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

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 biológicas. 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 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 PAMf(insertar referencia aca) y el modelo de Henikoff and Henikoff de donde nacen las matrices BLOSUM(insertar referencia aca). 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 biol´ogicas 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 sustituci´on 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 sustituci´on generalizadas en su desarrollo.
4. Analizar la metodolog´ıa y los resultados de los algoritmos desarrollados mediante su aplicación en diferentes casos de estudio.

3. MARCO TEÓRICO

3.1. Matrices de Sustitución

Las matrices de sustitución son un conjunto de valores que permiten cuantificar la probabilidad de que un residuo sea sustituido por otro en un alineamiento, son derivadas de un análisis estadístico de la sustitución de residuos de conjuntos de datos de alineamientos seguros de secuencias altamente relacionadas. Las matrices de sustitución para secuencias de nucleótidos son relativamente simples. Un valor positivo o un alto puntaje es dado para un acierto y un valor negativo o bajo puntaje para un fallo. Esta asignación se basa en la suposición que podría no ser realista; observaciones muestran que transiciones (sustituciones entre purinas y purina o entre pirimidinas y pirimidinas) ocurren más frecuentemente que las transversiones (sustituciones entre purinas y pirimidinas).

3.2. PAM

Las matrices PAM fueron las primeras construidas , quienes recogían alineamientos de 71 grupos de secuencias de proteínas muy 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 en 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 (de ahí su nombre) 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 es definida como 1 % de las posiciones de aminoácidos que se han cambiado. para construir una tabla de sustitución PAM1, 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. [6]

3.3. 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 huecos 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 (como es el caso de las matrices PAM, obtenidas al multiplicar por sí misma un determinado número de veces una matriz inicial).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.

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]

3.4. 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 de DNA y RNA, 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, y grupo carboxilo(-COOH). 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]

3.5. 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, adenina, timina, citosina(G,A,T y C). Cada base está unida a un grupo fosfato y a una azúcar desoxirribosa para formar un nucleótido. La única cosa que hace diferente 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]

3.6. Alineamientos 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 compararon mediante la búsqueda de patrones de caracteres comunes y establecer residuo-residuo 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 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. Para las secuencias divergentes y secuencias de longitudes variables, este método puede no ser capaz de generar resultados óptimos, ya que falla para reconocer regiones locales muy similares entre las dos secuencias. 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 para conservar patrones de ADN 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. Simplificación de alfabetos de aminoácidos por medio del algoritmo de Branch and Bound y matrices de sustitución

Las proteínas y ADN son generalmente representados por secuencias de letras. En un número de circunstancias los alfabetos simplificados(donde una o más letras pueden ser representadas por el mismo símbolo) han demostrado su potencial utilidad en varios campos de la bioinformática incluyendo búsqueda de patrones que se producen a un ritmo inesperado, el estudio del plegamiento de proteínas y la búsqueda de consenso en alineamiento de múltiples secuencias . El tema principal abordado en este trabajo es la posibilidad de encontrar un enfoque general que permita un análisis exhaustivo de todos los posibles alfabetos simplificados, utilizando matrices de sustitución.

Una matriz de sustitución para alfabetos estructurales basada en alineamiento estructural de proteínas homólogas y sus aplicaciones. En el presente trabajo, hemos demostrado, usando un alfabeto estructural, la utilidad de codificación de una estructura 3D en el espacio 1D mediante el uso de una matriz de sustitución.

