Revista de Matemática: Teoría y Aplicaciones 2008 **15**(1) : 27–40
cimpa – ucr issn: 1409-2433

GENERALIZACIÓN DE DOS MÉTODOS DE DETECCIÓN DE CONGLOMERADOS. APLICACIONES EN BIOINFORMÁTICA

Laureano Rodríguez Corvea* Gladys Casas Cardoso†
Ricardo Grau Abalo‡ Mario Pupo Meriño §

Recibido/Received: 26 Febrero 2007; Aceptado/Accepted: 31 Oct 2007

Resumen

En este artículo se propone una modificación a los métodos de detección de conglomerados de enfermos en el tiempo, para que puedan utilizarse en otras ramas del saber. Se muestran especificaciones particulares en los métodos Scan y Grimson. Los datos originales se transforman en una secuencia binaria donde los "unos" representan la categoría de interés y los "ceros" se corresponden con los demás casos. Las transformaciones particulares de los algoritmos se implementaron en el paquete Mathematica. Además se presentan varias aplicaciones interesantes del campo de la bioinformática.

Palabras clave: Conglomerados temporales, método Scan, método de Grimson, aplicaciones bioinformáticas.

Abstract

This paper presents a modification to the temporal disease clusters methods in order to apply them to other sciences. Particular specifications are shown in Scan and temporal Grimson techniques. The original data are transformed in a binary sequence: the value "one" represents the interest category and value "zero" corresponds to all other cases. Computational transformations of the algorithms are implemented using Mathematica Package. Besides some interesting bioinformatics applications are presented.

*Departamento de Informática, Facultad de Ciencias Médicas de Sancti Spíritus, Cuba. E-Mail: corvea@uclv.edu.cu

†Grupo de Bioinformática, Facultad de Matemática, Física y Computación, Universidad Central "Marta Abreu" de Las Villas, Cuba. E-Mail: gladita@uclv.edu.cu

‡Misma dirección que G. Casas, E-Mail: rgrau@uclv.edu.cu

§Misma dirección que G. Casas, E-Mail: mpupo@uclv.edu.cu

Keywords: Temporal disease clusters, Scan method, Grimson method, Bioinformatics applications.

Mathematics Subject Classification: 62P10

1 Introducción

El conocimiento matemático del mundo moderno avanza más rápido que nunca. Teorías que en sus inicios eran completamente distintas, se han reunido para formar teorías más completas y abstractas. Numerosos problemas importantes ya han sido resueltos con éxito, no obstante en la actualidad aparecen constantemente nuevas dificultades que la ciencia deberá solucionar. Los estudios que se realizan sobre los diferentes genomas constituyen un ejemplo concreto de esta realidad [2].

La secuenciación de genomas ha generado un amplio catálogo de miles de millones de pares de bases o de aminoácidos. Para ayudar a los investigadores a determinar el sentido de esa enorme cantidad de datos han hecho falta muchos instrumentos matemáticos e informáticos [2] y [3].

La Estadística es una de las ramas de la matemática que más posibilidades tiene de aportar instrumentos confiables para obtener conocimientos a partir de grandes volúmenes de datos. Su desarrollo se intensificó de manera notable a partir de 1960 con la aparición de de las computadoras digitales. Numerosos métodos que surgieron para resolver problemas de una ciencia particular, se han generalizado y su campo de aplicación se ha diversificado grandemente [8].

Por otra parte, la detección de conglomerados de enfermos, a partir del desarrollo y aplicación de métodos estadísticos específicos, constituye un problema epidemiológico en el que se ha venido trabajando también desde hace relativamente poco tiempo [13]. Las primeras publicaciones al respecto aparecieron en 1964 por Knox [16] y a partir de esa fecha han tenido un incremento exponencial [12]. El objetivo fundamental de estas técnicas es detectar la presencia de un exceso de casos diagnosticados de una determinada enfermedad en espacio, tiempo o considerando ambos escenarios a la vez, [5] y [7]. Las técnicas clásicas de detección de conglomerados, (métodos jerárquicos, o de las k medias), no resuelven el problema de manera correcta, por lo que fue necesario desarrollar e implementar métodos matemáticos más específicos [15].

