Metodología y algoritmo para la construcción de una matriz de sustitución generalizada para alfabetos arbitrarios que describen secuencias protéicas

YERMINSON DONEY GONZALEZ MUÑOZ

200843846

yerminson.gonzalez@correounivalle.com

Irene Tischer, Ph.D

irene.tischer@correounivalle.edu.co

Facultad de Ingeniería

Escuela de Ingeniería de Sistemas y Computación

Programa Académico de Ingeniería de Sistemas

Santiago de Cali, Noviembre 30 de 2014

Índice

Resumen

En este documento se presenta una propuesta para desarrollar e implementar una metodología para la construcción de una matriz de sustitución generalizada para alfabetos arbitrarios que describen secuencias protéicas. El desarrollo de esta propuesta se realiza mediante el seguimiento de un conjunto de pasos que van desde el entendimiento conceptual del problema, el desarrollo de la metodología y la implementación de los algoritmos correspondientes que permitan la construcción de una matriz de sustitución generalizada, que se utilizará en un algoritmo de alineamiento que sirva para probar algunos casos de estudio.

1. INTRODUCCIÓN

Las matrices de sustitución han sido una herramienta muy importante en el alineamiento de secuencias ya que permiten obtener mejores alineamientos debido a la información que contienen. Las matrices de sustitución más interesantes son las relacionadas con proteínas y se crearon a partir de diferentes modelos como el de Dayhoff de donde nacen las matrices PAM y el modelo de Henikoff and Henikoff de donde nacen las matrices BLOSUM. Básicamente nos muestran un matriz donde se puntúa que tan bueno es sustituir un aminoácido por otro.

2. OBJETIVOS

2.1. Objetivo General

Desarrollar una metodología que permita construir una matriz de sustitución generalizada para alfabetos arbitrarios que describen secuencias biológicas e implementar los algoritmos subyacentes.

2.2. Objetivos Específicos

1. Estudiar las matrices de sustitución usadas actualmente para establecer los elementos básicos en la construcción de una matriz de sustitución que permitan su generalización.
2. Diseñar una metodología para la construcción de una matriz de sustitución generalizada, así como los algoritmos que permitan su construcción utilizando alfabetos arbitrarios.
3. Desarrollar un algoritmo para el alineamiento de secuencias en alfabetos arbitrarios utilizando matrices de sustitución generalizadas en su desarrollo.
4. Analizar la metodología y los resultados de los algoritmos desarrollados mediante su aplicación en diferentes casos de estudio.

3. MARCO TEÓRICO

3.1. Nucleótidos

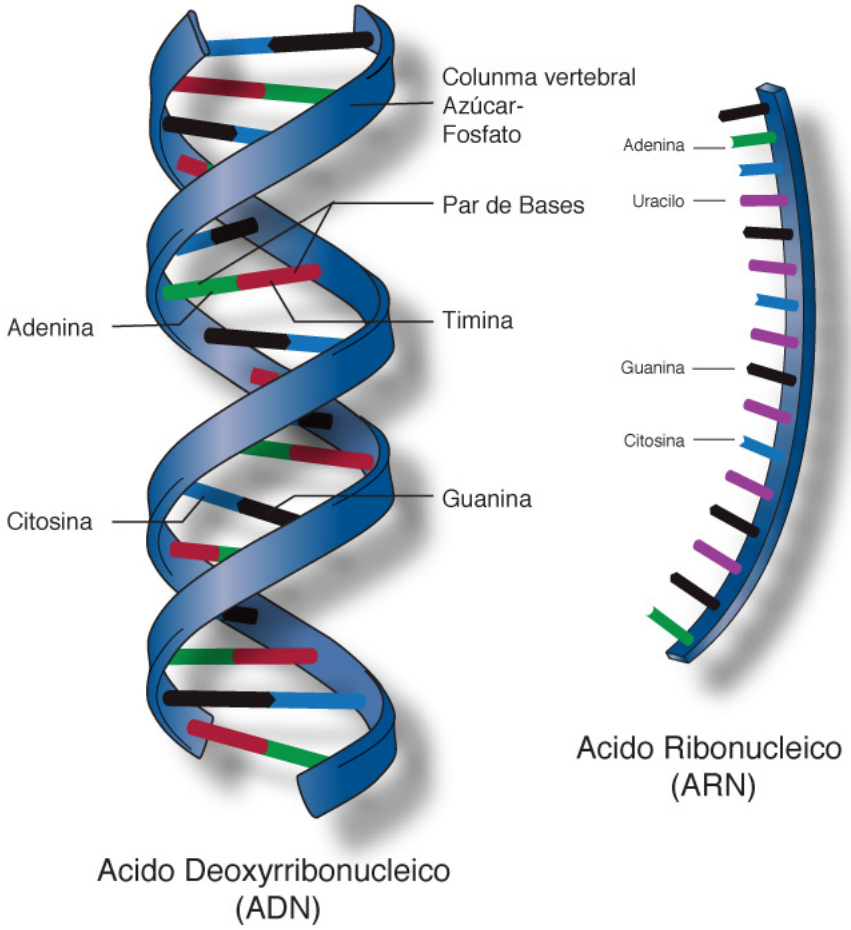
Los genes por sí mismos contienen su información como una secuencia específica de nucleótidos que son encontrados en el ADN de las moléculas. Solo cuatro diferentes bases son usadas en las moléculas de ADN: guanina (G), adenina (A), timina (T) y citosina (C). En el ARN, la base nitrogenada uracilo (U) ocupa el lugar de la timina. Cada base está unida a un grupo fosfato y a una azúcar desoxirribosa en el caso del ADN o ribosa en el caso del ARN para formar un nucleótido. El elemento que nos permite diferenciar un nucleótido de otro es la base nitrogenada que contiene. Al representar la estructura de cada una de las bases nitrogenadas podemos encontrar diferencias muy marcadas con lo que puede distinguir fácilmente entre un nucleótido y otro. [8]

Un nucleótido es la pieza básica de los ácidos nucleicos. El ADN y el ARN son polímeros formados por largas cadenas de nucleótidos. Un nucleótido esta formado por una molécula de azúcar (desoxirribosa en el ADN y ribosa en el ARN) unido a un grupo fosfato y una base nitrogenada. Las bases nitrogenadas se pueden agrupar en dos grandes grupos purinas formado por adenina y guanina y pirimidinas formado por guanina, timina y uracilo.

