propias de las técnicas. Una alternativa a las herramientas tradicionales de comparación de secuencias es el Análisis de Clústeres Hidrofobicos (HCA) [GABORIAUD C, BISSERY V, BENCHETRIT T & MORNON JP, 1987], no obstante, debido a su falta de automatización en su utilización, no existen marcos de navegación sobre bases de datos que permitan analizar de forma rápida, los grandes volúmenes de información que se encuentran almacenados en estas.

En este trabajo se propone reorganizar la información presente en el PDB teniendo en cuenta los datos que se pueden extraer mediante el uso de la técnica HCA y de la utilización de una técnica de organización de datos denominada redes SOM (Mapas Autoorganizativos) [KOHONEN, 1995] a través de la cual se desarrolló un esquema de reorganización de secuencias y consulta de las mismas basadas en HCA para la base de datos PDB.

Este documento se estructura en tres capítulos los cuales se relacionan a continuación:

El primer capítulo hace referencia al marco teórico de la temática, en él se indican los conceptos y métodos empleados para el desarrollo del algoritmo de

reorganización de secuencias, mediante la aplicación y el análisis de la técnica HCA para el manejo de la información presente en el PDB.

El segundo capítulo se estructura y explica en detalle la construcción de los algoritmos de reorganización de la información, así como también se valida mediante la utilización de una técnica estándar de comparación de la mejora en tiempos de consulta sobre el PDB con la utilización de el algoritmo desarrollado.

Finalmente se presentan las conclusiones y recomendaciones producto del desarrollo de este proyecto de investigación.

MARCO TEÓRICO

1.1 LAS PROTEÍNAS

Las proteínas son importantes polímeros biológicos formados por la construcción de bloques llamados aminoácidos [ALAIN, 2002]. Poseen amplia variabilidad estructural y funciones biológicas muy diversas. Para su clasificación se puede recurrir a criterios físicos, químicos, estructurales o funcionales. Toda proteína, desde las humanas hasta las que forman las bacterias unicelulares, son el resultado de las distintas combinaciones entre 20 aminoácidos esenciales (ver ANEXO A).La representación lineal de una proteína es denominada secuencia, en la que los aminoácidos se unen a lo largo de una sola línea (con un inicio denotado como N-termino y el final C-termino).

1.1.1 AMINOÁCIDOS

Unidades estructurales fundamentales, como indica su nombre, tienen dos grupos funcionales característicos: el grupo carboxilo o grupo ácido (-COOH), y el grupo amino (-NH2).

1.1.1.1 CLASIFICACIÓN DE LOS AMINOÁCIDOS

Según su grado de hidrofobicidad se clasifican en [ALAIN, 2002]:

- Grupo I: Aminoácidos apolares (hidrofobicos).Si están en gran abundancia en una proteína, la hacen insoluble en agua. Entre ellos: Prolina, Alanina, Valina, Leucina, Isoleucina, Metionina, Tirosina y Fenilanina.

- Grupo II: Aminoácidos polares no ionizables (hidrofilicos). Contrariamente al grupo anterior, si una proteína los tiene en abundancia será soluble en agua entre ellos:Serina, Triptofano, Treonina, Metionina, Lisina, Histidina, Glutamina, Glutamico, Glicina, Cisteína, Aspartico, Asparagina, Arginina.

1.1.2 ESTRUCTURAS DE LAS PROTEÍNAS

La estructura de una proteína es la disposición espacial que adopta una molécula proteica, es responsable de la función de la proteína y depende tanto de los aminoácidos presentes en ella como de las interacciones químicas entre estos, dichas interacciones pueden provocar en la proteína plegamientos específicos. Se clasifica jerárquicamente en cuatro niveles interdependientes:

1.1.2.1 ESTRUCTURA PRIMARIA

Se refiere a la secuencia lineal de aminoácidos que componen la proteína. Permite determinar sus características químicas y biológicas, y especifica los niveles estructurales superiores. Su alteración puede modificar la configuración general de la proteína, dando lugar a una diferente. Afectando también su función, determinando que esta pueda o no realizarla. [GARRIDO, 2002]

Figura 1. Estructura primaria de una proteína

Fuente:http://www.smallscalechemistry.colostate.edu/PowerfulPictures/ProteinStructure.pdf

1.1.2.2 ESTRUCTURA SECUNDARIA

El segundo nivel estructural hace referencia a su disposición espacial [GUILLEN, 2009]. Puede variar entre distintos motivos, siendo los principales:

- Estructura en Hélice Alfa:

Se trata de la forma más simple y común, donde la proteína adopta una forma helicoidal y posee 3,6 aminoácidos por vuelta. [ALAIN, 2002]

Figura 2.Hélice alfa Estructura Secundaria

Fuente:http://www.smallscalechemistry.colostate.edu/PowerfulPictures/ProteinStructure.pdf

- Estructura en Lámina Beta:

Se origina cuando la molécula proteica, o una parte de esta, adopta una disposición en zig-zag. [GUILLEN, 2009]

Figura 3. Hoja Plegada Beta

Fuente:http://www.smallscalechemistry.colostate.edu/PowerfulPictures/ProteinStructure.pdf

1.1.2.3 ESTRUCTURA TERCIARIA

Cuando las fuerzas al interior de la cadena polipeptídica provocan que la molécula se vuelva todavía más compacta, sufriendo pliegues en el espacio y adoptando una forma tridimensional, se constituye una estructura terciaria. [ALAIN, 2002]

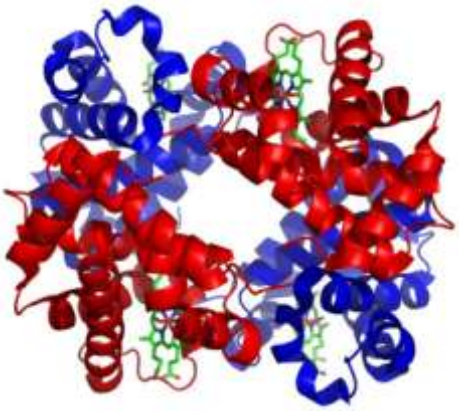
Figura 4. Estructura terciaria de las Proteínas

Fuente: http://lib.bioinfo.pl/courses/view/501

1.1.2.4 ESTRUCTURA CUATERNARIA

Se forma por la agregación de dos o más cadenas de polipeptidos, llamadas subunidades o monómeros, las cuales pueden tener idénticas o diferentes secuencias de aminoácidos. [ALAIN, 2002]

Figura 5. Estructura cuaternaria de una proteína



Fuente: http://lib.bioinfo.pl/courses/view/501

1.2 COMPARACIÓN DE SECUENCIAS

Comparar secuencias se presenta como un primer paso hacia el análisis bioinformatico a nivel estructural y funcional de nuevas secuencias descubiertas, ya que permite realizar inferencias sobre la evolución de una nueva proteína con base en proteínas existentes en bases de datos de información biológica. Se fundamenta en el proceso de alineamiento de secuencias [TORRES Miguel, PAEZ Cárdenas, MARTINEZ Alicia, RODRIGUEZ Enrique, 2012].El alineamiento de secuencias aminoacidicas o (ácidos nucleicos) tiene como objetivo encontrar la posición relativa de dos de ellas, en la que se produzca un mayor número de coincidencias entre sus componentes, valorando su similaridad. Las relaciones entre secuencias pueden ser: (homologas, ortólogas, xenologas o parálogas) se puede realizar de dos formas:

1.2.1 COMPARACIÓN POR IDENTIDADES

Utiliza una matriz en la que se coloca una secuencia en forma vertical, la otra de forma horizontal y se recorre cada una de las diagonales, acumulando el mayor número de coincidencias ver figura 6. La diagonal que presente el mayor número de identidades representará el desplazamiento relativo que mejor alinea las secuencias. [CARUGO, 2006].Su uso es muy limitado y no es la más eficiente para establecer la relación entre secuencias [DOOLITLE, 1981], ya que la sustitución de un aminoácido por otro de propiedades similares, puede no tener gran influencia en la función de la proteína, se debe considerar la inserción y pérdida de residuos.

Figura 6. Comparación por identidades

	0	1	2	3	4	5	6	7	8	9	10	11
0		T	C	A	G	A	C	G	A	T	T	G
-1	A			1		1			1			
-2	T	1								2	1	
-3	C		2				1					
-4	G				1			2				3
-5	G				3			1				1
-6	A			1		4			2			
-7	G				2			2				
-8	C		1				1					
-9	T	1								3	1	
-10	G				2			3				2

```
pos:      12345678901
X:        TCAGACGATTG   n=11
          | |   | |
pos:  Y:  ATCGGAGCTG    n=10
```

Fuente: Elaboración propia

1.2.2 COMPARACIÓN POR SEMEJANZA

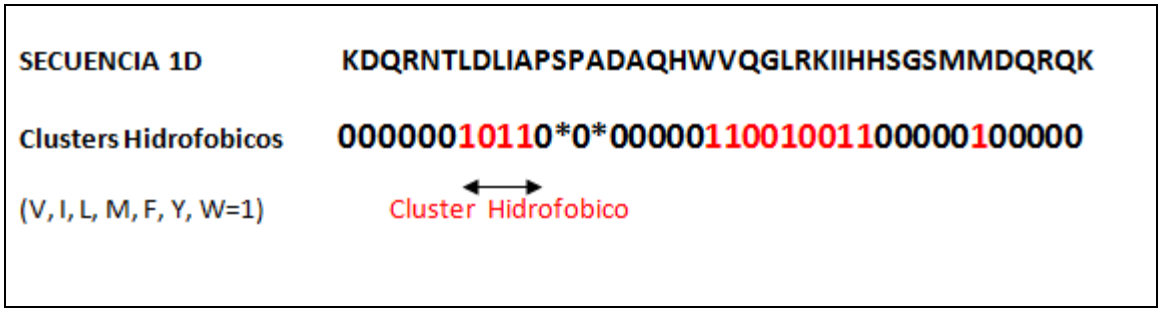
Para encontrar la similitud entre secuencias, es necesario alinear las regiones semejantes entre sí. Las técnicas actuales se basan en determinar una cierta función distancia entre dichas secuencias, que busca informar la proximidad que existe entre ellas. Algunas de estas son: métodos visuales, fuerza bruta, programación dinámica, métodos heurísticos, alineamiento global, alineamiento local, FASTA y BLAST entre otros [BRUNGGER Y LABHEAD,2003] Aplicar estas técnicas en la estructura primaria de una proteína no brinda la información necesaria para definir su funcionalidad, ya que no se describen la forma como los aminoácidos se interrelacionan entre sí con base en la estructura secundaria y terciaria [AUNG Z., Y K.L. TAN, 2007], es por esto que se hace necesario aplicar métodos alternativos de identificación.

1.3 ANÁLISIS DE CLUSTERS HIDROFÓBICOS (HCA)

A diferencia de las técnicas antes mencionadas las cuales se centran en la estructura primaria de una proteína, HCA se basa en el estudio de la estructura secundaria. Donde este es la base teórica de HCA. Este método, se enfoca en los casos donde los métodos lineales, como BLAST o FASTA no ofrecen ninguna información concluyente dicha información se presenta cuando los procesos de comparación arrojan un porcentaje de similitud entre el 25% y el 35%. Dicho rango de valores de baja similaridad es conocida como twligth-zone. [SILVA, 2007] [GABORIAUD, C., BISSERY, V., BENCHETRIT, T & MORNON, and J.P, 1987]

Esta técnica compara secuencias divergentes de proteínas a través de la información de la estructura que está implícita en la construcción de clusters hidrofobicos. Conformados por los aminoácidos (V, I, L, F, M, Y, W), designados con valor 1 y los demás con 0, creando una codificación binaria ver figura 7 con subsecuencias de proteínas que definen las regiones hidrofóbicas [EUDES, 2007].

Figura 7. Secuencia lineal y Código Binario



Fuente: Elaboración propia

Además de este tipo de codificación binaria se encuentra también la codificación Peitsch code (código P) y la codificación (Q), siendo P la más utilizada, ya que describe los clusters de forma más sencilla, especialmente aquellos de gran longitud.

1.3.1 CODIFICACIÓN PEITSCH CODE (CÓDIGO P)

Trascripción decimal del código binario que permite una útil descripción de los clusters en términos de almacenamiento y manipulación [EUDES, 2007].

1.3.2 CODIFICACIÓN Q

Este código contiene el alfabeto B = {V, M, U, D} donde la correspondencia se da de la siguiente manera: V a 11, M a 101, U a 1001 y D a 10001 [EUDES, 2007].

Todos estos esquemas de codificación, permiten mediante una construcción grafica bidimensional, ver figura 6 comparar proteínas cuya similitud en la secuencia no es suficiente para inferir conclusiones respecto a su funcionalidad, pero que desde su contenido estructural si se pueden hallar dichas conclusiones. Sin embargo, esta metodología a pesar de dar resultados en donde otras no los ofrecen. No es una técnica que haya sido explorada desde el punto de vista de automatización y por ende consultas sistemáticas sobre bases de datos no son posibles.

1.4 BASE DE DATOS DE INFORMACIÓN BIOLÓGICA

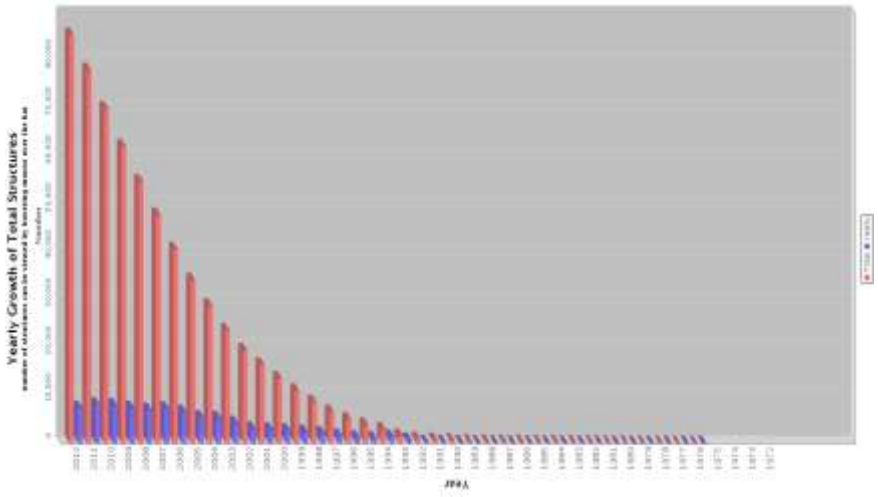
Las bases de datos biológicas son instrumentos de gran importancia para los científicos, ya que ayudan a comprender y explicar una serie de fenómenos biológicos, desde la estructura biomolecular y su interacción, hasta el

metabolismo completo de los organismos y la comprensión de la evolución de las especies. Este conocimiento ayuda a facilitar la lucha contra las enfermedades, el desarrollo de medicamentos, y el descubrimiento de las relaciones básicas entre las especies en la historia de la vida entre otras. Dicho conocimiento se encuentra distribuido entre múltiples bases de datos, generales y especializadas como lo son InterPro, PROSITE, ENZYME, PIR y PDB. Haciendo que sea difícil garantizar la coherencia de la información.