En este trabajo se modifican dos de las técnicas de detección de aglomeraciones en el tiempo: el método de Scan y el de Grimson, con el propósito de ampliar su campo de aplicación. Las modificaciones presentadas constituyen instrumentos alternativos novedosos para la solución de diversos problemas, en este artículo se muestran soluciones a varios problemas clásicos de bioinformática. La programación de las técnicas se realizó en todos los casos con utilizando el paquete Mathematica.

2 Ideas de las transformaciones propuestas

En epidemiología, un conglomerado o "cluster" temporal de enfermos es un exceso de casos diagnosticados muy cercanos en el tiempo [6].

En los métodos originales de detección, la variable de interés es el tiempo en que ocurre el evento. Dicho evento pudiera ser la fecha de diagnóstico de la enfermedad o incluso, la fecha en la que aparecieron los primeros síntomas si esta es suficientemente precisa [14]. El primer paso del algoritmo consiste en ordenar cronológicamente los datos obtenidos. Posteriormente se divide el eje que representa el tiempo total considerado en intervalos fijos que puede ser años, meses o días. A partir de este punto cada método sigue sus propios pasos para determinar si existen conglomerados.

Estos algoritmos pueden transformarse para ampliar su campo de aplicación. La idea que se defiende en este trabajo es generalizarlos de manera que ellos puedan utilizarse para detectar conglomerados en un sentido más universal.

Para lograrlo se propone ordenar los datos por algún criterio determinado que depende del campo de aplicación. Si se trabaja con fechas, los datos se ordenan cronológicamente, si se trabaja con secuencias de bases que representan algún gen completo, o una porción de este, sería correcto asumir que tal juego de datos ya está ordenado.

El segundo paso consiste en transformar dicha secuencia en una secuencia análoga, pero dicotómica. El valor uno se colocará cada vez que aparezca la categoría de interés: una base, un aminoácido o una subsecuencia determinada dentro de una secuencia del ADN o de proteínas una fecha y otro evento que se considere. El valor cero se asociará a todas las demás categorías, [17]. Los datos transformados se representan en una línea. El nuevo problema que surge es el de determinar si en la secuencia formada por ceros y unos existen conglomerados de unos.

Por ejemplo, supóngase que se tiene una determinada secuencia de un gen y que dentro de ella resulta de interés determinar si existen conglomerados de la subsecuencia GCG. La transformación de la secuencia original en una dicotómica se realiza como se muestra en la figura 1.

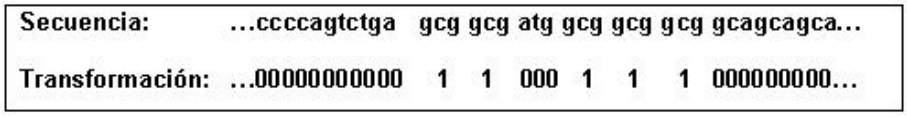

Figura 1: Ejemplo de conversión de una porción de la secuencia de un gen.

Obsérvese que la categoría de interés: subsecuencia GCG, se sustituyó por un uno, mientras que el resto de los casos considerados se sustituyó por el valor cero.

3 Los métodos Scan

Los métodos Scan surgieron con el propósito de detectar aglomeraciones dentro de períodos de tiempo consecutivos, pues pudiera suceder que un conglomerado temporal se extienda por dos o más intervalos [14].

3.1 El método Scan sobre una línea

El método clásico define un intervalo o ventana de tamaño fijo que se mueve, con un determinado paso, por la línea del tiempo que se analiza. En el caso generalizado que se propone en este artículo, la ventana se movería por la secuencia de unos y ceros en que se transformó el problema original. La idea del método radica en que si existe un conglomerado, el número máximo de la categoría de interés (unos) hallado en una ventana, debe ser muy grande al compararla con las cantidades que aparecen en la mayoría de las ventanas restantes.

De manera general se pueden definir las hipótesis de la forma siguiente:

H_0 : La categoría representada por unos se distribuye uniformemente dentro de la secuencia considerada.

H_1 : Existe al menos un conglomerado dentro de alguna de las ventanas analizadas [18].

Para la formulación matemática se definen:

t : amplitud de la ventana,

T : período de tiempo total que se analiza, (en el método original), o largo de la secuencia, (en la modificación propuesta)

$L = \frac{T}{t}$: fracción que representa el período de tiempo (o de longitud) total que se analiza con relación al ancho de la ventana

n : cantidad de enfermos (o de unos) diagnosticados en T

λ : número esperado de casos (o de unos) por unidad en un proceso de Poisson

$w_{y,y+t}$: cantidad de enfermos (o de unos) en la ventana $[y, y+t)$.