3.2. Ácidos nucleicos

Los ácidos nucleicos son un tipo importante de macromoléculas presentes en todas las células y virus. Las funciones de los ácidos nucleicos tienen que ver con el almacenamiento y la expresión de información genética. El ácido desoxirribonucleico (ADN) codifica la información que la célula necesita para fabricar proteínas. Un tipo de ácido nucleico relacionado con él, llamado ácido ribonucleico (ARN), presenta diversas formas moleculares y participa en la síntesis de proteínas.

La pentosa es diferente en cada uno de los ácidos nucleicos : Desoxirribosa en el ADN y Ribosa en el ARN , las bases nitrogenadas son similares pero en el ARN se cambia timina por uracilo y la estructura de las cadenas es diferente el ADN forma una doble hélice mientras el ARN forma solo una cadena.



3.3. Aminoácidos

Aminoácidos son los bloques de construcción de las proteínas. Como el ADN y ARN, proteínas son sintetizadas como polímeros lineales(cadenas) compuestas de pequeñas moléculas. A diferencia del ADN y ARN, en los cuales hay cuatro nucleótidos de los cuales escoger, proteínas son construidas de 20 aminoácidos con una variedad de tamaños , formas y propiedades químicas.

Cada aminoácido está compuesto esencialmente por un grupo amino(-NH2) , un carbono alfa, grupo carboxilo(-COOH) y una cadena (habitualmente denominada cadena lateral o radical R) de estructura variable que determina la identidad y las propiedades de cada uno de los diferentes aminoácidos. Existen cientos de radicales por lo que se conocen cientos de radicales pro lo que se conocen cientos de aminoácidos pero solo 20 forman parte de las proteínas y tienen codones específicos en el código genético.

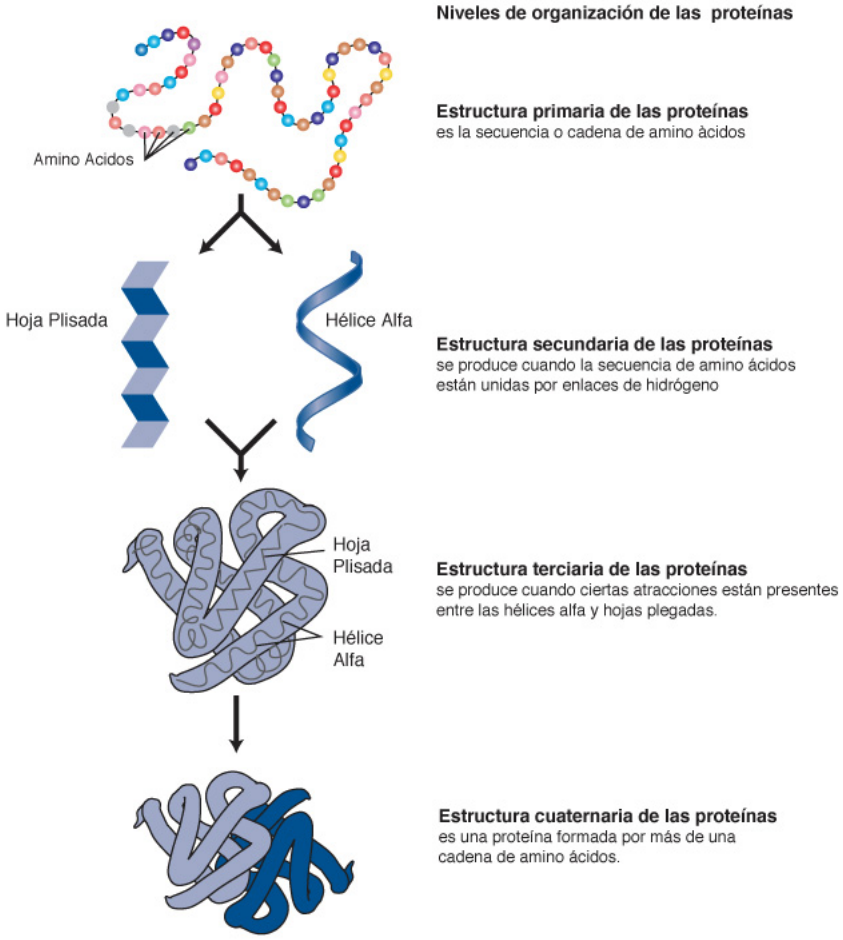
Los aminoácidos están con frecuencia agrupados en tres categorías. Los aminoácidos hidrofóbicos, los cuales tienen cadenas laterales compuestas principalmente o enteramente de carbono e hidrógeno, tienen pocas probabilidades de formar enlaces de hidrógeno con las moléculas de agua. Los aminoácidos polares, los cuales con frecuencia contienen oxígeno y/o nitrógeno en sus cadenas laterales, forman enlaces de hidrógeno con agua de una manera muy fácil. Finalmente, los aminoácidos cargados llevan una carga positiva o negativa en el pH biológico. [8]

3.4. Proteínas

Las proteínas son una clase importante de moléculas que se encuentran en todas las células vivas. Una proteína se compone de una o mas cadenas largas de aminoácidos cuya secuencia corresponde a la secuencia de ADN del gen que la codifica. Las proteínas desempeñan gran variedad de funciones en la célula, incluidas estructurales (citoesqueleto), mecánicas (musculo), bioquímicas (enzimas), y de señalización celular (hormonas). Las proteínas son también parte esencial de la dieta.

Las proteínas desempeñan un papel fundamental para la vida y son las biomoleculas mas versátiles y diversas, imprescindibles en el crecimiento del organismo.

Las proteínas pueden ser estudiadas de acuerdo a su nivel de organización empezando con la estructura primera la cual es la secuencia o cadena de aminoácidos que describe la proteína, la estructura secundaria que se produce cuando la secuencia de aminoácidos están unidas por enlaces de hidrógeno formando hélices alfa y hojas plegadas beta, la estructura terciaria que se produce cuando ciertas atracciones están presentes entre las hélices alfa y las hojas plegadas beta y finalmente la estructura cuaternaria que es una proteína formada por mas de una cadena de aminoácidos.



3.5. Matrices de Sustitución

Una matriz de sustitución o puntuación describe el ritmo al que un carácter en una secuencia cambia a otro carácter con el tiempo. Las matrices de sustitución se ven usualmente en el contexto de alineamiento de secuencias de aminoácidos o ADN, donde la similitud entre las secuencias depende del tiempo desde su divergencia y de los ritmos de sustitución según se presenta en la matriz.

Este tipo de matrices son mas usuales en los alineamientos de secuencias de aminoácidos (proteínas) que en los de nucleótidos ya que en las matrices de sustitución para secuencias de nucleótidos son relativamente simples. Un valor positivo o una puntuación alta es dada para un acierto y un valor negativo o puntuación baja es dada para un fallo, para el caso de los gaps también se manejaría una puntuación negativa.