1.4.1 PROTEIN DATA BANK (PDB)

Una base de datos biológica a destacar es el PDB, la cual alberga el contenido estructural de proteínas y ácidos nucleicos. Administrado por diversas organizaciones asociadas, responsables del depósito, mantenimiento, procesamiento y libre suministro de estos datos biológicos para la comunidad científica.

Figura 8. Crecimiento de las estructuras de la PDB (Protein Data Bank)



Fuente: http://www.rcsb.org

Inicialmente el PDB contenía tan sólo 7 estructuras de proteínas. Desde entonces ha experimentado un crecimiento exponencial llegando a un número de 85582estructuras relativas a gran cantidad de moléculas biológicas [PDB].Esta base de datos sigue siendo hoy en día referente de comparación para el rendimiento de los diferentes algoritmos de computación, pues contiene información no redundante del espacio de proteínas, por ello fue usada en este trabajo para validar el algoritmo computacional que se desarrolló.

Cuando se dispone de una vasta cantidad de información, se hace necesario abordar el estudio de la misma desde nuevas perspectivas, si bien existen diversas metodologías enfocadas en el análisis de información, estos procesos pueden aún ser optimizados. Con el fin de satisfacer esta necesidad, se propone el empleo de disciplinas que tengan por objeto el procesamiento de grandes conjuntos de información biológica.

1.5 MINERÍA DE DATOS

Para poder tener un adecuado acceso a toda la información biológica disponible en las bases de datos antes mencionadas, se hace uso de la minería de datos la cual es vista como una herramienta o disciplina orientada al procesamiento y análisis de información. Ciencia que tiene por objeto la extracción de información no trivial, previamente desconocida, residente de manera implícita en grandes almacenes de datos. Contiene todo un conjunto de técnicas como lo son: series de tiempo, redes bayesianas, árboles de decisión, clustering, y detección de anomalías; encaminadas a la extracción de conocimiento procesable contenido en las bases de datos [MINERIA DE DATOS]

1.5.1 TÉCNICA DE MINERÍA DE DATOS CLUSTERING

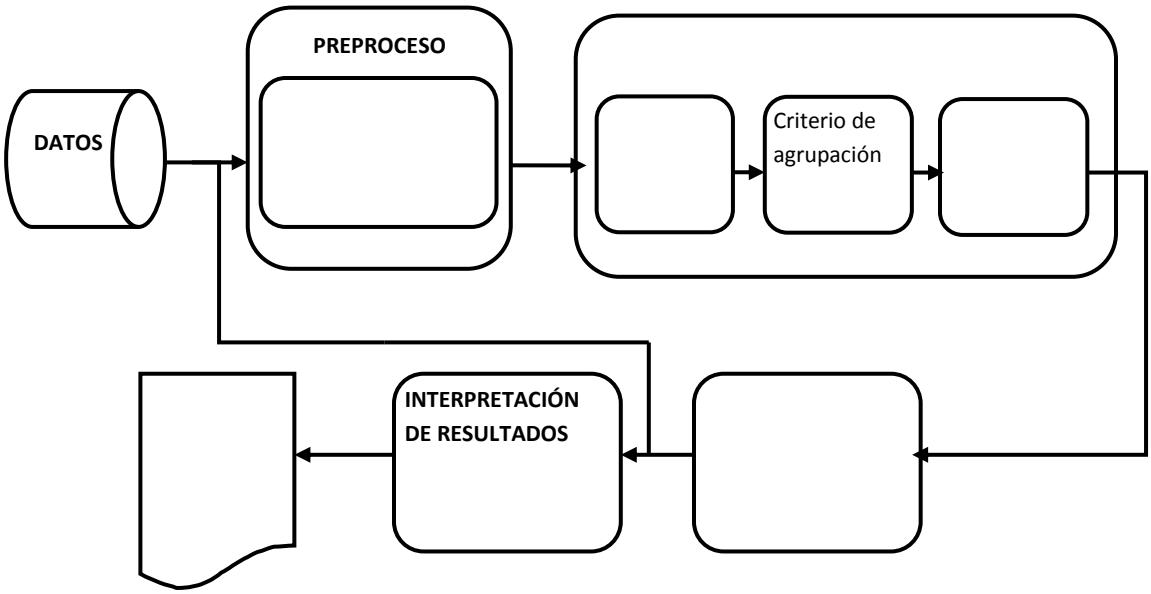
Una de las técnicas propias de la minería de datos y la cual se consideró para la reorganización de una base de datos haciendo uso de la técnica HCA fue

clustering la cual se describe a continuación, la cual divide un conjunto de objetos de tal forma que los miembros de cada grupo son similares de acuerdo a alguna métrica. El agrupamiento por similitud consiste en trasladar una medida intuitiva de semejanza dentro de una medida cuantitativa. El objetivo es hacer una clasificación en grupos, de manera que los elementos dentro de un conjunto sean afines entre si y distintos de los de otros grupos. En definitiva, se puede afirmar que el objetivo final de las técnicas de agrupamiento es organizar, a partir de medidas de similitud o disimilitud, los datos con algún [FAYYAD U M, PIATETSKY G, & SMYTH P, 2012] [RIMELQUE, RUIZ & GIBERT, 2006].

1.5.2 FASES DE ANÁLISIS CLUSTERING

En los métodos de clasificación no supervisados, cuando se desea aplicar alguna de las técnicas de agrupamiento a un problema concreto, se debe tener en cuenta una serie de fases para obtener un resultado adecuado. Estas fases se pueden resumir en [GALLARDO, 2009]:

Figura 9. Etapas del proceso Clustering



Fuente: Elaboración propia

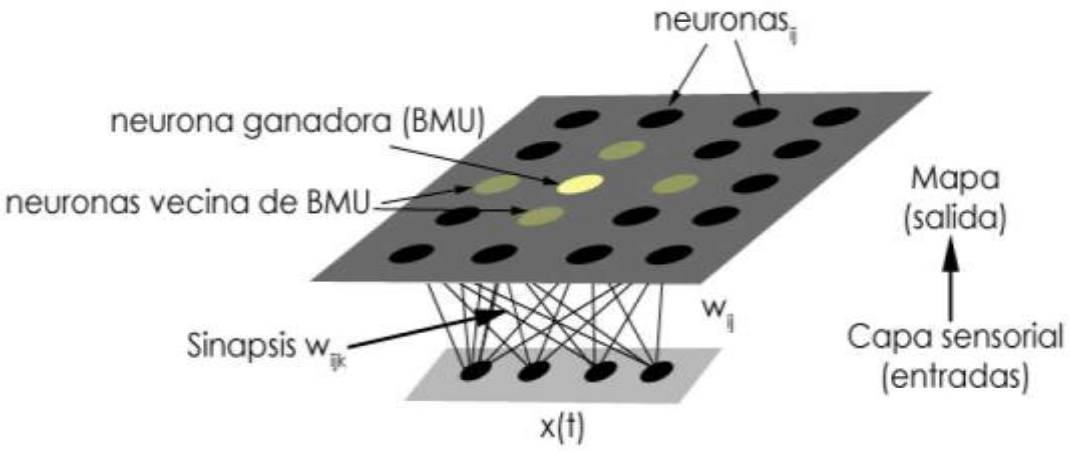
- Elección de variables: Las cuales dependerán del planteamiento del problema).
- Pre proceso: Tratamiento previo de los datos con el fin de garantizar su adecuado procesamiento.
- Extracción de características: Se encarga de seleccionar aquellos datos que son más relevantes para la clasificación y reducir la dimensión de los datos de entrada.
- Selección de características: Consiste en escoger un subconjunto de datos de entrada, eliminando aquellas características con poca o ninguna información predictiva, es posible realizarla mediante: Principal Component Analysis (PCA), la cual es una técnica estadística de síntesis de la información, o reducción de la dimensión (número de variables).
- Selección de instancias: Pretende elegir los ejemplos que sean relevantes para una aplicación y lograr el máximo rendimiento.
- Diseño del algoritmo: Esta etapa es la principal y se divide en tres fases:
  - Medida de similitud o distancia: Permite saber cuánto se asemejan o difieren dos vectores de características o dos objetos.
  - Criterio para el agrupamiento: Depende del tipo de grupos que se quiera encontrar y de sí los parámetros de entrada son conocidos.
  - Selección del algoritmo: Elección de un algoritmo que se adapte a los requisitos del problema. Uno que permite realizar la clasificación de los datos es el SOM, cuya característica más importante es el concepto de aprendizaje en un vecindario o agrupación próximo a la neurona ganadora.
- Validación de los resultados: Una vez obtenidos, es necesario verificar que estos son correctos y que no hay fallos en la clasificación.
- Interpretación de los resultados**:** El objetivo final de las técnicas de agrupamiento es obtener algún significado o alguna estructura que ofrezca información anteriormente desconocida.

1.5.3 ALGORITMO DE AGRUPAMIENTO CLUSTERING

1.5.3.1 MAPAS AUTOORGANIZATIVOS SOM

Dentro de las herramientas de clustering se optó por un tipo de red neuronal no supervisada que busca rasgos comunes, regularidades, correlaciones o categorías entre los datos de entrada, y los incorpora a su estructura interna de conexiones. Por tanto, las neuronas deben auto organizarse en función de los estímulos (datos) procedentes del exterior. Estas compiten unas con otras con el fin de llevar a cabo una tarea dada. Se pretende que cuando se presente a la red un patrón de entrada, sólo una de las neuronas de salida (o un grupo de vecinas) se active. Por tanto, compiten por activarse, quedando finalmente una como vencedora, y anuladas todas las otras, que son forzadas a sus valores de respuesta mínimos. El objetivo de este aprendizaje es categorizarlos datos que se introducen en la red. Se clasifican valores similares en la misma categoría y, por tanto, deben activar la misma neurona de salida [HULLE, 2001].

Figura 10. Mapa Autoorganizativo de Kohonen [Bedregal, 2009]



Fuente: Fuente: Bedregal,”Agrupamiento de Datos utilizando técnicas MAM-SOM”,2008

La figura 10 representa un ejemplo de red SOM de dos capas, donde cada neurona de competición es una categoría y cada neurona de entrada está conectada con cada una de las células de la capa de competición. Previamente a la evaluación del conjunto de datos realizados por el SOM, es necesario entrenarla. Antes de iniciar el entrenamiento, a los vectores prototipo se les asignan valores iníciales. Esta inicialización puede ser realizada de tres formas distintas:

•Inicialización Aleatoria: Donde los vectores de peso son inicializados con valores aleatorios pequeños.

• Muestra de Inicialización: Donde los vectores de peso son inicializados tomando valores aleatorios pertenecientes al conjunto de entrada.

• Inicialización Lineal: Donde los vectores de peso son inicializados de manera ordenada a lo largo del subespacio lineal generado por los dos vectores propios principales del conjunto de entrada.

En cada paso de entrenamiento, un vector muestra del conjunto de entrada, es elegido al azar y una medida de similaridad es calculada entre este y todos los vectores de peso del mapa. La Best-MatchingUnit (BMU), es la unidad cuyo vector de peso tiene la mayor similaridad con la muestra de entrada. La similaridad es generalmente establecida por una medida de distancia. Luego de encontrar la BMU, los vectores prototipo del SOM son actualizados y estos, junto con sus vecinos topológicos son movidos más cerca del vector de entrada en el espacio de entrada. Esto acerca las BMU prototipo y sus vecinos al vector muestra.

Luego de analizar dos de los métodos más ampliamente utilizados para la realización Clustering, K-means y mapas Autoorganizativos SOM, y de comparar su funcionamiento y ventajas y desventajas, se optó por aplicación del último. A continuación se listan las razones principales para esta elección:

[LOCKEHEIDE SANDRA, 2007]

- A diferencia de K-means, que se enfoca en el análisis de las medias de los datos a clusterizar, SOM trabaja con distancias, conservando la topología de los datos. Realizando su clasificación de acuerdo a la distribución y no a la cantidad de información dentro de los mismos.

- Debido a que SOM preserva la topología del espacio de los datos, proyecta datos altamente dimensionales a un esquema de representación de baja dimensión.

- SOM utiliza un algoritmo de entrenamiento no supervisado, a diferencia de k-means, que para una adecuada clasificación de los datos requiere fuertemente de la asignación previa de los centroides. Facilitando su implementación.

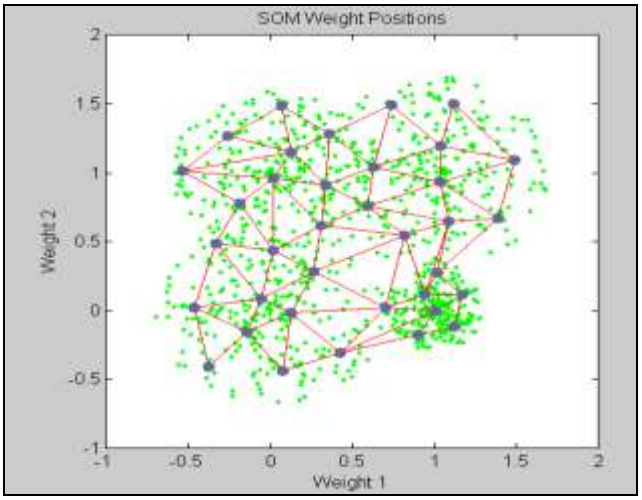
- SOM puede tomar un conjunto de datos de entrenamiento reducido, reduciendo la complejidad computacional y haciéndolo, a diferencia de K-means, menos sensible a valores atípicos.

- SOM es un algoritmo bastante robusto, ya que todos los vectores prototipo de características son afectados por todos los datos.

Adicional a esto, SOM cuenta con un completo conjunto de herramientas de visualización, facilitando análisis posteriores del conjunto de datos:

- **Mapa de puntos som:** Muestra la ubicación de los puntos de datos a agrupar y los vectores de peso como se indica en la figura 11.