El uso de representación 1D de estructura 3D combinada con nuestra matriz de sustitución y programación dinámica simple permitio obtener buenos resultados incluso en el caso de dos homólogos con longitudes de secuencia muy diferentes, se logró localizar regiones de similitud estructural, destacar el cambio sutil en conformaciones dentro de las regiones alineadas e identificar a regiones de ninguna similitud estructural. [2]

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

Nosotros proponemos un método general para derivar matrices de sustitución para aminoácidos de los campos de fuerza de baja resolución. A diferencia de los métodos populares actuales, el enfoque no se basa en los argumentos evolutivos o la alineación de secuencias o estructuras. En su lugar, los residuos son computacionalmente mutados y ellos contribuyen al total de energía/puntuación que es recolectada. El promedio de estos valores sobre cada posición dentro de un conjunto de proteínas resulta en un matriz de sustitución.[3]

4.3. Fold-specific substitution matrices for protein classification

Los métodos que se centran en estructuras secundarias tales como matrices de puntuación para posiciones específicas y los modelos ocultos de markov, han probado la utilidad de asignar proteínas a familias. Sin embargo, para asignar proteínas a un atributo de clase dentro de una familia estos métodos pueden introducir más parámetros de los necesarios. Nosotros describimos un método para organizar proteínas en una familia que presenta menos una reducción en el número de parámetros. Esto conduce a la definición de una matriz de sustitución de atributos de clase (CLASSUM) similar a la BLOSSUM. [4]

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

Las matrices de sustitución de aminoácidos juegan un papel central en métodos de alineamiento de prote´ınas. Matrices log-odds est´andar , tales como la series de PAM y BLOSSUM , son construidas a partir de grandes conjuntos de alineamientos de proteínas teniendo implícito en el fondo las frecuencias de los aminoácidos. Sin embargo , estas matrices frecuentemente son usadas para comparar proteínas con diferencias muy marcadas en la composición de los aminoácido, tales como proteínas de transmembrana o proteínas de organismos con una composición de bases de nucleótidos fuerte. Se ha dicho en otra parte que las matrices estándar no son ideales para este tipo de comparaciones y, además, por esta razón se ha presentado la transformación de una matriz estándar para su uso en un contexto de composición no estándar. [5]

5. TRABAJO REALIZADO

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. Mostrando 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 fisico quimicas de los aminoácidos otra manera de construir matrices de sustitución basadas 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 estan evolutivamente relacionadas) o no, por lo que se busca una puntuación para el alineamiento que refleje eso. Por esta razón 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 pueden 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 sirvieron como base para construir estas matrices.

Para la construcción de una matriz de sustitución para proteínas necesitamos alrededor de 200 valores teniendo en cuenta los 20 aminoácidos y que 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.

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.
- 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 de X a Y dado que las secuencias tienen un ancestro en común. El denominador indica la probabilidad de que haya una sustitución de X a Y dado que las secuencias son alineadas por azar.
- 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.
- 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 :

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

- 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:

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

- 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.

5.1.1. BLOSUM

Una de las matrices de sustitución más utilizada actualmente y que fue definida por Henikoff y Henikoff en 1992 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, 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.

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.

Las puntuaciones dentro de la matriz BLOSUM corresponde a log 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 homologa.

q_{ij}son las probabilidades ultimas de encontrar los aminoácidos i y j en cualquier secuencia de proteína de forma aleatoria.

El factor λ es un mero factor para asegurar que tras aplicarlo se obtendrá un necesario redondeo al entero más cercano. Con lo que la matriz contendrá valores dispersos fácilmente tratables.

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 más que el L % de los residuos idénticos es compartidos para BLOSUM-L.
- Contar el número de sustituciones a través de los diferentes cluster(en la misma familia) .
- Estimar la frecuencia usando conteos.

5.1.2. PAM

La matriz de sustitución PAM fue definida por Margaret O. Dayhoff en 1978 esta matriz es derivada de alineamientos globales de secuencias cercanamente relacionadas, es motivada por la evolució. 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:

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

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

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

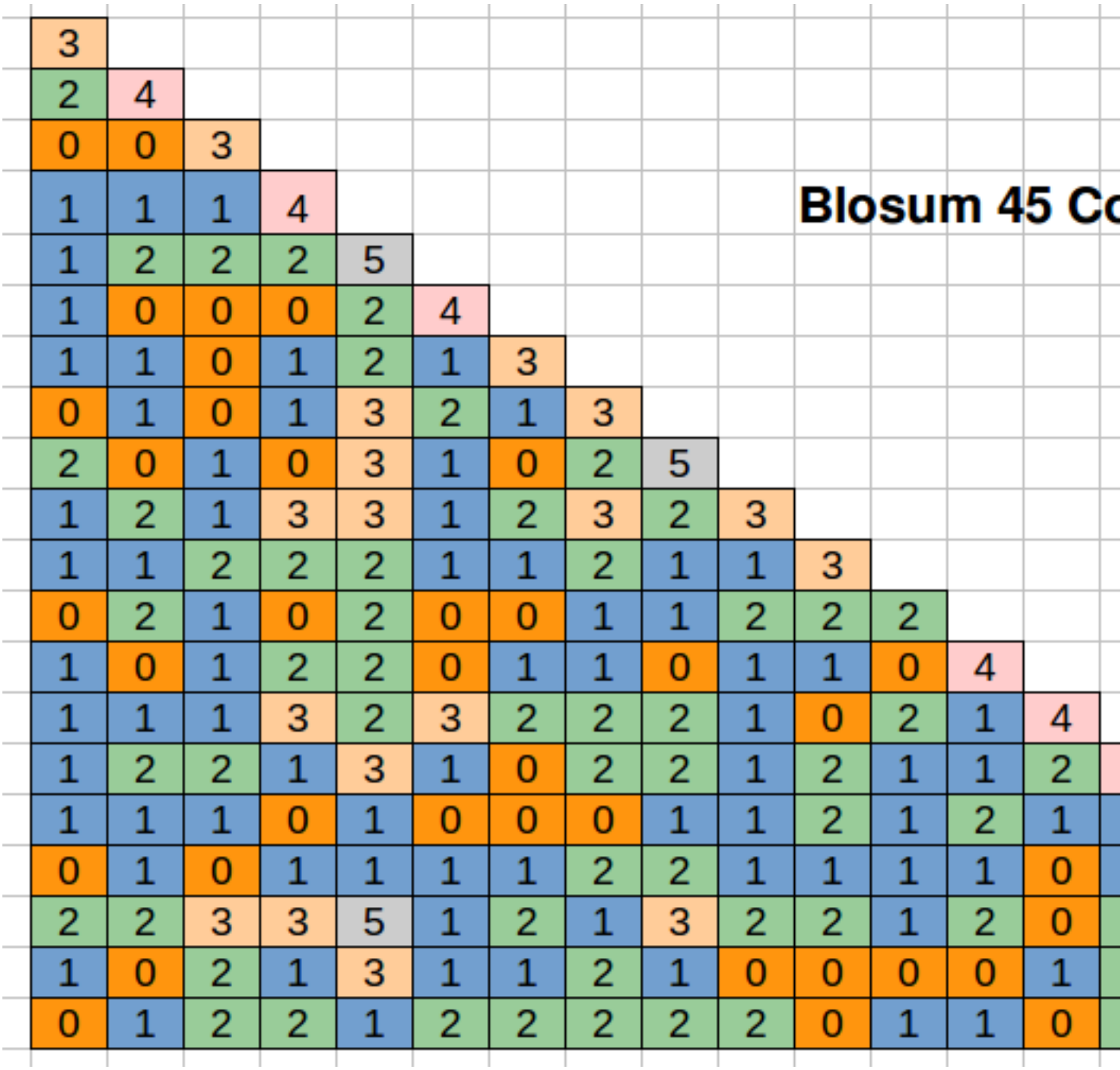
Debido a que en nuestra generalización buscamos tener un visión general es oportuno realizar una profundización sobre su implementación principalmente en los datos que son usados inicialmente para construir los clusters con el determinado % de identidad.