Las matrices de sustitución mas utilizadas son las matrices PAM y las BLOSUM. Ambas son matrices de log-probabilidades que se diferencian principalmente, en los fundamentos para el calculo inicial de las probabilidades de sustitución entre aminoácidos.

3.6. PAM

Las matrices PAM fueron las primeras construidas , quienes recogían alineamientos de 71 grupos de secuencias de proteínas estrechamente relacionadas. El nombre PAM viene de "mutación puntual aceptada", porque usa homólogos muy relacionados, las mutaciones observadas no se espera que cambie significativamente la función común de las proteínas. Es decir las sustituciones de aminoácidos son consideradas como aceptadas por la selección natural.

El valor de una determinada celda representa la probabilidad de la sustitución de un aminoácido por otro, conocida como mutación puntual. Puesto que la matriz se calcula observando diferencias entre proteínas muy cercanas evolutivamente (con, al menos, un 85 % de similitud), las sustituciones en cuestión no tienen efecto sobre la función de la proteína, por lo que se trata de mutaciones aceptadas en el proceso evolutivo. Estas secuencias de proteínas fueron agrupadas basadas en reconstrucción filogenética usando "maximum parsimony". Las matrices PAM fueron después derivadas basadas en la divergencia evolutiva entre las secuencias del mismo grupo. Una unidad (PAM 1) es definida como 1 % de las posiciones de aminoácidos que se han cambiado. Para construir una tabla de sustitución PAM 1, un grupo de secuencias estrechamente relacionadas con frecuencias de mutación correspondientes a una unidad PAM son escogidas. Basados en la información mutacional para este grupo de secuencias, una matriz de sustitución puede ser derivada a partir de esta matriz (PAM 1) elevandola a una determinada potencia por lo que la matriz (PAM 250) resultaria elevando a la 250 potencia la matriz (PAM 1). [6]

3.7. BLOSUM

BLOSUM (matriz de sustitución de bloques de aminoácidos) es una matriz de sustitución utilizada para el alineamiento de secuencias de proteínas. BLOSUM se usa para puntuar alineamientos entre secuencias de proteínas evolutivamente divergentes. Se basa en alineamientos locales, y se introdujo en 1992 por primera vez en un artículo de S. Henikoff y J. G. Henikoff. Recorrieron la base de datos BLOCKS analizando regiones muy conservadas de familias de proteínas (sin gaps en el alineamiento de secuencias) y comprobaron las frecuencias relativas de aparición de los aminoácidos y las probabilidades de sustitución entre ellos. Seguidamente calcularon una puntuación de log-probabilidad para cada una de las 210 posibles sustituciones de los 20 aminoácidos estándar.

Todas las BLOSUM se basan en alineamientos observados, y no son extrapoladas de comparaciones de proteínas cercanamente relacionadas. Existen bastantes conjuntos de matrices BLOSUM que utilizan diferentes bases de datos de alineamientos, y que se nombran con números. Las BLOSUM seguidas de un número alto están diseñadas para comparar secuencias cercanamente relacionadas, mientras que las BLOSUM con número bajo están diseñadas para comparar secuencias relacionadas de forma distante.

Por ejemplo, BLOSUM 80 se usa para alineamientos menos divergentes, mientras que BLOSUM 45 se usa para alineamientos más divergentes. A cada posible identidad o sustitución se le asigna una puntuación basada en las frecuencias observadas en el alineamiento de proteínas relacionadas. Se da una puntuación positiva a las sustituciones más probables, mientras que corresponde una puntuación negativa para sustituciones menos probables. El numero que acompaña a la matriz BLOSUM indica el porcentaje de similitud que existe entre las secuencias que se tomaron como base para la construcción de la matriz por tal motivo entre mayor sea el numero mas relación hay entre las secuencias que fueron consideradas para la construcción de la matriz.

BLOSUM 62 es la matriz calculada usando las sustituciones observadas entre proteínas que tienen, como mínimo, el 62 % de identidad en la secuencia, y se ha convertido en el estándar de la mayoría de los programas que utilizan este tipo de matrices.[7]

3.8. Alineamiento de Secuencias

La comparación de secuencias se encuentra en el corazón del análisis bioinformatico. Esto es un importante primer paso hacia el análisis estructural y funcional de secuencias recién determinadas. Como las nuevas secuencias biológicas están siendo generadas en proporciones exponenciales, la comparación de secuencias es cada vez más importante para extraer inferencias funcionales y evolutivas de una nueva proteína con proteínas ya existentes en las bases de datos. El principal proceso en este tipo de comparación es el alineamiento de secuencias. Este es el proceso por el cual secuencias se comparan mediante la búsqueda de patrones de caracteres comunes y se establecen r-r correspondencia entre secuencias relacionadas. Alineamiento de pares de secuencias es el proceso de alinear dos secuencias y es la base de búsqueda de similitud en bases de datos y alineación múltiple de secuencias.

El objetivo general del alineamiento de pares de secuencias es encontrar el mejor emparejamiento de dos secuencias, tal que hay máxima correspondencia entre residuos. Para lograr esta meta , una secuencia necesita ser desplazada respecto a la otra para encontrar la posición donde se encuentran las coincidencias máximas. En alineación global, dos secuencias a ser alineados se supone que las dos secuencias son generalmente similares en toda su longitud. La alineación se lleva a cabo de principio a fin de ambas secuencias para encontrar el mejor alineamiento posible a través de la longitud total entre las dos secuencias. Este método es más aplicable para la alineación de dos secuencias estrechamente relacionadas de más o menos la misma longitud. El alineamiento global tiene problemas con secuencias divergentes y secuencias de longitudes variables ya que falla para reconocer regiones locales muy similares entre las dos secuencias por lo que no se obtienen resultados óptimos. La alineación local, por otra parte, no asume que las dos secuencias en cuestión tienen similitud en toda la longitud. Esto sólo encuentra regiones locales con el más alto nivel de similitud entre las dos secuencias y alinea estas regiones sin considerar para la alineación el resto de las regiones de la secuencia. Este enfoque se puede utilizar para la alineación de secuencias más divergentes con el objetivo de buscar patrones de ADN conservados o secuencias de proteínas. Las dos secuencias a ser alineadas pueden ser de diferentes longitudes. Este enfoque es más apropiado para alinear secuencias biológicas divergentes que contienen sólo los módulos que son similares, que se conocen como dominios o motivos. [6]

4. ESTADO DEL ARTE

Las matrices de sustitución son muy usadas actualmente en alineamiento de secuencias de nucleótidos o proteínas y es así como muchos estudios e investigaciones se centran principalmente en la manera que se trabajará con la matriz de sustitución y no con el alfabeto a partir del cual se construirá buscando obtener propiedades distintas que se tienen con el uso de distintos alfabetos. A continuación mostramos una breve descripción de los trabajos que más se asemejan al trabajo que se realizará.