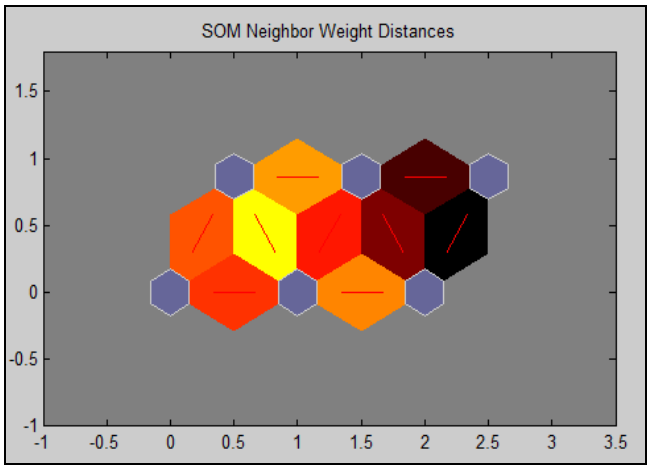
Figura 11 Herramienta Mapa de puntos Som



Fuente: http://www.mathworks.com/help/nnet/ug/self-organizing-feature-maps.html

- **plano de distancias som:** indica las distancias entre las neuronas vecinas. Esta figura utiliza el código de colores ver figura 12: Los hexágonos azules representan las neuronas. Las líneas rojas conectarlas neuronas vecinas. Los colores de las regiones que contienen las líneas rojas indican las distancias entre las neuronas .Los colores más oscuros representan grandes distancias. Los colores más claros representan distancias menores

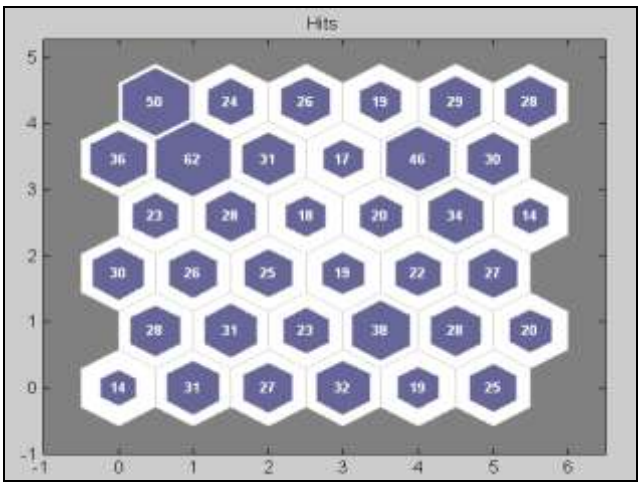
Figura 12 . Herramientas Plano de distancias SOM



Fuente: http://www.mathworks.com/help/nnet/ug/self-organizing-feature-maps.html

- **plano de puntos som:** Permite ver la distribución de los datos como se indica en la figura 13 a través de las neuronas en donde posteriormente son clasificados ver anexo B.

Figura 13. Herramientas Mapa de Distribución de Datos



Fuente: http://www.mathworks.com/help/nnet/ug/self-organizing-feature-maps.html

DESARROLLO DEL ALGORITMO COMPUTACIONAL

2.1 METODOLOGÍA

Para el desarrollo de este proyecto, se partió del estudio de la información contenida en el PDB, y de la teoría sobre la cual se sustenta la técnica HCA, utilizada para codificar los contenidos del PDB. Con el fin de comprender su funcionamiento, tanto del PDB como de HCA se prosiguió con la búsqueda y estudio de las distintas metodologías acordes que permitieran la solución de dicha problemática. Destacándose, dentro del campo de la Minería de Datos, la utilización de Mapas Autoorganizativos (SOM). El siguiente paso consistió en la elección una metodología que permitiera preparar los datos adecuadamente para la aplicación del SOM, siendo la elegida, el Vector de Composición de Momento (VCM) [RUAN J, WANG K, YANG J, KURGAN L.A. & CIOS K.J, 2005], Seleccionadas las herramientas y el enfoque con el cual abordar la problemática, se procedió a la ejecución de este proyecto.

2.2 PLANTEAMIENTO DEL PROBLEMA

HCA permite establecer el porcentaje de similaridad entre secuencias de proteínas altamente divergentes teniendo en cuenta su contenido estructural y no su secuencia de aminoácidos. Sin embargo, se basa en interacción humana experta y no tiene la capacidad de realizar comparaciones sistemáticas en bases de datos conteniendo grandes conjuntos de información.

Existen enfoques que intentan automatizar el proceso de comparación empleando esta técnica pero que no contemplan la navegación sobre las bases de datos con

información biológica, lo cual genera inconvenientes a la hora de realizar búsquedas exhaustivas sobre las mismas [SILVA, 2007].

En este trabajo, se planteó la necesidad de reducir los tiempos de consulta en una base de datos con 196.393 secuencias de proteínas, sobre la cual se necesita realizar procesos de anotación de proteínas provenientes del genoma del café. Esto implica comparar 24.994 secuencias de proteínas de este genoma simultáneamente con las 196.393 secuencias presentes en la base de datos que se tomó como referencia, para lo cual se necesitaría realizar cerca de 4.908’646.642 cotejos, los cuales, tomando como tiempo de ejecución por comparación228.1 ms, requerirían cerca de 35.5 años para la realización del proceso de anotación.

La técnica HCA empleada tiene como finalidad la comparación de secuencias teniendo en cuenta su contenido estructural, el cual se encuentra representado mediante tres formas de codificación anteriormente mencionadas. El código P, el código Q, y el código binario, como se mencionó en el capítulo anterior para poder analizar una base de datos con información biológica debe realizarse una codificación general de la base de datos referencia en estos tres formatos, y estructurar la información de manera que se realice un proceso de anotación empleando este tipo de información y que permita reducir los tiempos de comparación de las secuencias pertenecientes a algún genoma.

2.3 PLANTEAMIENTO DE LA SOLUCIÓN

Partiendo de un conjunto de 196.393 secuencias de proteínas almacenadas en el PDB y agrupadas de acuerdo a las tres distintas codificaciones empleadas por HCA, se procedió a realizar el pre procesamiento de los datos.

Primero, se filtraron los clusters, seleccionando únicamente aquellos cuya longitud original fuera mayor a 1, es decir, los que contuvieran más de un aminoácido. Esto

se hizo debido a que los aminoácidos hidrofóbicos aislados, al no estar asociados con ningún otro, no conforman realmente un cluster, además de poder estar contenidos en otros y en distintas posiciones. Por esto, se considera que no aportan información significativa al proceso de notación de secuencias, pero añadirían mayor carga de trabajo.

El siguiente paso fue codificar los clusters seleccionados utilizando VCM, el cual permite extraer información del contenido de los aminoácidos de un clúster o secuencia, así como la ubicación y la frecuencia de los aminoácidos dentro de la misma. El VCM funciona del siguiente modo.

Algoritmo 1VCM Código Q

- La cadena de aminoácidos hidrofobicos O={o_1, o_2, …, o_n}
- Sea A={#, V, M, U, D, *} el alfabeto de la codificación Q
- Sea N la longitud de la cadena O
- Sea A_i el-iesimo aminoácido, cuando los aminoácidos se ordenan de la forma que se muestra en A
- Para un w > 0 donde w ϵ Z, se define (x_1^w, x_2^w, …, x_6^w) como el Vector Composición de Momento de Orden w para el código Q
- $x_i^w = \frac{\sum_{j=1}^{W_i} n_{i,j}^w}{\prod_{d=0}^{w}(N-d)}$ para i=1, 2, …,6
- Donde ni,j es la j-esima posición del i-esimo aminoácido en O

Algoritmo 2VCM Código P

- La cadena de aminoácidos hidrofobicos O={o_1, o_2, …, o_n}
- Sea A={0, 1, 2, 3, 4, 5, 6, 7, 8, 9, *} el alfabeto de la codificación P
- Sea N la longitud de la cadena O
- Sea A_i el-iesimo aminoácido, cuando los aminoácidos se ordenan de la forma que se muestra en A
- Para un w > 0 donde w ϵ Z, se define (x_1^w, x_2^w, …, x_{11}^w) como el Vector Composición de Momento de Orden w para el código P
- $x_i^w = \frac{\sum_{j=1}^{W_i} n_{i,j}^w}{\prod_{d=0}^{w}(N-d)}$ para i=1, 2, …,11
- Donde ni,j es la j-esima posición del i-esimo aminoácido en O

Algoritmo 3 VCM Código Binario

- La cadena de aminoácidos hidrofobicos O={o_1, o_2, …, o_n}
- Sea A={0, 1} el alfabeto de la codificación Binaria
- Sea N la longitud de la cadena O
- Sea A_i el-iesimo aminoácido, cuando los aminoácidos se ordenan de la forma que se muestra en A

  Para un w > 0 donde w ϵ Z, se define (x_1^w, x_2^w) como el Vector Composición de Momento de Orden w para el código Binario
- $x_i^w = \frac{\sum_{j=1}^{W_i} n_{i,j}^w}{\prod_{d=0}^{w}(N-d)}$ para i=1, 2
- Donde ni,j es la j-esima posición del i-esimo aminoácido en O
-

De esta manera, el VCM redimensionó los conjuntos de clusters, que originalmente poseen longitudes variables, oscilando entre 2 y 85 aminoácidos. Obteniendo como resultado tres conjuntos de vectores de dimensiones 22x1 para todos los clusters codificados en código p, 12x1 para los correspondientes al código q y 4x1 para aquellos en código binario. A continuación se realizó el entrenamiento de 3 redes SOM para la clasificación de los conjuntos de vectores obtenidos, dando como resultado un índice, correspondiente al grupo al que pertenece cada cluster, y una representación o mapa, en el cual se observa la distribución topológica de los datos.

Luego de realizar pruebas con distinto número de grupos, se decidió clasificar los clusters en 6 conjuntos, ya que para este número, se observaba una mejor distribución de la información, evitando grupos vacíos o demasiado poblados. En la figura 14 se observa la distribución de los clusters por grupos, generada por el SOM (ver Anexo B) y en la figura 15 se visualiza de manera gráfica el comportamiento de la distribución.

Figura 14 Distribución de clusters correspondientes al código Q, por el algoritmo SOM

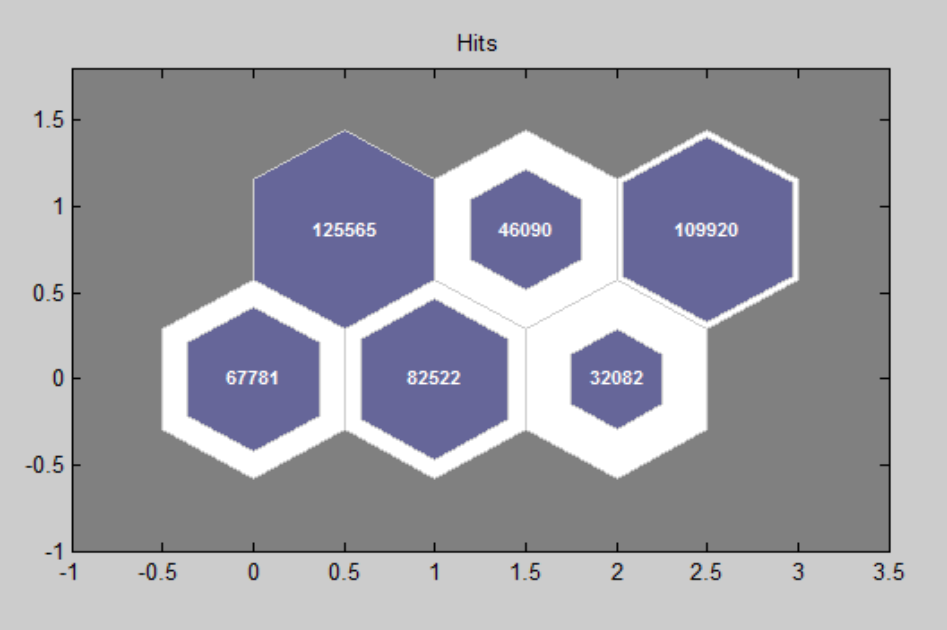
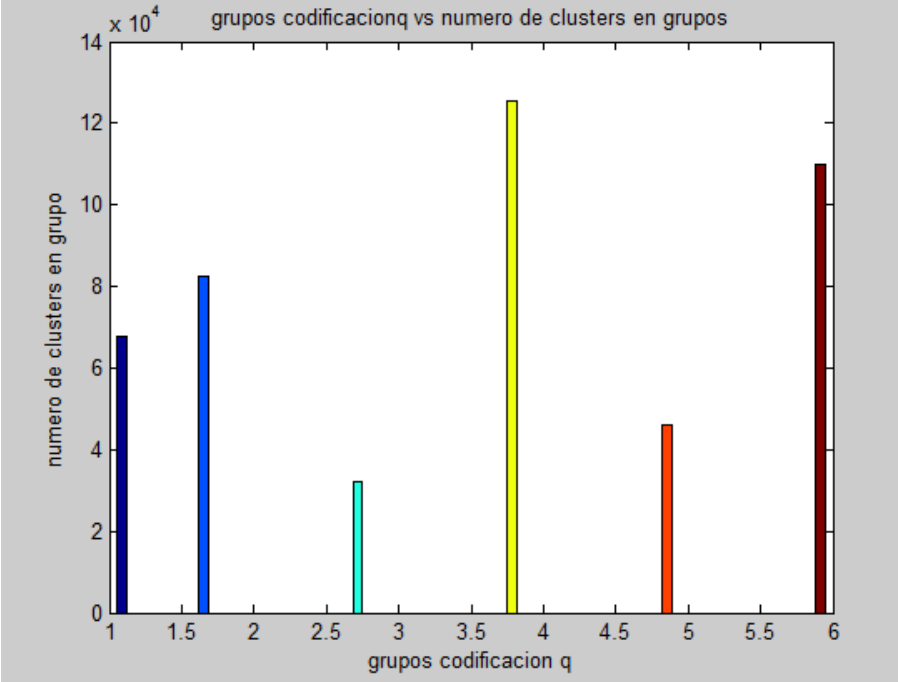


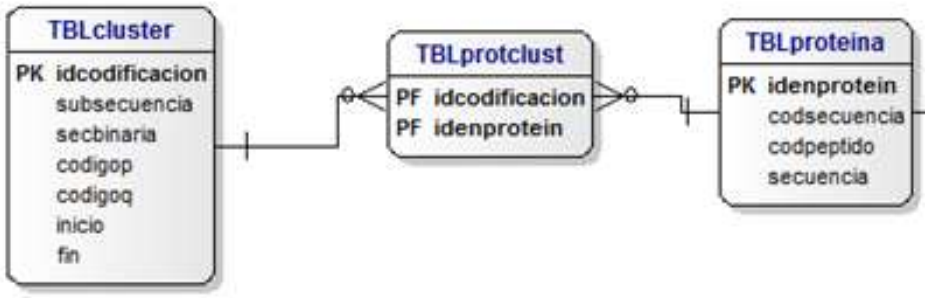
Figura 15. Histograma de los grupos clasificados en la codificación Q por el algoritmo Som



Una vez obtenida la clasificación hecha por el SOM, se procedió a reestructurar la base de datos actual (ver Anexo C).Obteniendo como resultado 3 bases de datos, una para cada tipo de codificación HCA.