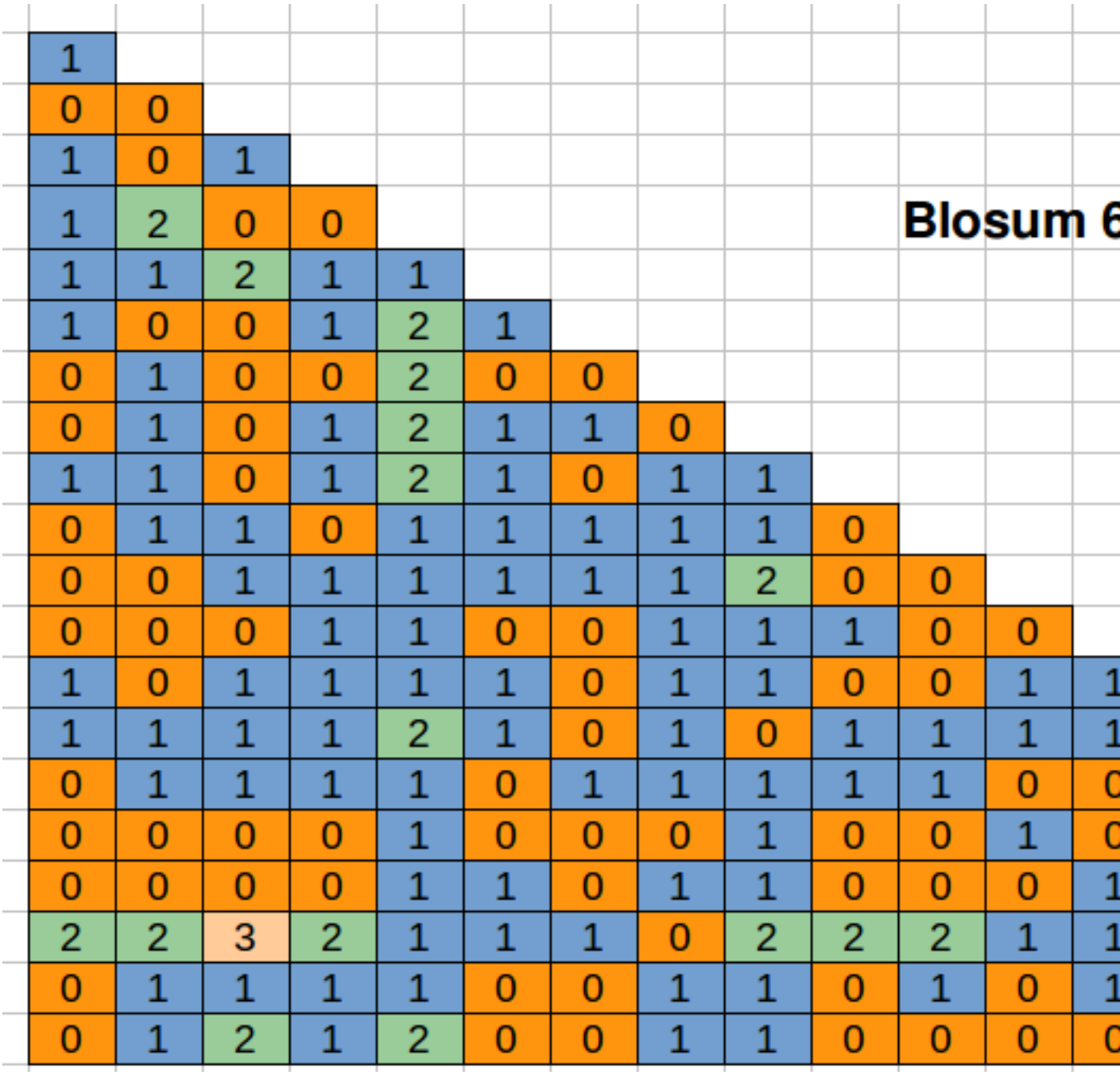
5.2. ELECCIÓN DEL ALGORITMO

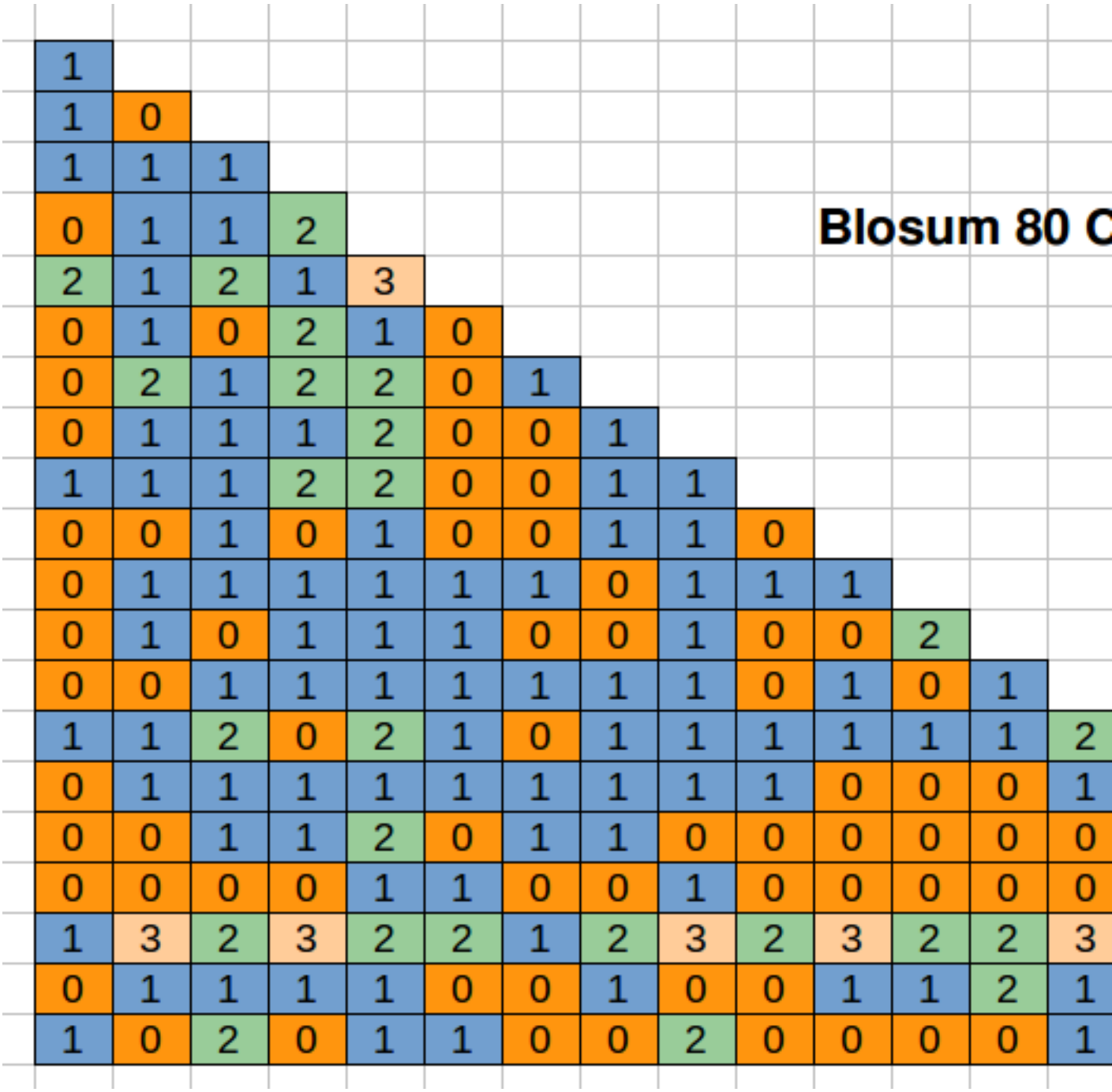
El algoritmo inicial sobre el cual se desarrollaron las matrices Blosum remonta al año 1991 esta base se ha mantenido con algunas modificaciones que se han ido realizando durante los años siguientes a su creación permitiendo corregir problemas en dicho algoritmo pero conservando su base y usando diferentes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

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 7.0 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.


Blosum 45 Co


Blosum 6



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