4.1. Matrices de similitud de aminoácidos basadas en campos de fuerza

Se propone un método general para derivar matrices de sustitución de aminoácidos a partir de campos de fuerza de baja resolución. A diferencia de los métodos populares para la obtención de matrices de sustitución el método planteado no utiliza argumentos evolutivos, alineamiento de secuencias o estructuras. La estrategia utilizada fue mutar los residuos y recolectar la contribución al total de la puntuación/energía total. El promedio de estos valores para cada posición entre un conjunto de proteínas resulta en una matriz de sustitución.

En este trabajo se muestra como se puede obtener matrices de sustitución basadas en campos de fuerza que actual sobre estructuras de tres dimensiones. La matriz obtenida permitirá realizar comparaciones entre diferentes campos de fuerza sin importar la diferencia que pueda existir entre los diferentes campos de fuerza. Estas matrices de acuerdo a como son calculadas tienen menos relación con la evolución o alineamientos de secuencias ya que están fundamentadas en propiedades de los residuos como es el caso de los campos de fuerza.

Como parte de los resultados varios ejemplos de matrices de sustitución asimétricas han sido calculados a partir de campos de fuerza usando diferentes acercamientos y el rendimiento obtenido para cada una de las matrices fue comparado con respecto a las matrices de sustitución convencionales.

4.2. Matrices de sustitución de pliegue especifico para clasificación de proteínas

Se describe un método para organizar proteínas en una familia que permita reducir un orden de magnitud en el numero de parámetros utilizados para la construcción de una matriz de sustitución. La base usada como medida de similitud en las matrices de sustitución convencionales es el logaritmo de la razón de oportunidades (log odd ratio). Esta característica se adapto para poder crear un atributo de diferenciación que conllevo a la definición de las matrices de sustitución CLASSUM(Class Attribute Substitution Matrices) con características similares a la matriz de sustitución BLOSUM.

El método utilizado fue aplicado para clasificar secuencias jerárquicamente en subgrupos lambda y kappa de la superfamilia de las inmunoglobulinas. Las posiciones que confieren una clase fueron identificadas basadas en el grado de variabilidad de un aminoácido en una posición. Donde las matrices CLASSUM tienen mejores resultados en la clasificación comparados a los resultados de la matriz de sustitucion BLOSUM 62.

Los resultados obtenidos sugieren que matrices de sustitución derivadas de información especifica de una familia pueden mejorar la resolución de métodos automáticos que usan matrices de sustitución para búsqueda y clasificación de proteínas. Con lo que se pueden clasificar proteínas dentro de una familia usando menos parámetros que los que serian utilizados por métodos de clasificación como redes neuronales , modelos ocultos de Markov o matrices de puntuación de posición especifica.

4.3. Construcción de matrices de sustitución de aminoácidos para comparación de proteínas con composiciones no estándar

Las matrices de sustitución de aminoácidos juegan papel importante en los métodos de alineamientos de proteínas. Las matrices estándar de log-odd probabilidades como la PAM y BLOSUM son construidas a partir de largos conjuntos de alineamientos teniendo implícitamente antecedentes de las frecuencias de los aminoácidos.

Estas matrices de sustitución han sido usadas para comparar proteínas con diferencias muy marcadas en la composición de los aminoácidos como lo son las proteínas de transmembrana o proteínas de organismos con composiciones de nucleótidos fuertemente sesgadas. Las matrices convencionales no son ideales para este tipo de comparaciones por lo que se construyo una matriz que puede ser utilizada en un ambiente donde no hay una composición estándar obteniendo mejores resultados mediante el ajuste matemático de las matrices ya existentes.

El ajuste composicional de las matrices de sustitución de aminoácidos en general tiende a incrementar la significación estadística de los alineamientos de proteínas con composiciones no estándar y frecuentemente mejora la exactitud de estos alineamientos también.

4.4. Un alfabeto estructural de proteínas y su matriz de sustitución CLESUM

Al usar un modelo mixto para la distribución de 3 ángulos pseudo enlazados formados por átomos de carbono alfa de 4 residuos consecutivos, los estados de la estructura local son discretizados como 17 letras conformadas de un alfabeto estructural de proteínas. Se construye una matriz de sustitución usando estas letras y basados en los alineamientos estructurales de la base de datos FSSP, que es una base de datos de estructuras alineadas mediante un proceso exhaustivo de todos contra todos comparando la estructura 3D de las proteínas en el PDB(Protein Data Bank).

La mayoría de métodos de predicción de estructura local usan tres estados de la estructura secundaria : helices , hebras y bucles. Sin embargo una estructura secundaria podría variar significativamente en sus estructuras 3D. Al restringir la conformaciones de los residuos locales a estados manejables se puede discretizar la conformación de una proteína para convertir la estructura 3D de la cadena principal a una secuencia 1D con los estados discretos que describen la estructura de la proteína parecidos a los aminoácidos.

La matriz de sustitucion construida se derivo de la misma manera que la matriz BLOSUM cambiando cierta informacion con la intencion de mostrar mas detalles por ejemplo el factor de escala utilizado fue de 20 en lugar de 2 la matriz construida y utilizada para comparar secuencias teniendo en cuenta su alfabeto estructural se denomino CLESUM.

4.5. Una matriz de sustitución para reconocimiento de pliegue de la proteína que incluye predicción de la estructura secundaria de una secuencia

En el reconocimiento de pliegue de proteínas, un secuencia de aminoácidos de prueba es comparado con una biblioteca de pliegues representativos de estructura conocida para identificar un homologo estructural. En caso donde la secuencia de aminoácidos de prueba y su homologo tienen una clara

similitud como secuencias las matrices de sustitución tradicionales han sido usadas para predecir la similitud estructural.

En el caso que la prueba es secuencialmente distante de su homologo se ha desarrollado una matriz de sustitución de 5 dimensiones (7 x 3 x 2 x 7 x 3) que lleva una representación 3D a 1D de una proteína llamada H3P2, calculada a partir de una base de datos de 119 pares estructurales.

Miembros de cara par estructural comparten un pliegue similar pero tienen una identidad de secuencia menor al 30 %. Cada posición de la secuencia de prueba es definida por uno de las 7 clases de residuo y 3 clases de estructura secundaria. Cada posición de un pliegue homologo es definida por 7 clases de residuo , 3 clases de estructura secundaria y 2 clases enterradas.