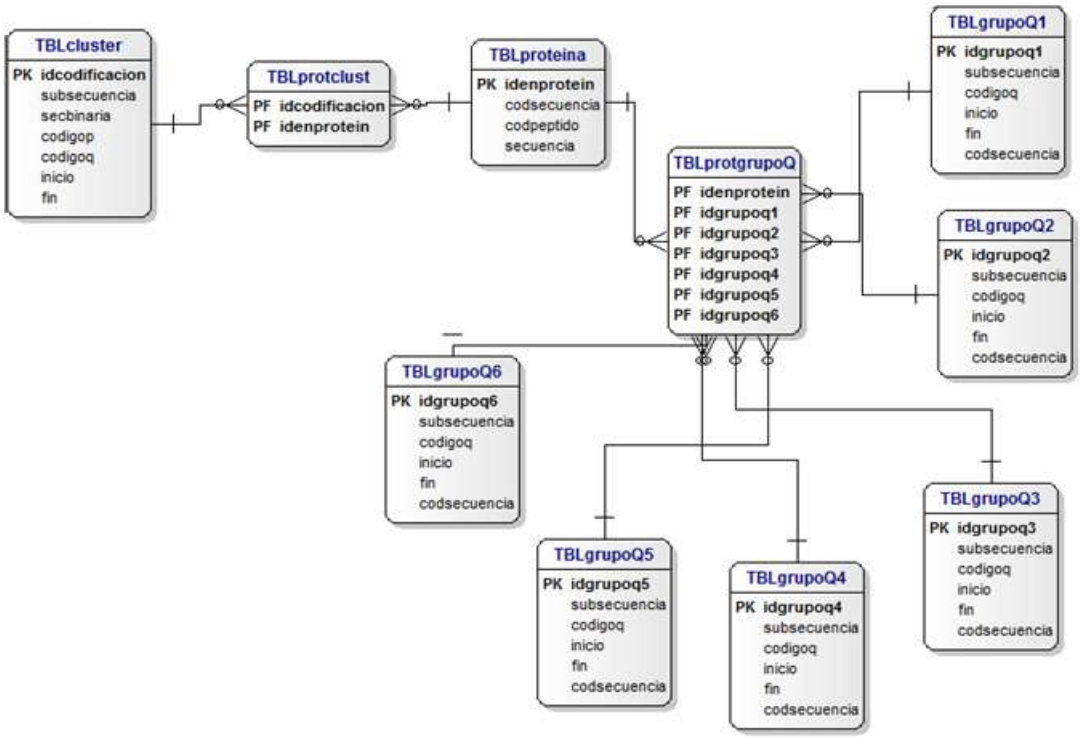
Cada una contiene 6 tablas, correspondientes a igual número de grupos, en los cuales el SOM clasificó los clusters hidrofobicos en su respectiva codificación (Q, P o Binaria), contiene también la subsecuencia original y un identificador único del cluster dentro de su grupo. Además, almacena información referente a la secuencia a la cual dicho cluster pertenece, como lo es su código de secuencia, y sus posiciones inicial y final dentro de la misma. Estas tablas se encuentran relacionadas entre sí y con sus secuencias origen mediante una tabla índice, facilitando la búsqueda de secuencias o clusters similares.

Figura 16 Base de datos original conteniendo secuencias de proteínas codificadas en HCA.



Fuente: Elaboración propia

Figura 17. Base de datos reestructurada para código Q

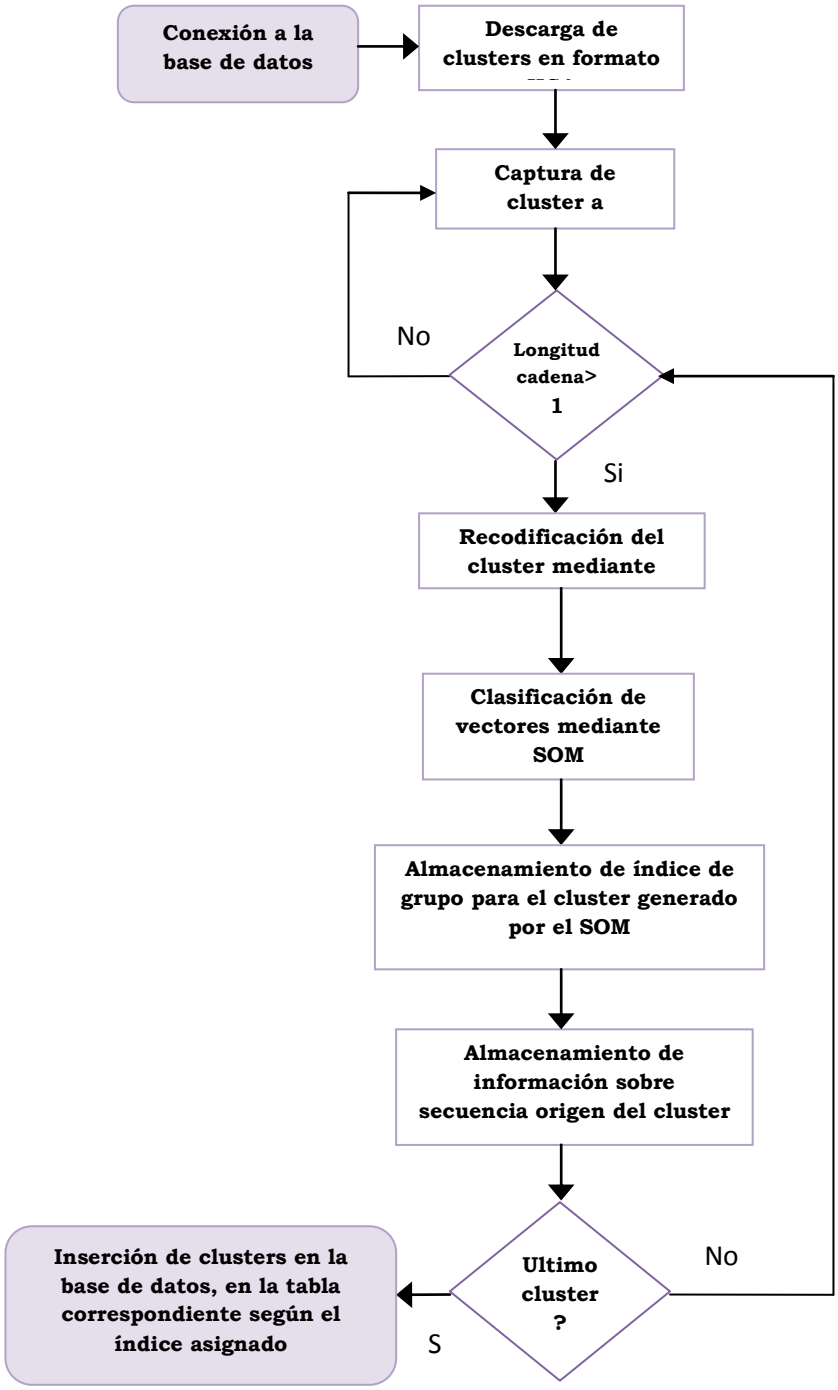


Fuente: Elaboración propia

Como se puede observar, se partió de la base de datos inicial obtenida del PDB, la cual se expandió, agregándole 7 nuevas tablas; 6 en las cuales se almacenaron los clusters según fueron clasificados por el SOM, y una tabla adicional, tblgrupoq, encargada de mantener estos clusters ya clasificados, relacionados con sus secuencias origen. Este mismo proceso se realizó para la codificación P y la codificación Binaria (ver Anexo C).

A continuación se puede observar el procedimiento general utilizado para la clasificación de las secuencias almacenadas en el PDB y codificadas en HCA.

Figura 18. Algoritmo general para clasificación de secuencias utilizando



Fuente: Elaboración propia

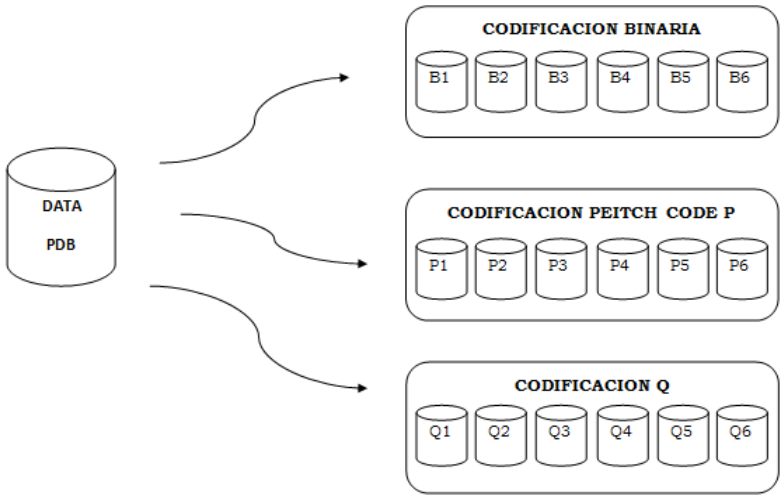
Finalmente, se procedió a realizar un conjunto de pruebas sobre las bases de datos obtenidas, tomando una serie de muestras aleatorias. Estas muestras fueron comparadas utilizando HCA e indicando un alto grado de similaridad entre los clusters agrupados. Validando así el agrupamiento realizado por el SOM.

2.4 RESULTADOS Y DISCUSIÓN

2.4.1 ENTRENAMIENTO Y PRUEBAS

La validación del rendimiento del algoritmo de clusterizacion se realizó empleando la técnica HCA, con la cual se hace evidente que el proceso de clasificación, fue aceptable. Se utilizó la base de datos PDB, que alberga un numero de 79.265 secuencias a las que a su vez, les fue aplicado un filtrado de datos, con el fin de eliminar información redundante, trabajando así con un numero de 24.664 secuencias codificadas en formato HCA, a su vez constituidas por 599.130 subsecuencias o “clusters”, datos con los que se llevó a cabo el proceso de agrupación como se indica en la figura 19

Figura 19. Proceso de clusterizacion



Fuente: Elaboración propia

Esta gráfica representa la nueva configuración que adquirieron los clusters almacenados en el PDB, divididos según el tipo de codificación (código Q, código P y código Binario).

Para evaluar la efectividad de la clasificación hecha por el algoritmo de agrupamiento, se extrajeron de forma aleatoria 384 clusters, designados como datos de evaluación.

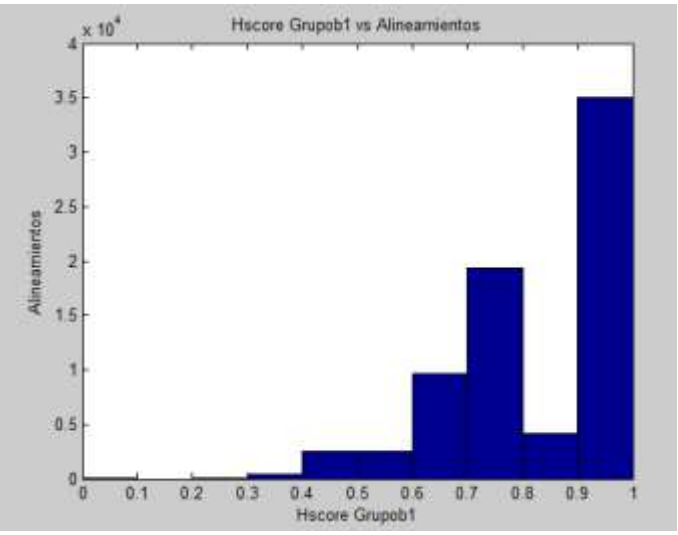
La prueba aplicada para valorar el rendimiento de los resultados generados por el algoritmo de clasificación fue el porcentaje de similaridad HCA Score (Hscore) generado por el enfoque de [Silva, 2007], el cual mide el porcentaje de aminoácidos hidrofobicos topológicamente conservados, sirviendo como medida de similaridad entre dos secuencias a comparar. En el cálculo del Hscore, las regiones de dos proteínas se dividen en un determinado número de secciones, tomando como base la similitud en la forma y la distribución de los clusters contenidos en dichas áreas. La segmentación es arbitraria, pero se elige de modo que ayude a maximizar el emparejamiento de clusters hidrofobicos entre las secuencias correspondientes en una sección. Los aminoácidos emparejados entre estos clusters son identificados, aquellos aminoácidos individuales no encontrados dentro de un cluster son ignorados. Se cuentan los aminoácidos hidrofobicos emparejados (MHA) en cada sección de las dos proteínas y el número total de aminoácidos hidrofobicos (THA) en todos los clusters. Una vez obtenidos estos valores, se procede a calcular el Hscore:

$$\text{HCA Score (\%)} = (2 \times \text{MHA}/\text{THA}) \times 100$$

Es considerado que si el Hscore entre dos proteínas es superior a 60%, entonces dichas proteínas son similares [KRAIWATTANAPONG J, OOI T, KINOSHITA S, SUGIMURA I, SAWABE T & EZURA Y, 2000].

.Esta prueba fue realizada a cada uno de los grupos correspondientes a las tres codificaciones HCA (ver Anexo D), allí se observó que la mayoría de los clusters almacenados en estos grupos presentan un HCA score que oscila entre 60% y 100%, cumpliendo con el requisito de HCA respecto a similaridad y confirmando que el proceso de agrupamiento realizado por el SOM fue realizado de forma adecuada.

Figura 20. Análisis de similaridad para grupob1, codificación binaria.



Fuente:Elaboracion Propia

2.4.2 RESULTADOS OBTENIDOS CLASIFICACIÓN CLUSTERS

Adicionalmente para contrastar los porcentajes obtenidos mediante la utilización de HCA, se calcularon los valores de la varianza y la media del Hscore para cada codificación. Estos valores arrojan una idea del rendimiento del algoritmo de clusterizacion, logrando los siguientes resultados (ver Tabla 1).

Tabla1. Rendimiento alcanzado por el clasificador de acuerdo a los diferentes tipos de codificación.