Para realizar la prueba de la matriz de sustitución obtenida se realizo una validación cruzada para comparar la matriz H3P2 con las matrices : GONNET , PAM 250, BLOSUM 62y una matriz de sustitución únicamente de estructura secundaria. Los resultados mostraron que para secuencias distantemente relacionadas la matriz H3P2 detecto mas estructuras homologas y mas alta confiabilidad que las otras matrices de sustitución basados en la relación sensibilidad contra especificidad.

4.6. Una matriz de sustitución para alfabetos estructurales basada en alineamientos estructurales de proteínas homologas y sus aplicaciones

Análisis de estructuras de proteínas basado en estructuras de motif cortas se han usado ampliamente y se han mostrado como algo útil en la predicción de la estructura dela proteína , reconstrucción de la cadena principal, descripción y predicción de pequeños lazos y largos fragmentos.

Usando una representación 1D de una una estructura 3D combinada con una matriz de sustitución y programación dinámica simple se logro localizar regiones de similitud estructural, resaltar un cambio sutil en las regiones de similitud estructural e identificar regiones donde no hay similitud estructural.

4.7. Matriz de sustitución de aminoácidos para la identificación de conformación de proteínas

Las matrices de sustitución juegan un rol esencial el alineamiento de secuencia de proteínas, una tarea fundamental en bioinformática. Las matrices mas utilizadas como PAM fueron derivadas se secuencias homologas y las matrices BLOSUM fueron derivadas de segmentos alineados en PROSITE por lo que no integraron información relacionada con la conformación de las proteínas en la construcción de las matrices.

Usando las bases de datos PDB_SELECT ,que contiene secuencias de aminoácidos, y DSSP ,que contiene estructuras secundarias de proteínas, se creo una base de bloques de conformación de estructura los cuales representa la relación secuencia estructura, los miembros en un bloque son idénticos en conformación y altamente similares en secuencia. A partir de este base de datos de bloques se derivo una matriz de sustitución de aminoácidos de conformación especifica llamada CBSM60. La matriz muestra un mejorado rendimiento en en la búsqueda de conformación de segmentos y en la detención de homólogos.

En la construcción del la propia base de datos bloque para la búsqueda de similitudes se inicio con la matriz de sustitución BLOSUM 62 y después de obtener el primer resultado de la matriz CBSM60 fue utilizado nuevamente para la creación de una nueva versión de la matriz CBSM60 y el proceso se repitió hasta que una convergencia final fue encontrada.

4.8. Calculo de matrices de sustitución para análisis comparativo genómico

Las matrices de sustitución describen el ritmo de mutación de un carácter en una secuencia biológica a otro carácter y son muy importantes para muchas tareas como análisis filogenético y alineamiento de secuencias. Se presenta un novedoso algoritmo para construir una matriz de sustitución de nucleótidos el método esta fundamentado en la teoría de la información básicamente el algoritmo iterativamente usa compresión y construye la matriz de sustitución a partir del alineamiento y se aplica nuevamente la matriz encontrada en el alineamiento para encontrar uno mejor hasta un punto de convergencia donde el alineamiento no se pueda mejorar.

Hasta donde se sabe este es el primer algoritmo que realiza el calculo de una matriz de sustitución para secuencias del tamaño de un genoma sin asumir nada previamente o usar datos de previos alineamientos.

El método ha sido aplicado sobre información real con distancias filogenéticas y composición de nucleótidos diferentes la cual engañaría a métodos estadísticos clásicos.

5. REVISIÓN Y COMPARACIÓN DE LAS MATRICES DE SUSTITUCIÓN MAS UTILIZADAS

Dentro de la primera parte de este trabajo se buscaba realizar un estudio de las matrices de sustitución que son utilizadas actualmente con el fin de encontrar la más adecuada a partir de la cual se pueda desarrollar la metodología que se quiere proponer. Para ello se tomaron como principales referencias la matriz PAM y la matriz BLOSUM. Mostramos las características fundamentales y de manera muy general como son construidas para finalmente establecer una comparación entre ambas matrices y seleccionar la más adecuada como base para realizar matrices de sustitución generalizadas utilizando alfabetos arbitrarios.

5.1. Matrices de sustitución

Las matrices de sustitución juegan un papel muy importante en la bioinformática ya que al momento de comparar secuencias se debe tener en cuenta la influencia que puede tener la evolución y como se pueden presentar diferentes tipos de cambios durante el proceso. Para las proteínas en el proceso de sustitución se considera más aceptable el reemplazo de aminoácidos similares que aquellos que son divergentes con respecto a sus características físico químicas .

Existen dos maneras de construir matrices de sustitución para proteínas : una de ellas es basada en las propiedades de cada aminoácido como lo son :residuo hidrofóbico,carga electronegatividad y el tamaño viendo así la importancia de conocer las propiedades físico químicas de los aminoácidos otra manera de construir matrices de sustitución. La segunda manera se basa en el código genético con lo que se busca el número necesario de sustituciones para pasar de un codón a otro teniendo en cuenta las dos secuencias alineadas.

La otra manera más común de construir matrices de sustitución es mediante la observación de la proporción de sustitución real entre varios aminoácidos en la naturaleza, por lo que si una sustitución entre dos aminoácidos es observada frecuentemente será puntuada positivamente por lo que se maneja una idea muy sencilla pero basados en algo que ocurre en la naturaleza.

Por esta razón la segunda manera es la más utilizada ya que es más intuitiva y se ajusta mejor a lo que realmente puede estar ocurriendo entre diferentes secuencias de proteínas. Lo que se desea conocer es si dos secuencias son homólogas, es decir están evolutivamente relacionadas, o no lo está; por lo que se busca una puntuación para el alineamiento que refleje la razón con la que un residuo puede cambiar a otro teniendo en cuenta alineamientos realizados anteriormente. Por este motivo y basados en alineamientos sin gaps se han desarrollado dos matrices muy utilizadas para puntuar alineamientos de secuencias en bioinformática : la matriz PAM y la matriz BLOSUM que de manera diferente buscan puntuar cada una de las sustituciones que se puede dar de un aminoácido a otro teniendo en cuenta una análisis inicial sobre una gran cantidad de secuencias que estaban de cierto modo relacionadas y que sirven como base para construir estas matrices.