Grupos	Código Q		Código P		Código R (Binario)	
	% Media	% Varianza	% Media	% Varianza	% Media	% Varianza
Grupo 1	91.10	2.06	83.67	2.82	85.16	2.50
Grupo 2	86.97	2	80.46	2.56	85.04	1.97
Grupo 3	90.31	2.07	80.39	2.62	100	0
Grupo 4	79.05	2.14	80.32	3.06	85.59	2.32
Grupo 5	87.94	2.24	82.45	2.54	93.28	1.60
Grupo 6	91.11	3.70	89.53	3.45	100	0

En la Tabla 1 se observa que el valor de similaridad medio para cada grupo es bastante alto, oscilando entre 79,05% y 91,11% para el código Q, entre 80,31% y 89,53% para el código P y entre 85,04% y 100% para la codificación Binaria. Se observa también una baja dispersión en los tres casos, garantizando que la media obtenida es homogénea. Todo lo anterior proporciona seguridad para afirmar, de nuevo, que el agrupamiento realizado por el SOM fue particularmente eficiente.

2.4.3 RESULTADOS OBTENIDOS SIMILITUD ENTRE SECUENCIAS

En lo referente a la agilización del proceso de notación de secuencias de proteínas, se evidencia la mejora en cuanto a selección y análisis de información, dado que ahora el proceso de comparación no es realizado contra toda la información almacenada en la PDB, sino con aquellas secuencias que presentan un alto grado de similaridad,

Para ello se diseñó un método de prueba de mejora en el rendimiento, para ello se tomaron 20 secuencias por tipo de codificación a las que se les realizo una búsqueda de secuencias similares así como su tiempo de comparación lo cual se expone en las tablas 3,5 y 7.

TABLA 2.PRUEBA # 1

Sec. 1 101m	>>MVLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRVKH LKTEAEMKASEDLKKHGVTVLTALGAILKKKGHHEAELKPLAQSHATKHKIPI KYLEFISEAIIHVLHSRHPGNFGADAQGAMNKALELFRKDIAAKYKELGYQG			
Nº	**Clusters**	**Código Q**	**Grupo**	**Secuencias que lo contienen**
1	MVL	VV	1	73
2	WQLVLHVWAKVEAD V	MVVMVUD	2	1
3	ILIRLF	VVMV	2	1
4	LEKFDRVKHL	UUU	5	1
5	VTVLTALGAIL	MVUUV	4	1
6	IKYLEFISEAIIHVL	MVMVDVMV	2	2
7	MNKALELFRKDIAAK YKELGY	DMVDDUM	5	1
			Total de secuencias*	**74**

* El total de secuencias puede diferir de la suma de las Secuencias que contienen el cluster, ya que distintos clusters pueden estar contenidos en la misma secuencia.

Tabla 3.Resultados para 20 cadenas de secuencias codificación Q

Idenprotein	codsecuencia	Secuencia	N ° sec a comparar	Tiempo [s]
4	11ba	>>KESAAAKFERQHMDSGNSPSSSSNYCNLMMCCRKMTQGKCKPVNTFVHESLADVKAVCSQKKVTCKNGQTNCYQSKSTMRITDCRETGSSKYPNCAYKTTQVEKHIIVACGGKPSVPVHFDASV	39	8,8959
18	192l	>>MNIFEMLRIDEGLRLKIYKDTEGYYTIGIGHLLTKSPSLAAAKAALAAAIGRNTNGVITKDEAEKLFNQDVDAAVRGILRNAKLKPVYDSLDAVRRAALINMVFQMGETGVAGFTNSLRMLQQKRWAAAAAALAKSRWYNQTPNRAKRVITTFRTGTWDAYKNL	2038	464,8678
24	1a0b	>>TTEENSKSEALLDIPMLEQYLELVGPKLITDGLAVFEKMMPGYVSVLESNLTAQDKKGIVEEGHKIKGAAGSVGLRHLQQLGQQIQSPDLPAWEDNVGEWIEEMKEEWRHDVEVLKAWVAKATKK	1666	380,0146
50	1a2i	>>APKAPADGLKMEATKQPVVFNHSTHKSVKCGDCHHPVNGKEDYRKCGTAGCHDSMDKKDKSAKGYYHVMHDKNTKFKSCVGCHVEVAGADAAKKKDLTGCKKSKCHE	193	44,0233
144	1aac	>>DKATIPSESPFAAAEVADGAIVVDIAKMKYETPELHVKVGDTVTWINREAMPHNVHFVAGVLGEAALKGPMMKKEQAYSLTFTEAGTYDYHCTPHPFMRGKVVVE	279	63,6399

186	1afa	>>AIEVKLANMEAEINTLKSKLELTNKL HAFSMGKKSGKKFFVTNHERMPFSK VKALCSELRGTVAIPRNAEENKAIQE VAKTSAFLGITDEVTEGQFMYVTGGR LTYSNWKKDQPDDWYGHGLGGGED CVTIVDNGLWNDISCQASHTAVCEFP A	68	15,5108
347	1auz	>>SLGIDMNVKESVLCIRLTGELDHHT AETLKQKVTQSLEKDDIRHIVLNLEDL SFMDSSGLGVILGRYKQIKQIGGEMV VCAISPAVKRLFDMSGLFKIIRFEQSE QQALLTLGVAS	1	0,2281
1032	1d2z	>>LSSKYSRNTELRRVEDNDIYRLAKI LDENSCWRKLMSIIPKGMDVQACSG AGCLNFPAEIKKGFKYTAQDVFQIDE AANRLPPDQSKSQMMIDEWKTSGKL NERPTVGVLLQLLVQAELFSAADFVA LDFLNESTPARPVDGPGALISLELLE	101	23,0381
2019	1g87	>>AGTYNYGEALQKSIMFYEFQRSGD LPADKRDNWRDDSGMKDGSDVGVD LTGGWYDAGDHVKFNLPMSYTSAML AWSLYEDKDAYDKSGQTKYIMDGIK WANDYFIKCNPTPGVYYYQVGDGGK DHSWWGPAEVMQMERPSFKVDASK PGSAVCASTAASLASAAVVFKSSDPT YAEKCISHAKNLFDMADKAKSDAGYT AASGYYSSSSFYDDLSWAAVWLYLA TNDSTYLDKAESYVPNWGKEQQTDII AYKWGQCWDDVHYGAELLLAKLTNK QLYKDSIEMNLDFWTTGVNGTRVSY TPKGLAWLFQWGSLRHATTQAFLAG VYAEWEGCTPSKVSVYKDFLKSQIDY ALGSTGRSFVVGYGVNPPQHPHHRT AHGSWTDQMTSPTYHRHTIYGALVG	4494	1025,08 14

		GPDNADGYTDEINNYVNNEIACDYNA GFTGALAKMYKHSGGDPIPNFKAIEKI TNDEVIIKAGLNSTGPNYTEIKAVVYN QTGWPARVTDKISFKYFMDLSEIVAA GIDPLSLVTSSNYSEGKNTKVSGVLP WDVSNNVYYVNVDLTGENIYPGGQS ACRREVQFRIAAPQGTTYWNPKNDF SYDGLPTTSTVNTVTNIPVYDNGVKV FGNEP		
4092	1njk	>>MGSSHHHHHHSSGRENLYFQGH MQTQIKVRGYHLDVYQHVNNARYLE FLEEARWDGLENSDSFQWMTAHNIA FVVVNININYRRPAVLSDLLTITSQLQ QLNGKSGILSQVITLEPEGQVVADALI TFVCIDLKTQKALALEGELREKLEQM VKGH	9	2,0529
6301	1v30	>>MSYKEKSVRIAVYGTLRKGKPLHW YLKGAKFLGEDWIEGYQLYFEYLPYA VKGKGKLKVEVYEVDKETFERINEIEI GTGYRLVEVSTKFGKAFLWEWGSKP RGKRIKSGDFDEIRLEHHHHHH	61	13,9141
7174	1x5b	>>GSSGSSGMPLFTANPFEQDVEKA TNEYNTTEDWSLIMDICDKVGSTPNG AKDCLKAIMKRVNHKVPHVALQALTL LGACVANCGKIFHLEVCSRDFATEVR AVIKNKAHPKVCEKLKSLMVEWSEEF QKDPQFSLISATIKSMKEEGITFPPAG SQTSGPSSG	1477	336,903 7
13821	2ra9	>>GQHTLKQFAADSALTTTTPLCSEV PLFDINALGDWTYLGTSLPAKFAKLF ASILHCIDDEYFLITPVEKVRVQVEDA PLLIVDFERAQPHSLLNVSTSIGTLHH	106	24,1786

		NVDIKQMKLTDDSVYLPLERGLWGKL GRACYYNFVNEFNLSDLNEQ		
15295	2y78	>>"MGSSHHHHHHSSGLVPRGSHMT VVTTESGLKYEDLTEGSGAEARAGQ TVSVHYTGWLTDGQKFDSSKDRNDP FAFVLGGGMVIKGWDEGVQGMKVG GVRRLTIPPQLGYGARGAGGVIPPNA TLVFEVELLDV	3	0,6843
15477	2yug	>>GSSGSSGLDIVGIWWTVSNFGEIS GTIAIEMDKGAYIHALDNGLFTLGAPH REVDEGPSPPEQFTAVKLSDSRIALK SGYGKYLGINSDGLVVGRSDAIGPRE QWEPVFQDGKMALLASNSCFIRCNE AGDIEAKNKTAGEEEMIKIRSCAERET	2654	605,377 4
17150	3d4t	>>MRGSHHHHHHGSDDDDKAELRLL MFEQPGCLYCARWDAEIAPQYPLTD EGRAAPVQRLQMRDPLPPGLELARP VTFTPTFVLMAGDVESGRLEGYPGE DFFWPMLARLIGQAEPGQ	368	83,9408
18038	3f7k	>>AIHAVCVLKGDSPVTGTIHLKEEGD MVTVTGEITGLTPGKHGFHVHEFGD NTNGCTSAGGHFNPHGKEHGAPED ENRHAGDLGNVVAGEDGKAVINMKD KLVKLTGPDSVIGRTLVVHVDEDDLG RGGHEQSKITGNAGGRLACGVIGITK E	10	2,281
19073	3hj7	>>MSGEKGYWFELPVPALLPLPNGY AIISEFGEHYPRKQAGNDWFVVDPAS VSLPLRVRTRRRGDRMVLKGTGGTK KLKEIFIEAKIPRMERDRWPIVEDADG RILWVPGLKKSAFEAQNRGQARYILL QYQAMNSLEHHHHHH	5466	1246,79 46
20262	3kjj	>>MAHHHHHHMDIRYFGTTPRYSEA	2	0,4562

		VGANGLIFLSGMVPENGETAAEQTA DVLAQIDRWLAECGSDKAHVLDAVIY LRDMGDYAEMNGVWDAWVAAGRTP ARACVEARLARPEWRVEIKITAVKRD AATA		
22167	3owr	>>GSKEDLPAYEEAEITKVGAYHRFY SGDKDAITGENIVAEKELDRTNNIDSE HGVATAVFTIPAAGGKFTEAERAKVS LSNLVVYVNVSTAARVTPLDGSPKFG VPADWTREHKYSVMAADGTKKIWTV KVTLNK	1723	393,016 3

Tabla 4. Prueba #1 secuencia codificación p

Sec. 4 11ba	>>"KESAAAKFERQHMDSGNSPSSSSNYCNLMMCCRKMTQGKCKPVNT FVHESLADVKAVCSQKKVTCKNGQTNCYQSKSTMRITDCRETGSSKYPN CAYKTTQVEKHIIVACGGKPSVPVHFDASV			
Nº	**Clusters**	**Código P**	**Grupo**	**Secuencias que lo contienen**
1	YCNLMM	39	2	2
2	VNTFVHESLADVK AV	19529	2	2
3	"MRI	5*	6	31
4	"VEKHIIV	71	1	1
5	VHFDASV	81	1	7
			Total de secuencias*	**34**

Tabla 5. Resultados para 10 cadenas de secuencias codificación P

Idenprotein	codsecuencia	Secuencia	N° sec a comparar	Tiempo [s]
8	135l	>>KVYGRCELAAAMKRLGLDNYRGYSLGNWVCAAKFESNFNTHATNRNTDGSTDYGILQINSRWWCNDGRTPGSKNLCNIPCSALLSSDITASVNCAKKIASGGNGMNAWVAWRNRCKGTDVHAWIRGCRL	1230	280,563
258	1alc	>>KQFTKCELSQNLYDIDGYGRIALPELICTMFHTSGYDTQAIVENDESTEYGLFQISNALWCKSSQSPQSRNICDITCDKFLDDDITDDIMCAKKILDIKGIDYWIAHKALCTEKLEQWLCEKE	1650	376,365
1383	1eal	>>AFTGKYEIESEKNYDEFMKRLALPSDAIDKARNLKIISEVKQDGQNFTWSQQYPGGHSITNTFTIGKECDIETIGGKKFKATVQMEGGKVVVNSPNYHHTAEIVDGKLVEVSTVGGVSYERVSKKLA	316	72,0796
3735	1m8t	>>HLVQFNGMIRCTIPGSIPWWDYSDYGCYCGSGGSGTPVDELDRCCQVHDNCYTQAQQLTECSPYSKRYSYDCSEGTLTCKADNDECAAFVCDCDRVAAICFAGAPYNKENINIDTTTRC	1225	279,4225
8050	1zhq	>>ASYKVNIPAGPLWSNAEAQQVGPKIAAAHQGNFTGQWTTVVESAMSVVEVELQVENTGIHEFKTDVLAGPLWSNDEAQKLGPQIAASYGAEFTGQWRTIVEGVMSVIQIKYTF	391	89,1871
11536	2j6y	>>MDFREVIEQRYHQLLSRYIAELTKTSLYQAQKFSRKTIEHQIPPEEIISIHRKVLKELYPSLPEDVFHSLDFLIEVMIGYGMAYQEHQTLRGIQQEIKSEIEIAANVQQTL	1114	254,1034

17336	3dl3	>>MSHLRIPKNWTIQRSTPFFTKDNVPEALLTHHNTAVDVFGQICVMEGVVTYYGFANSEATEPEIKVVINAGQFATSPPQYWHRIELSDDAQFNINFWSDQDKSGKKMFNTKLEHHHHHH	4151	946,8431
20381	3kuf	>>AELTVEVRGSNGAFYKGFIKDVHEDSLTVVFENNWQPERQVPFNEVRLPPPPDIKKEISEGDEVEVYSRANDQEPCGWWLAKVRMMKGEFYVIEYAACDATYNEIVTFERLRPVNQNKTVKKNTFFKCTVD	9	2,0529
21409	3n53	>>MSLKKILIIDQQDFSRIELKNFLDSEYLVIESKNEKEALEQIDHHHPDLVILDMDIIGENSPNLCLKLKRSKGLKNVPLILLFSSEHKEAIVNGLHSGADDYLTKPFNRNDLLSRIEIHLRTQNYYSDLRKNEGHHHHHH	1284	292,8804
24032	3tzu	>>GPGSMTDRKIPGDRSYTADHEWIDIAPGAATPDGPVRVGITSVAVEALGDLVFVQLPEVGETVSAGESCGEVESTKTVSDLIAPASGQIVEVNTAAVDDPATIATDPYGAGWLYSVQPTAVGELLTASEYAGQNGLS	15	3,4215