Para la construcción de una matriz de sustitución para proteínas necesitamos alrededor de 400 valores teniendo en cuenta los 20 aminoácidos pero como durante la construcción de la matriz solo estamos teniendo en cuenta el reemplazo que se de de un aminoácido por otro es decir la dirección de la substitución no influye por lo que A —> B es igual a B —> A con lo que el resultado será una matriz simétrica que nos facilita el total de operaciones que se realizan para la construcción de una matriz de sustitución y que finalmente sera un total de 200 valores utilizados para la construcción de la matriz de sustitución.

A continuación se muestra la idea básica detrás de la construcción de una matriz de sustitución para proteínas basada en el ritmo de sustituciones observadas:

Considera el más simple de los alineamientos es decir un alineamiento global sin gaps de dos secuencias X y Y , de longitud n.

M probabilidad de que los residuos alineados tengan un ancestro en común

R probabilidad de que los residuos fueron alineados por azar

En la puntuación de este alineamiento se evalúa :

$$\frac{Pr(X,Y|M)}{Pr(X,Y|R)}$$

El numerador indica la probabilidad de que haya una sustitución dentro de la secuencia X a la secuencia Y dado que las secuencias tienen un ancestro en común(M). El denominador indica la probabilidad de que haya una sustitución de X a Y dado que las secuencias son alineadas por azar(R).

La matriz de sustitución que puntuará una alineamiento mediante la estimación de esta proporción para cada uno de los pares de aminoácidos que se encuentran en las dos secuencias alineadas.

Sea q_a la frecuencia del aminoácido a.

Sea q_{xi} y q_{yi} los residuos en una posición especifica de ambas secuencias X y Y.

Considera el caso donde el alineamiento de X y Y es aleatorio:

$$Pr(X,Y|R) = \prod_i^n q_{xi} \prod_i^n q_{yi}$$

Sea Pab la probabilidad de que a y b sean derivados de un ancestro común.

El caso donde el alineamiento es debido a un ancestro común es :

$$Pr(X,Y|M) = \prod_i^n P_{xiyi}$$

Esto como resultado de las observaciones relacionadas.

La razón de oportunidades de estas alternativas está dada por:

$$\frac{Pr(X,Y|M)}{Pr(X,Y|R)} = \frac{\prod_i^n P_{xiyi}}{\prod_i^n q_{xi} \prod_i^n q_{yi}} = \frac{\prod_i^n P_{xiyi}}{\prod_i^n q_{xi} q_{yi}}$$

Aplicando log probabilidades:log

$$log \frac{Pr(X,Y|M)}{Pr(X,Y|R)} = log \sum_i^n \frac{\prod_i^n P_{xiyi}}{\prod_i^n q_{xi} q_{yi}}$$

Aplicando log probabilidades podemos realizar un mejor análisis ya que estamos sumando sobre las razones de cada una de las posibles sustituciones y no sobre las multiplicaciones lo que nos podría generar un mayor error debido a que al realizar multiplicaciones de estos valores puede haber una perdida de precisión y desborde en las cantidades al finalizar los cálculos.

5.1.1. BLOSUM

Se basa en alineamientos de secuencias de proteínas evolutivamente divergentes para ello se tomó la base de datos BLOCKS y se analizaron regiones muy conservadas de familias de proteínas luego comprobaron las frecuencias relativas de la aparición de aminoácidos y la probabilidad de sustitución entre ellos. Las matrices BLOSUM se describen de la siguiente manera BLOSUM seguidas de un número ejemplo : BLOSUM 62, el número que acompaña el nombre de la matriz indica el porcentaje de identidad entre las secuencias de proteínas utilizadas para la construcción de la matriz. Por lo tanto BLOSUM 62 une a todas las proteínas en un alineamiento >= 62 de identidad. El número hace referencia al mínimo porcentaje de similitud entre los bloques usados para construir la matriz..Un número alto indica que usando esta matriz se comparan secuencias estrechamente relacionadas y un número bajo indica que usando esta matriz se comparan secuencias relacionadas de manera distante.

Las puntuaciones dentro de la matriz BLOSUM corresponden a log odd probabilidades. Y cada posición de la matriz está representada de la siguiente manera en términos generales:

$$a_{ij} = \left(\frac{1}{\lambda}\right) \log\left(\frac{P_{ij}}{q_i q_j}\right)$$

P_{ij} es la probabilidad de que dos aminoácidos i y j reemplacen uno a otro en una secuencia homóloga.

q_i y q_j son las probabilidades ultimas de encontrar los aminoácidos i y j en cualquier secuencia de proteína de forma aleatoria.

El factor λ de proporcionalidad para obtener puntuaciones que difieren después del redondeo para pares distintos de aminos.

En la matriz puntuaciones positivas indican sustituciones conservadas y las puntuaciones negativas indican sustituciones no conservadas.

A continuación se muestra de manera general como se realiza el cálculo de la matriz BLOSUM:

- Agrupar varias secuencias en un cluster siempre que el porcentaje de similitud sea mayor o igual que el valor L % definido para la matriz BLOSUM L, es decir se cumpla con esta similitud a lo largo de todas las secuencias utilizadas para la construcción de la matriz de sustitucion.
- Contar el número de sustituciones dentro de un mismo cluster .
- Estimar la frecuencia basado en los conteos.

5.1.2. PAM

Esta matriz es derivada de alineamientos globales de secuencias cercanamente relacionadas, es motivada por la evolución. En si la matriz PAM es una matriz usada para buscar relaciones entre secuencias divergentes que fue construida a partir de secuencias estrechamente relacionadas.

El modelo de evolución de las matrices PAM tiene las siguientes características:

- Cada posición cambia independientemente del resto.

- La probabilidad de mutación es la misma en cada posición.
- La evolución no recuerda(sin memoria).

Ahora definiremos la medida de distancia PAM : Sea S_1 y S_2 dos secuencias de proteínas con $|S_1| = |S_2|$ decimos que S_1 y S_2 están a una distancia x PAM S_1 muy probablemente fue producida a partir de S_2 con x mutaciones por 100 aminoácidos. Hay que diferenciar la unidad de medida PAM con el porcentaje de identidad de una secuencia por ejemplo la matriz PAM 250 tiene en cuenta secuencias con el 20 % de identidad. Pares de secuencias con un medida PAM > 250 probablemente no son homólogos ya que para estos pares % de identidad de secuencia es menor al 20 %.

El cálculo de las matrices PAM se basó en árboles filogenéticos que se construyeron a partir de 71 familias de proteínas estrechamente relacionada, las proteínas a ser estudiadas fueron seleccionadas sobre la base que tuvieran alta similitud con sus predecesores.

Las puntuaciones dentro de la matriz PAM corresponde a log probabilidades. El valor de una determinada celda representa la probabilidad de la sustitución de un aminoácido por otro a esto denominamos mutación puntual, así , la matriz se calcula observando diferencias en las proteínas muy cercanas evolutivamente por lo que las sustituciones en cuestión no tienen efecto sobre la función de la proteína . Y cada posición de la matriz está representada de la siguiente manera en términos generales:

$$M_x\,(i,j) = \log\left(\frac{f_{ij}}{f_i f_j}\right)$$

f_{ij}es la probabilidad de que dos aminoácidos i y j reemplacen uno a otro en una secuencia homologa. Proporción de mutación.

$f_i f_j$son las frecuencias relativas de los aminoácidos i y j en cualquier secuencia de la proteína.

Las frecuencias de las mutaciones de PAM X dependen linealmente de las frecuencias de las mutaciones de PAM 1 de modo que PAM X es calculada al repetir la multiplicación de matrices de PAM 1 por ella misma.

$$M_x = (M_1)^x$$

PAM	BLOSUM
Calculada a partir de alineamientos globales	Cálculo a partir de alineamientos locales
Secuencias de proteínas usadas en alineamiento tienen $> 99\,\%$ de identidad	Puede seleccionar un nivel de similitud entre las secuencias usadas en el análisis
La matriz más utilizada es la PAM 250	La matriz más usada es BLOSUM 62
Las matrices son extrapolaciones matemáticas de la matriz PAM 1	Cada matriz es el resultado del análisis de un alineamiento de bloques conservados (análisis real)
Es posible elaborar un modelo evolutivo y así generar nuevas matrices a partir de la primera	No permite generar un modelo evolutivo
	Permite detectar las mejores secuencias con relación biológica ya que se basa en alineamiento de secuencias existentes y no en valores extrapolados.

BLOSUM80	BLOSUM62	BLOSUM45
PAM1	PAM120	PAM250
Menos divergente	< ——————————— >	Mas divergente

Debido a que en nuestra generalización buscamos tener un visión general es oportuno realizar una profundización sobre su implementación principalmente en los datos que son usados inicialmente para construir los clusters con el determinado % de identidad.

5.2. Elección de la matriz a utilizar y el algoritmo

5.2.1. Comparación entre BLOSUM y PAM

Debido a que en este trabajo deseamos mostrar una metodología que permita la construcción de matrices de sustitución generalizadas y evaluando entre las matrices PAM y BLOSUM se puede comprobar que resulta mas coherente la utilización de las matrices BLOSUM si la idea es generalizar y poder llevar acabo estudios con diferentes grupos de proteínas ya que los resultados estarán estrechamente relacionados con las fuentes utilizadas para la construcción de las matrices es por esta razón que hemos decidido utilizar las matrices BLOSUM ya que se basan en alineamientos de secuencias de proteinas existentes.

5.2.2. Matriz seleccionada : BLOSUM y su algoritmo

El algoritmo inicial sobre el cual se desarrollaron las matrices BLOSUM remonta al año 1991 esta base se ha mantenido con algunas modificaciones que se han ido realizando durante los años siguientes a su creación permitiendo corregir problemas en dicho algoritmo pero conservando su base y usando diferentes bases de datos como entrada que corresponden a las distintas versiones de BLOCKS.

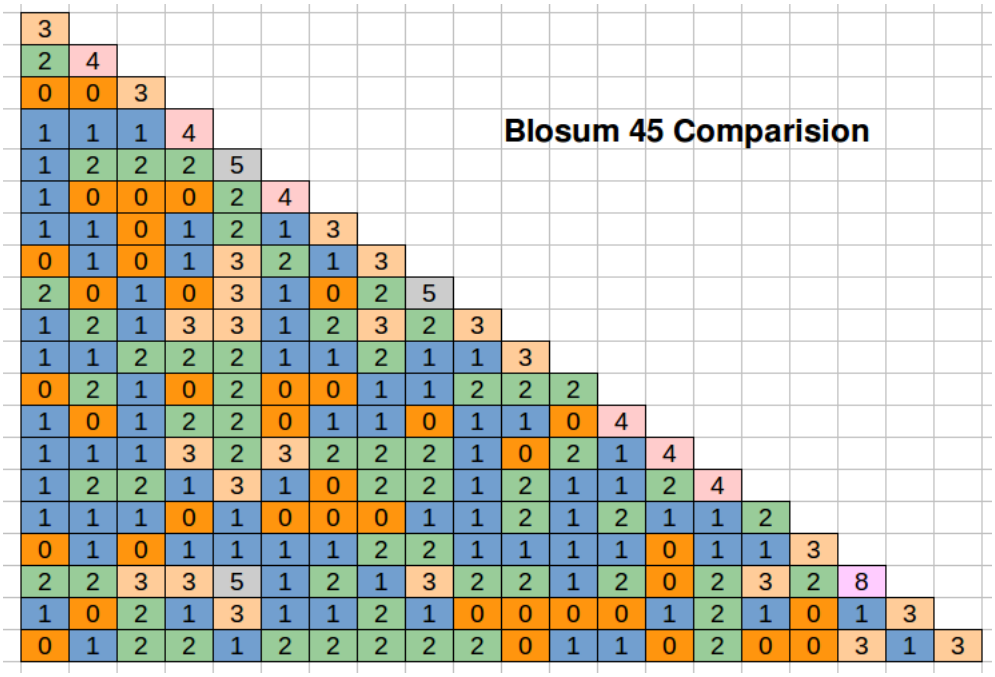
La base de datos BLOCKS contiene multiples alineamientos de regiones conservadas en familia de proteinas. La base de datos BLOCKS esta basada en entradas de InterPro con secuencias de SWISS_PROT y TrEMBL tambient tiene referencias a PROSITE, PRINTS , SMART, PFAM y ProDOM. La base de datos fue construida usando el sistema PROTOMAT que mediante el uso de un algoritmo de MOTIFS y mediante un proceso interactivo se construyeron todos los bloques que conforman la base de datos que describen cada una de las familias.