Tabla 6. prueba #1 codificación binario

Sec. 8 135l	>>"KVYGRCELAAAMKRLGLDNYRGYSLGNWVCAAKFESNFNTHATNRNTDGSTDYGILQINSRWWCNDGRTPGSKNLCNIPCSALLSSDITASVNCAKKIASGGNGMNAWVAWRNRCKGTDVHAWIRGCRL"			
Nº	**Clusters**	**Código Binario**	**Grupo**	**Secuencias que lo contienen**
1	VY	11	6	1222
2	LAAAMKRLGLDNYRGYSLGNWV	1000100101001001010011	1	1
3	FESNF	10001	1	5
4	YGILQINSRWW	10110100011	4	5
5	"LCNI	1001	4	5
6	LLSSDITASV	1100010001	1	4
7	MNAWVAW	1001101	2	5
8	VHAWI	10011	2	1
			Total de secuencias*	**1229**

Tabla 7.resultados para 10 cadenas de secuencias codificación binaria

Idenprotein	**codsecuencia**	**Secuencia**	**N° seca comparar**	**Tiempo [s]**
8	135l	>>KVYGRCELAAAMKRLGLDNYRGYSLGNWVCAAKFESNFNTHATNRNTDGSTDYGILQINSRWWCNDGRTPGSKNLCNIPCSALLSSDITASVNCAKKIASGGNGMNAWVAWRNRCKGTDVHAWIRGCRL	1229	280,3349
660	1bqk	>>ADFEVHMLNKGKDGAMVFEPASLKVAPGDTVTFIPTDKGHNVETIKGMIPDGAEAFKSKINENYKVTFTAPGVYGVKCTPHYGMGMVGV	239	54,5159

		VQVGDAPANLEAVKGAKNPKKAQERLDAALAALGN		
1911	1fvu	>>DCPSGWSSYEGNCYKFFQQKMNWADAERFCSEQAKGGHLVSIKIYSKEKDFVGDLVTKNIQSSDLYAWIGLRVENKEKQCSSEWSDGSSVSYENVVERTVKKCFALEKDLGFVLWINLYCAQKNPFVCKSPPP	1167	266,1927
2129	1gmz	>>DLWQFGKMILKETGKLPFPYYVTYGCYCGVGGRGGPKDATDRCCFVHDCCYGKLTSCKPKTDRYSYSRKDGTIVCGENDPCRKEICECDKAAAVCFRENLDTYNKKYMSYLKSLCKKXADDC	2857	651,6817
6485	1vhf	>>MSLILVYSTFPNEEKALEIGRKLLEKRLIACFNAFEIRSGYWWKGEIVQDKEWAAIFKTTEEKEKELYEELRKLHPYETPAIFTLKVENVLTEYMNWLRESVLEGGSHHHHHH	3	0,6843
15242	2y1l	>>MRGSHHHHHHGSDLGKKLLEAARAGRDDEVRILMANGADVNAEDASGWTPLHLAAFNGHLEIVEVLLKNGADVNAVDHAGMTPLRLAALFGHLEIVEVLLKNGADVNANDMEGHTPLHLAAMFGHLEIVEVLLKNGADVNAQDKFGKTAFDISIDNGNEDLAEILQKLN	28	6,3868
17051	3cwf	>>SNAETSDQRKAEEHIEKEAKYLASLLDAGNLNNQANEKIIKDAGGALDVSASVIDTDGKVLYGSNGRSADSQKVQALVSGHEGILSTTDNKLYYGLSLRSEGEKTGYVLLSASEKSDGLKGE	3503	799,0343
20069	3k51	>>RSVAETPTYPWRDAETGERLVCAQCPPGTFVQRPCRRDSPTTCGPCPPRHYTQFWNYLERCRYCNVLCGEREEEARACHATHNRACRCRTGFFAHAGFCLEHASCPPGAGVIAPGTPSQNTQCQPCPPGTFSASSSSSEQCQPHRNCTALGLALNVPGSSSHDTLCTSTGHHHHHH	7383	1684,062

21018	3m8e	>>MGSSHHHHHHSSGLVPRGSHMNRDHFY TLNIAEIAERIGNDDCAYQVLMAFINENGEAQ MLNKTAVAEMIQLSKPTVFATVNSFYCAGYID ETRVGRSKIYTLSDLGVEIVECFKQKAMEMR NL "	4521	1031,24
22507	3pqj	>>MTREDMEKRANEVANLLKTLSHPVRLMLV CTLVEGEFSVGELEQQIGIGQPTLSQQLGVL RESGIVETRRNIKQIFYRLTEAKAAQLVNALY TIFCAQEKQA	1646	375,452 6

Al contrastar los resultados relativos a los tiempos de comparación de secuencias de proteínas, se aprecia una mejora evidente, al pasar de 35,5 años en la comparación de una sola secuencia contra toda las secuencias almacenadas en el PDB, a un tiempo promedio de3,9457 minutos para el código Q, 4,3282 minutos para el código P y 8,5826 minutos para el código Binario, de acuerdo a los tiempos en las tablas 3, 5 y 7, al comparar la secuencia problema con las secuencias de mayor grado de similaridad. Agilizando de esta manera el proceso de notación de secuencias de proteínas.

CONCLUSIONES Y RECOMENDACIONES

3.1 CONCLUSIONES

- Se desarrolló un algoritmo que permite agilizar los procesos de notación de secuencias de proteínas, mediante la aplicación en conjunto de los algoritmos VCM y SOM, agrupando clusters de secuencias de proteínas codificadas en HCA con un alto grado de similaridad.

- Se comprobó la viabilidad de la aplicación de minería de datos a bases de datos consecuencias de proteínas, mediante la evaluación de los resultados obtenidos, con la metodología de comparación HCA.

- Se demostró que la aplicación de una correcta minería de datos a un gran volumen de secuencias de proteínas codificadas en HCA, permite reducir la cantidad de comparaciones a realizar, acortando el tiempo necesario para el proceso de anotación.

- Se evidenció que la utilización del algoritmo SOM para la clasificación de secuencias de proteínas codificadas en HCA, aumentó considerablemente la probabilidad de encontrar secuencias con alto porcentaje de similaridad, indicando que la propuesta realizada en este proyecto es viable y debería ser implementada.

3.2 RECOMENDACIONES

- Ampliar la base de datos de proteínas ingresando nuevas cadenas de secuencias como casos de estudio, de manera que facilite al investigador el acceso a un amplio repositorio de secuencias.

- Continuar con la línea de trabajo, puesto que los resultados obtenidos indican un gran potencial para el desarrollo de herramientas de comparaciones eficientes y mucho más específicas.

- La aplicación del algoritmo de agrupamiento implementado en otras bases de datos biológicas, validaría su utilidad en áreas de estudio afines.

- Se recomienda la implementación de una página web que provea acceso a la metodología de clasificación implementada, con sus correspondientes formularios de consulta, para el fin de simplificar su uso y maximizar su aprovechamiento.

- Se recomienda el desarrollo de herramientas complementarias, que permitan la realización de comparaciones aún más específicas.

BIBLIOGRAFIA

ABASCAL Federico, Análisis de genomas métodos para la predicción y anotación de la función de las proteínas", Centro Nacional de Biotecnología, Madrid 2003.

ALAIN J Cozzone, Proteins: Fundamental Chemical Properties, France 2002.

BEDREGAL Carlos Eduardo, Agrupamiento de Datos utilizando técnicas MAM-SOM, UCSP Universidad Católica San pablo Julio de 2008

BRUNGGER A y LAB Head, Algorithms for sequence Alignments. Lecture notes, 2003

CALLEBAUT I,LABESSE P, DURAND P, POUPON A,CANARD L, CHOMILIER J, HENRISSAT B & MORNON J P, Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives, France 1997

CARUGO Oliviero**,** Rapid Methods for Comparing Protein Structures and Scanning Structure, Viena Austria 2006

DOOLITLE R.F, Similar amino acid sequences: chance or common ancestry, 1981

EUDES R, LE TK, DELETTRÉ J, MORNON JP, CALLEBAUT I., A generalized analysis of hydrophobic and loop clusters within globular protein sequences BMC structural Biology , 2007.

FAYYAD U M, PIATETSKY G, & SMYTH P, From Data Mining To Knowledge Discovery in Databases, 2012

GABORIAUD C, BISSERY V, BENCHETRIT T & MORNON JP, Hydrophobic cluster analysis: an efficient new way to compare and analyse amino acid sequences. France 1987

GALLARDO Margarita, aplicación de técnicas de Clustering para la mejora del aprendizaje, Madrid 2009

GARRIDO Natalia, Biología Estructural de Proteínas, 2002

GUILLEN Victoria, Estructura y Propiedades de las Proteínas, España 2009.

HANEKAMP Theodor, protein sequence Alignments University of Wyoming.MOLB5650.Spring, 2002

HULLE Marc, Self-Organizing Maps, 2001

KOHONEN T, Self-Organizing Maps Series in Information Sciences, 1995

KRAIWATTANAPONG J, OOI T, KINOSHITA S, SUGIMURA I, & SAWABE T "Hydrophobic cluster analysis and classification of sixteen bacterial alginate lyases", World Journal of Microbiology & Biotechnology 16: 219-224, 2000

LOCKEHEIDE Sandra, segmentación de los contribuyentes que declaran iva utilizando técnicas de DataMining 2007

MAIZEL J V, & LENK R P, Enhanced graphic matrix analysis of nucleic acid and protein sequences, USA 1981

PEARSON W. R & LIPMAN D .J, Improved tools for biological sequence comparison, USA 1988

RAMÍREZ Alberto, Metodología de la Investigación Científica," Pontificia Universidad Javeriana, Colombia, 2004

RIMELQUE José, RUIZ Roberto & GIBERT Karina, minería de datos conceptos y tendencias, España 2006

RUAN J, WANG K, YANG J, KURGAN L.A. & CIOS K.J, accurate and consistent method for prediction of helix and strand content from primary protein sequences, 2005

SELLERS Peter, on the theory and computation of evolutionary distances SIAM, New York 1974

SILVA P J, assessing the reliability of sequence similarities detected through hydrophobic cluster analysis, Porto Portugal 2007

TATUSOVA Tatiana A & MADDEN Thomas L, BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences, USA 2011

TORRES Miguel, PAEZ Cárdenas, MARTINEZ Alicia, RODRIGUEZ Enrique, "Algoritmos para la búsqueda de texto", Colombia

ZEYAR Aung & KIAN -Lee Tan, Rapid retrieval of protein structures and scanning structure databases", 2007

PDB Banco de datos de proteínas 2013 Disponible on line: http://www.rcsb.org/pdb/home/home.do fecha de consulta octubre 2012.

GENBANK Banco de datos de proteínas 2013 Disponible on line: http://www.ncbi.nlm.nih.gov/genbank/ fecha de consulta octubre 2012.

Minería de datos Disponible on line: http://dis.unal.edu.co/profesores/jgomez/courses/data_mining/Mineria.pdf,

Consultado Noviembre de 2012.

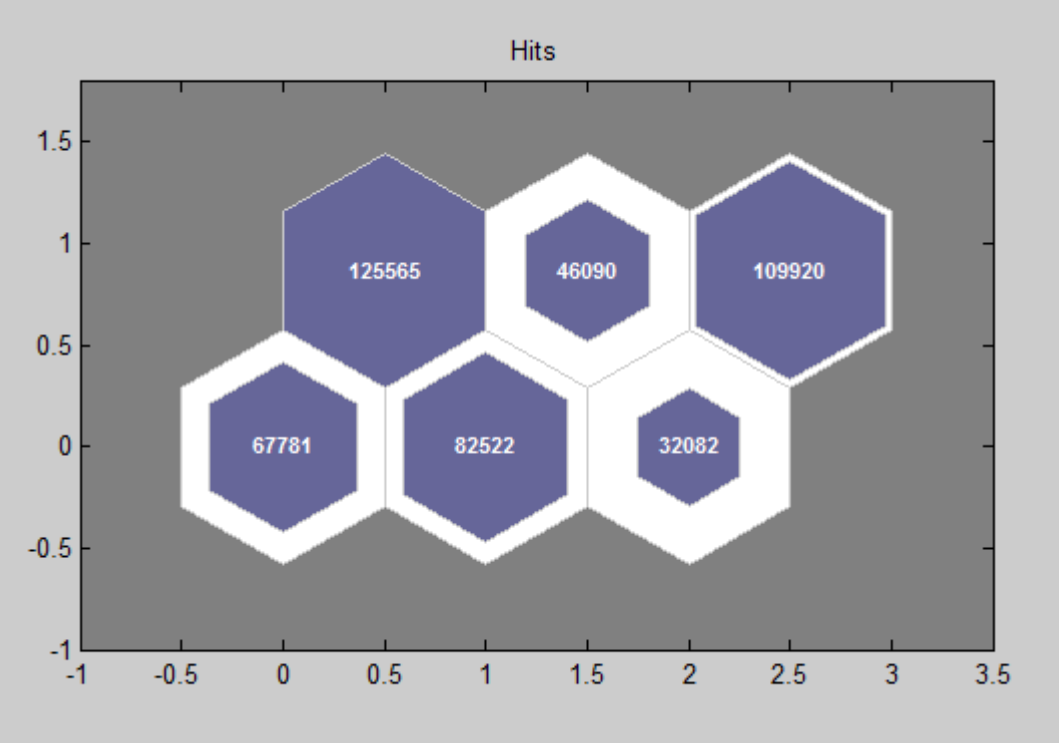
ANEXOS

ANEXO A. LISTADO DE AMINOÁCIDOS ESENCIALES

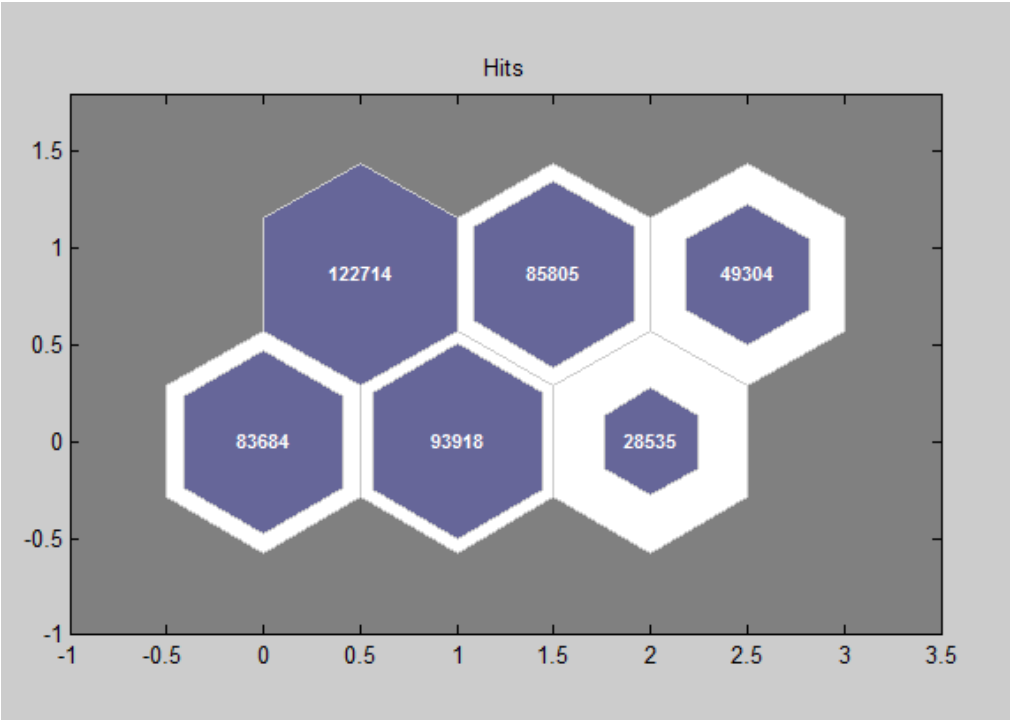
NOMBRE	ABRV	SÍMBOLO
Alanina	Ala	A
Arginina	Arg	R
Asparagina	Asn	N
Aspártico	Asp	D
Cisteina	Cys	C
Fenilalanina	Phe	F
Glicina	Gly	G
Glutámico	Glu	E
Glutamina	Gln	Q
Histidina	His	H
Isoleucina	Ile	I
Leucina	Leu	L
Lisina	Lys	K
Metionina	Met	M
Prolina	Pro	P
Tirosina	Tyr	Y
Treonina	Thr	T
Triptófano	Trp	W
Serina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

ANEXO B. DISTRIBUCIÓN DE LOS CLUSTERS COMO RESULTADO DE LA RED SOM

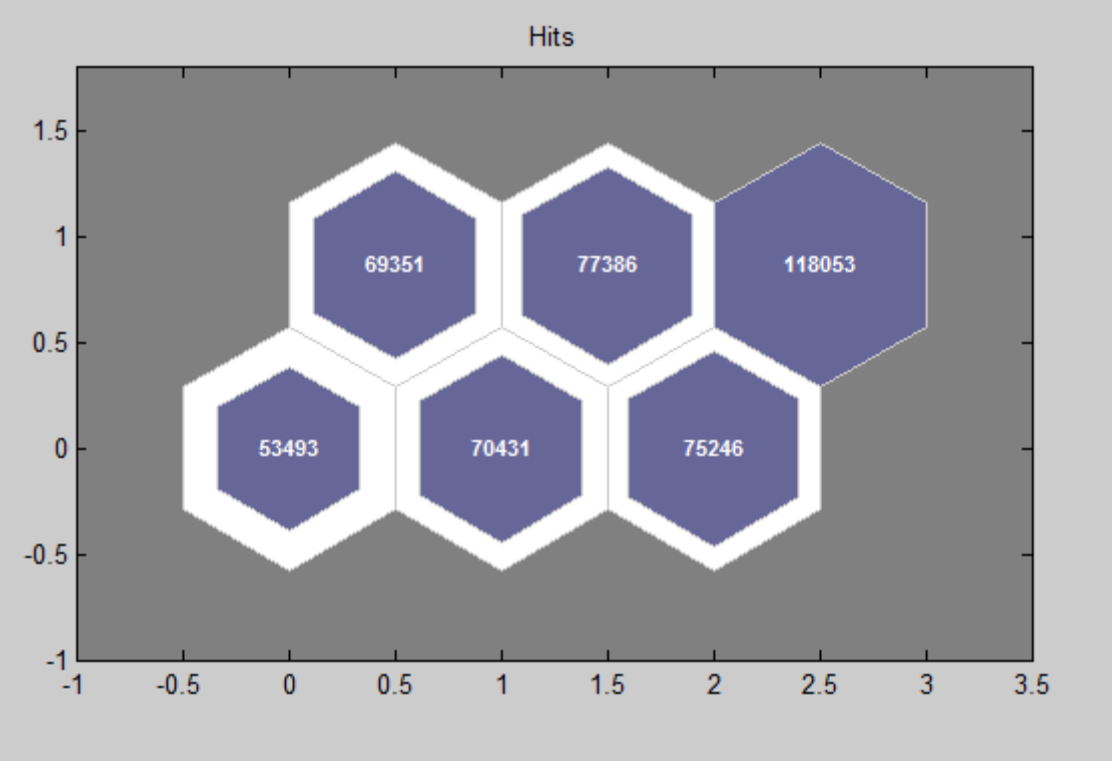
SOM Sample Hits Código Q



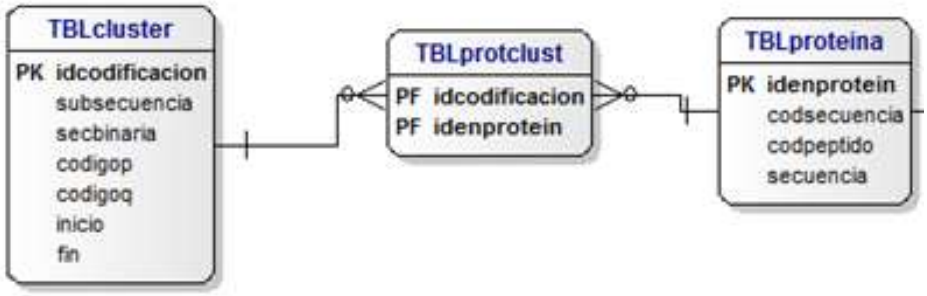
SOM Sample Hits Código Binario



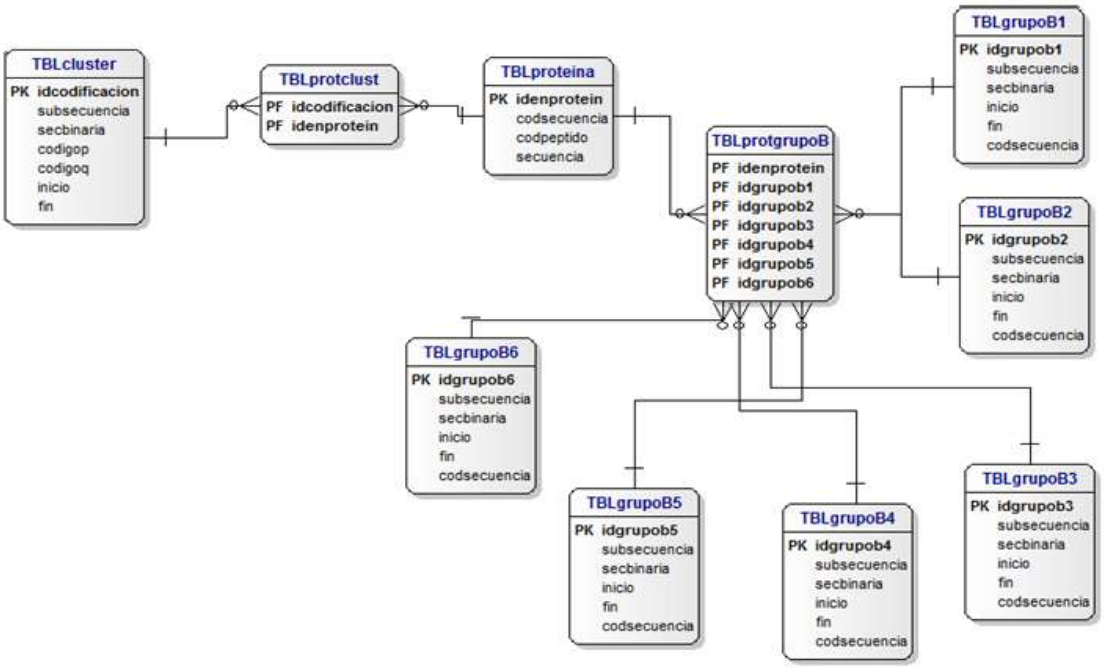
SOM Sample Hits Código P



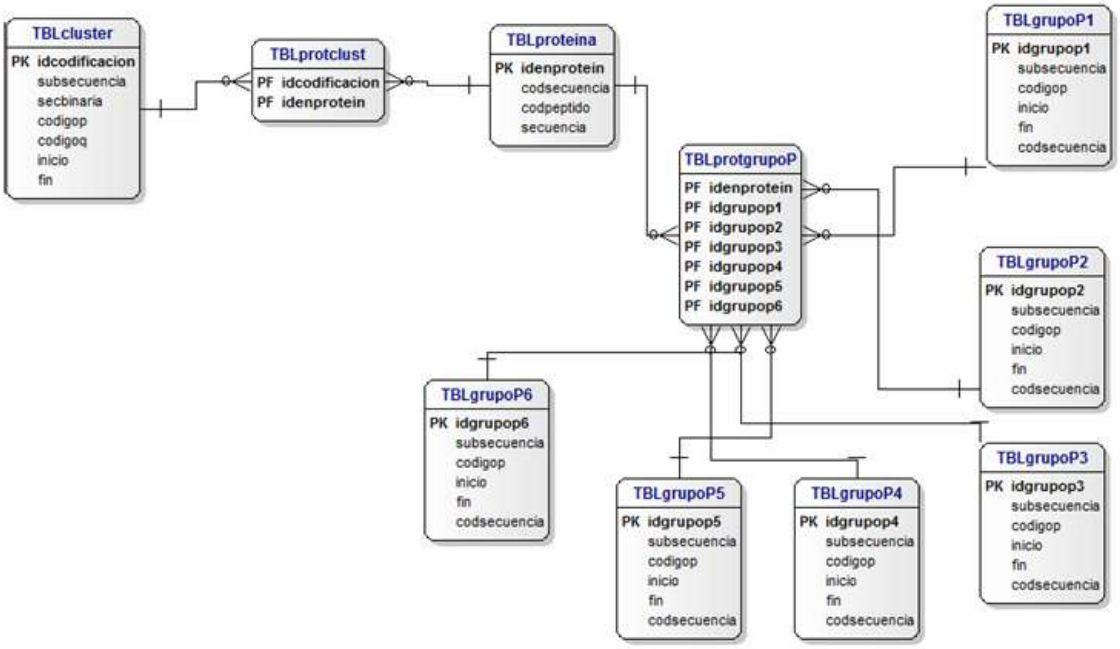
ANEXO C. ESTRUCTURA PDB ORIGINAL



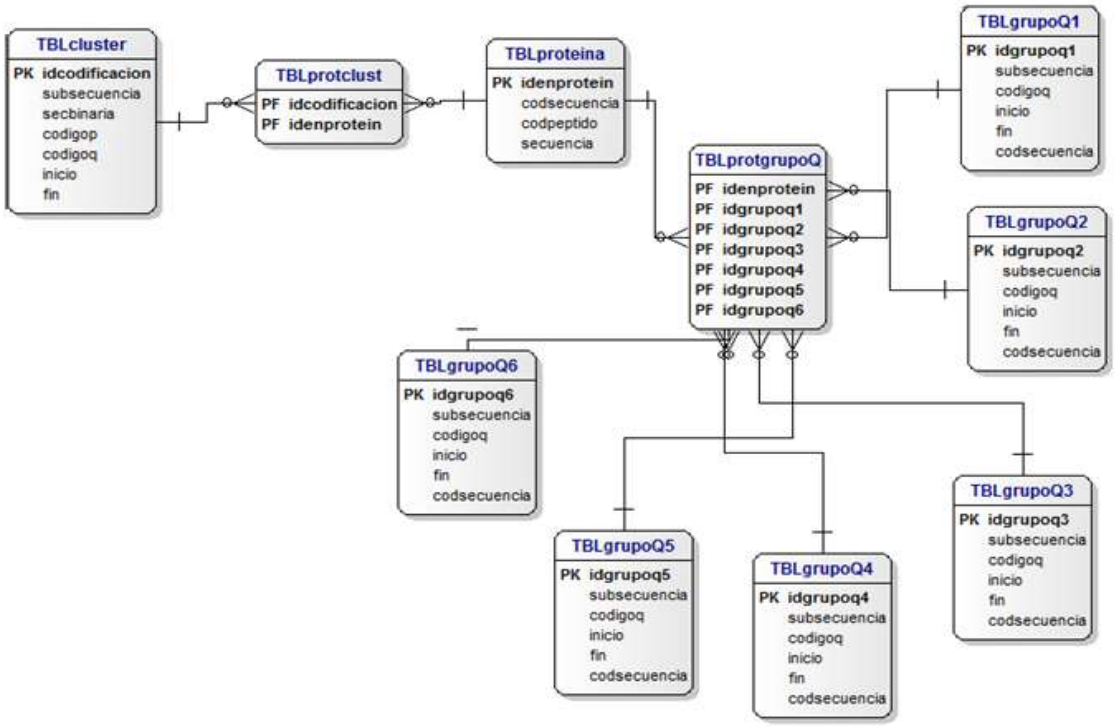
REESTRUCTURA DEL PDB PARA CODIGO BINARIO