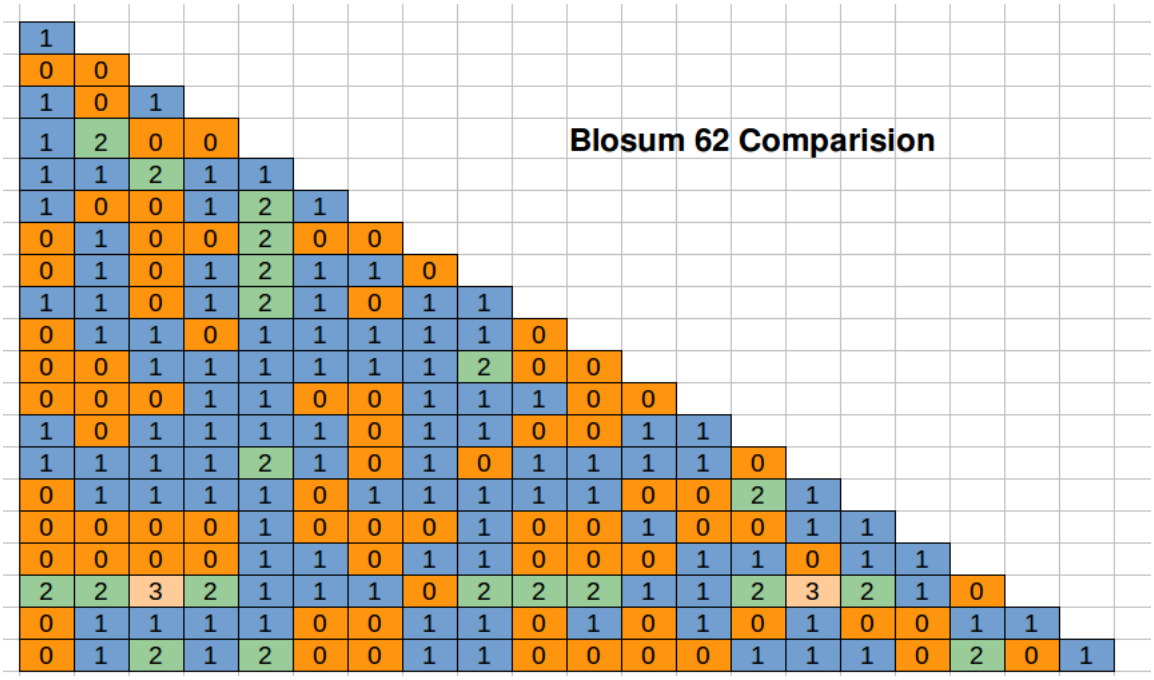
Inicialmente se planteaba construir los bloques usando el sistema PROTOMAT pero debido a la complejidad y al cambio que ha habido en cada una de las bases de datos con el tiempo se ha decidido trabajar con la ultima versión de BLOCKS que corresponde a la versión 14.3 del 2007.

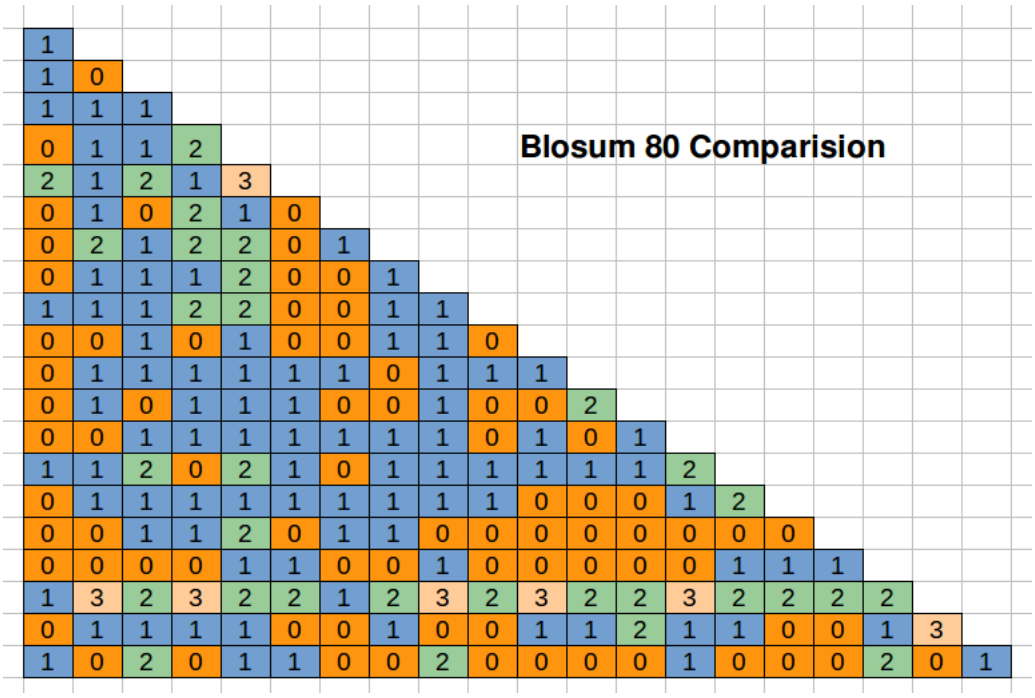
Para el desarrollo de este trabajo de grado hemos tomado como base un algoritmo que ha sido revisado y corregido algunos problemas que presentaba el algoritmo original, irónicamente los resultados que arroja la versión corregida son peores cuando hablamos de búsquedas en bases de datos usando como base las matrices de sustitución arrojadas por el algoritmo es por esta razón que basados en los alcances de este proyecto se decide realizar tres pruebas para verificar que el algoritmo revisado que a pesar de tener problemas con las matrices cuando son utilizadas para búsquedas conservan la misma información que requerimos para el proceso de generalización.

Para realizar estas pruebas tomamos como base de datos Blocks 14.3 y el algoritmo Blosum revisado y construimos las matrices de sustitución Blosum 45 , Blosum 62 y Blosum 80 después de obtener

estas matrices comparamos los resultados arrojados con las versiones que son manejadas actualmente como estándar de las mismas matrices he hicimos una comparación viendo que diferencias había entre la versión antigua y la nueva versión de cada una de las matrices de sustitución. A continuación se presentan tres imágenes correspondientes a la comparación de cada una de las matrices en donde cada valor en la matriz tiene un color asociado para reconocer a simple vista que tan divergente es la matriz arrojada de su antepasado.







Como podemos ver en cada una de las imagenes las diferencias no van mas allá de 3 unidades en la mayoría de las matrices exceptuando la matriz Blosum 45 que presenta valores mas divergentes cabe resaltar que cada uno de estos valores están determinados por los parámetros con los cuales se corrió el algoritmo y la base de datos utilizada para las matrices originales se uso Blocks 5.0 para estas nuevas matrices se uso la base de datos Blocks 7.0 que es la que se tendrá de base para realizar todo el desarrollo durante el trabajo de grado.

A partir de estos resultados podemos ver que es altamente viable usar el algoritmo revisado como base para las implementaciones de los algoritmos generalizados de las matrices de sustitución.

6. Conclusiones

7. Trabajos futuros

8. Referencias

9. Anexos