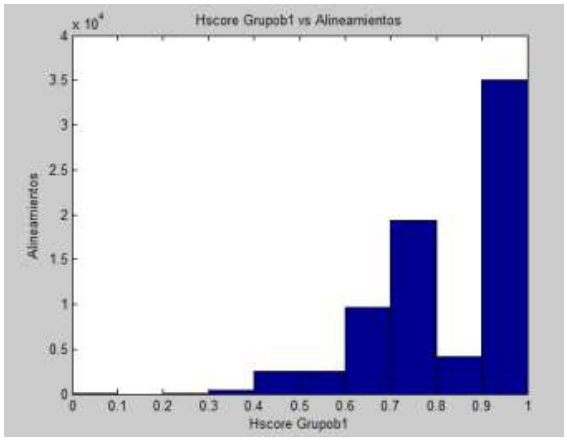
REESTRUCTURA DEL PDB PARA CODIGO P

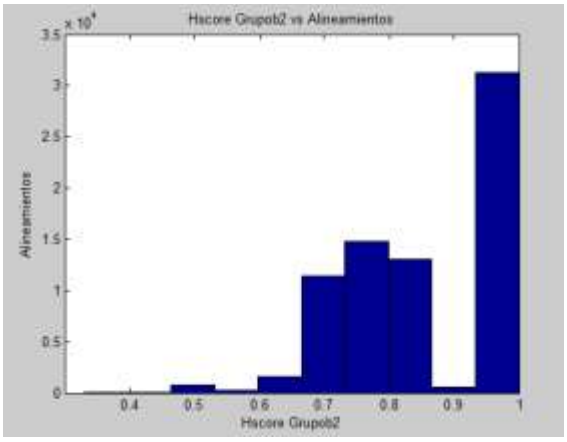


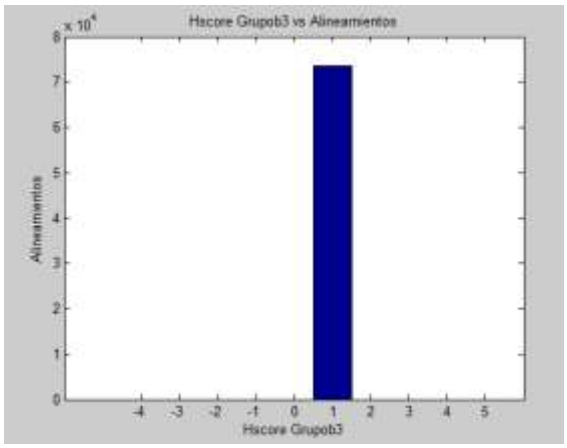
REESTRUCTURA DEL PDB PARA CODIGO Q

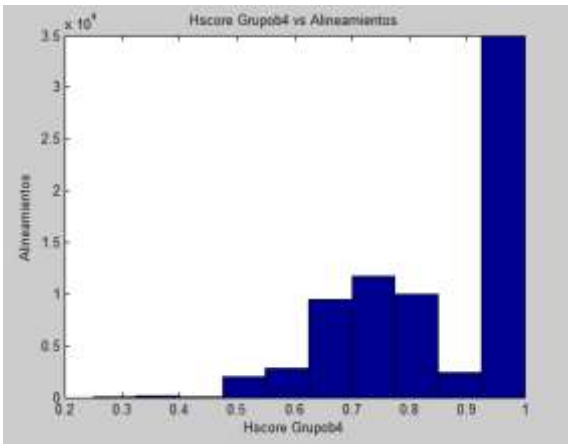


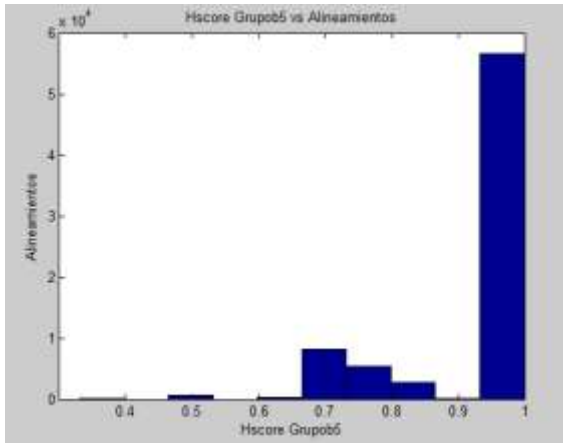
ANEXO D. PORCENTAJE HSCORE GRUPOS CODIFICACIÓN BINARIA

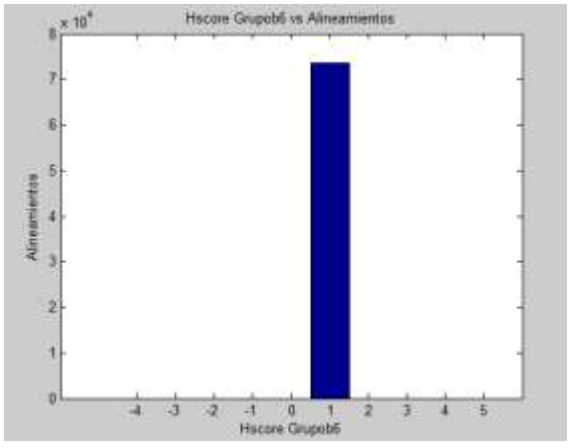




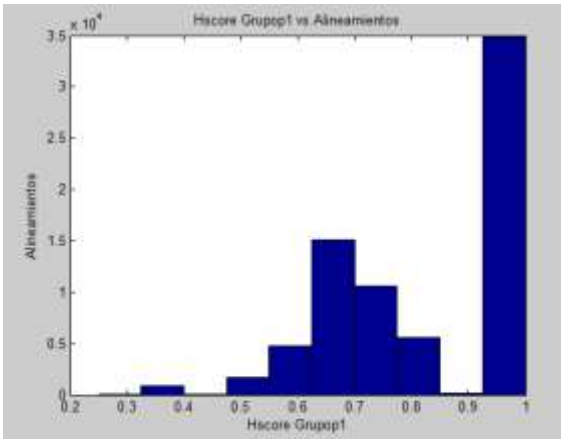


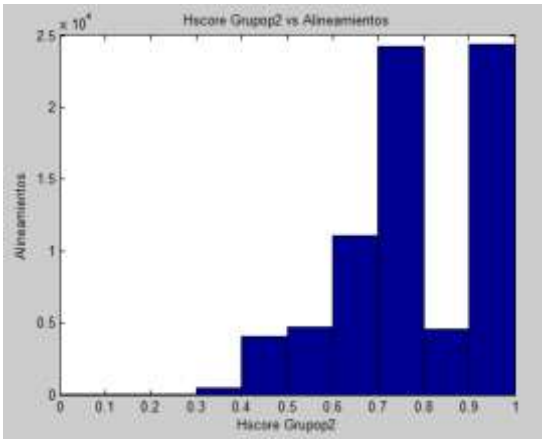


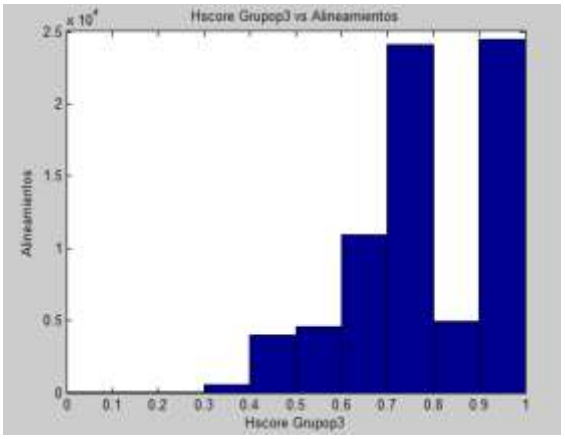


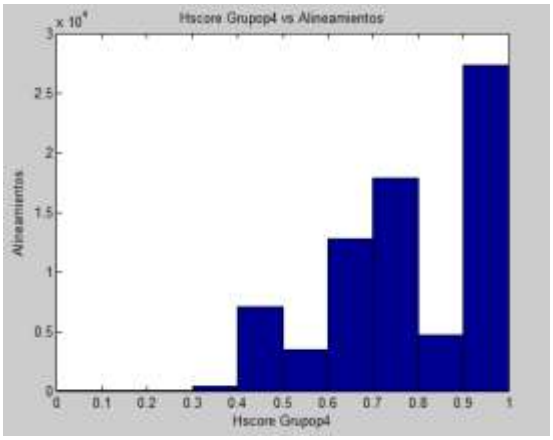


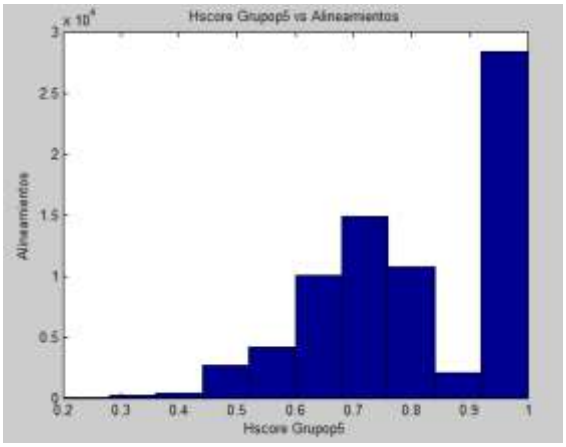
PORCENTAJES HSCORE GRUPOS CODIFICACION P

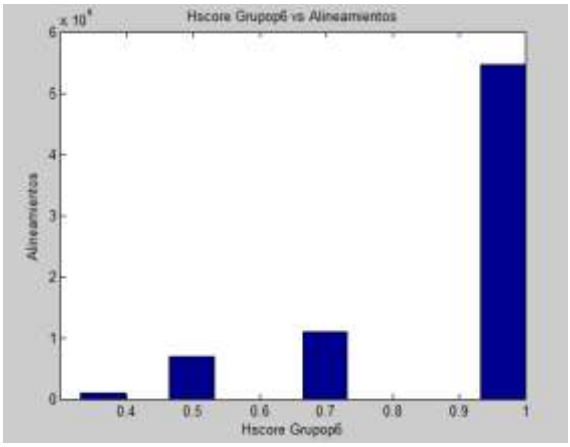




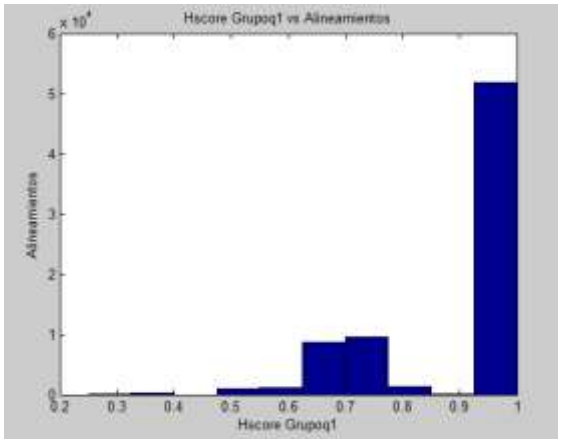


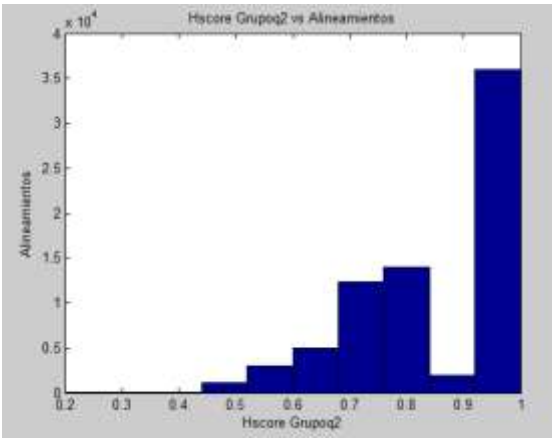


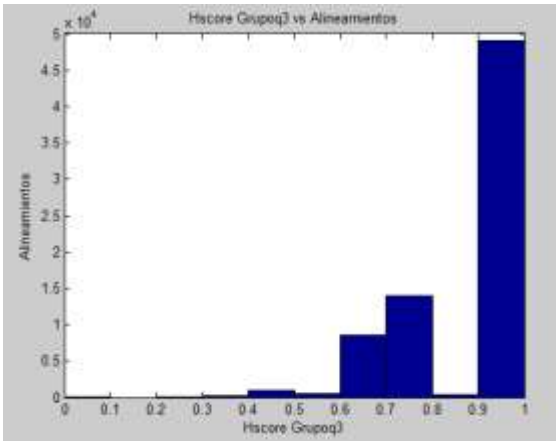


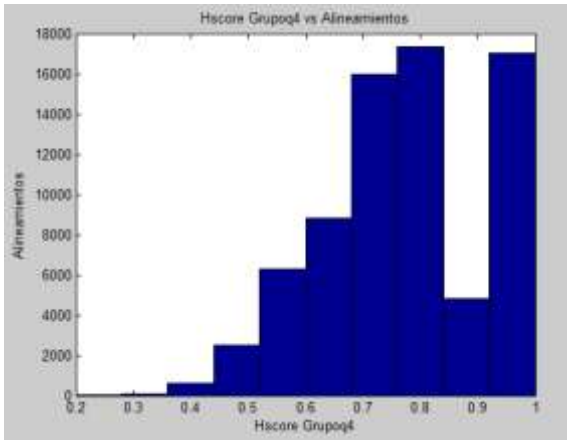


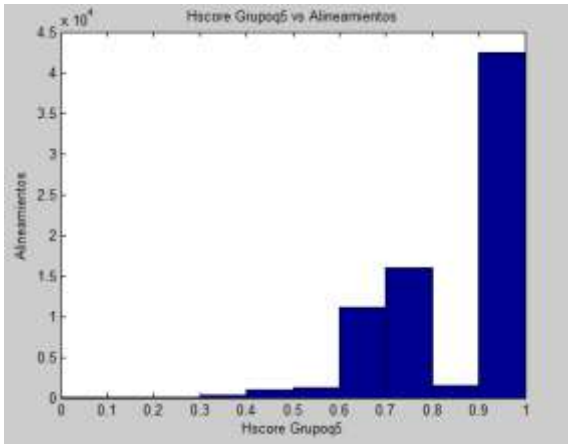
PORCENTAJES HSCORE GRUPOS CODIFICACION Q

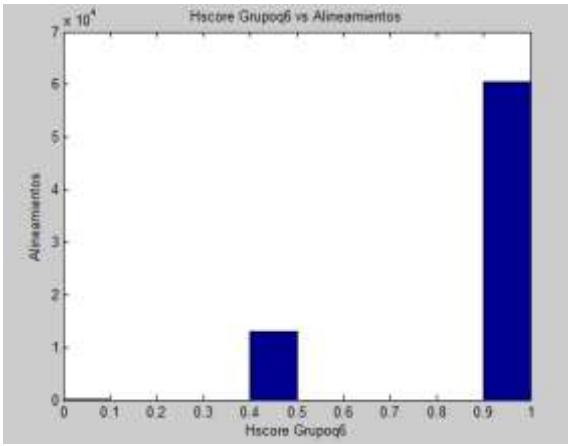

Hscore Grupoq1 vs Alineamientos
x 10^4
Alineamientos
Hscore Grupoq1


Hscore Grupoq2 vs Alineamientos
x 10^4
Alineamientos
Hscore Grupoq2


Hscore Grupoq3 vs Alineamientos
x 10^4
Alineamientos
Hscore Grupoq3


Hscore Grupoq4 vs Alineamientos
Alineamientos
Hscore Grupoq4


Hscore Grupoq5 vs Alineamientos
x 10^4
Alineamientos
Hscore Grupoq5


Hscore Grupoq6 vs Alineamientos
x 10^4
Alineamientos
Hscore Grupoq6