El estadístico: $w = w_t(T) = \max_{0 \le y \le T - t} \{w_{y,y+t}\}$ representa el número de casos (o de unos) que aparecen en una ventana cuando se mueve continuamente a lo largo del tiempo (o del eje longitudinal analizado). En la práctica, la ventana $[y, y+t)$ se mueve discretamente a partir de una sucesión de puntos equidistantes$y_1, y_2, \ldots y_k$ que cubren todo el período de análisis de amplitud T. Se denomina paso del Scan o paso del desplazamiento a: $\Delta y = y_k - y_{k-1}$. El estadístico anterior se estima por su versión discreta: $\bar{w} = \bar{w}_{t,\Delta t}(T) = \max_{1 \le i \le k\ t} \left\{w_{y_i, y_{i+t}}\right\}$

El test estadístico depende de varios de los parámetros explicados anteriormente y en esencia calcula la probabilidad p de que aparezcan w o más casos en una ventana. La fórmula que se utilizó para p es la propuesta en expresión 1 [19] y que se detalla en las funciones creadas con el Mathematica que aparecen en los anexos.

3.2 Algoritmo para el método Scan sobre una línea

A continuación se muestra el algoritmo para la aplicación del método de Scan generalizado.

Paso 1: Representar en una línea los datos transformados en ceros y unos.
Paso 2: Definir un intervalo o una ventana de longitud fija (cantidad de elementos), según criterio del especialista.
Paso 3: Desplazar con un paso fijo la ventana a lo largo de la línea de longitud y determinar en cada caso, la cantidad de unos asociados a ella.
Paso 4: Encontrar el número máximo de unos (categoría de interés) de la ventana móvil. Ese valor constituye el estadígrafo del test.
Paso 5: Calcular la probabilidad del test utilizando la fórmula propuesta en [19]:

$$p = P^*(w, \lambda L, \frac{1}{L}) = 1 - Q^*(w, \lambda L, \frac{1}{L}) \tag{1}$$

En el anexo aparecen las funciones más importantes que se programaron sobre el paquete Mathematica.

3.3 El método Scan sobre un círculo

Este método constituye una variación del anterior y se utiliza para problemas que tienen un comportamiento estacional. Los datos se encuentran ordenados a lo largo del eje de longitud y el círculo se forma uniendo el final con el inicial.

El algoritmo es en esencia, el mismo al lineal. La ventana se desplaza sobre el círculo y se determina en cada una el número de veces que aparece la categoría de interés (unos) asociados a ella. Con este desplazamiento circular se pretende incorporar al análisis la cercanía de posibles casos a finales del último período considerado con los del principio del primer período considerado, como si fueran períodos continuos.

Dado los detalles anteriores el algoritmo varía en el paso 5 en el cálculo de la probabilidad de observar w o más casos en un intervalo o ventana de tamaño fijo ya que Q no se estima de igual forma [19].

$$p = P^*(w, \lambda L, \frac{1}{L}) = 1 - Q^*(w, \lambda L, \frac{1}{L}) \tag{2}$$

En el anexo aparecen las funciones más importantes que se programaron sobre el paquete Mathematica

4 El método Grimson

El test de Grimson se considera uno de los métodos más generales y versátiles en la detección de conglomerados de enfermos. El puede aplicarse indistintamente al caso temporal, espacial y espacio-temporal [9]. Por analogías con el método Scan se explicará solamente la generalización de la versión temporal, aunque las mismas ideas pueden extenderse a los demás escenarios.

La idea central del método consiste en dividir la línea del tiempo en intervalos consecutivos no solapados a los que se les llamarán celdas. Aquellas celdas, cuya cantidad de unos (categoría de interés) exceda el valor esperado de una distribución de Poisson [12] o un umbral determinado a priori por un especialista, recibirá la categoría de marcada. Con

esta información podemos definir la hipótesis nula y la alternativa de la siguiente forma [10] y [11]:

H_0 : Las celdas marcadas se distribuyen uniformidad entre todas las celdas consideradas.

H_1 : Existe un gran número de celdas marcadas adyacentes.

El criterio de adyacencia entre las celdas se define de la forma usual: dos celdas son adyacentes si ambas son contiguas. Si se rechaza la hipótesis fundamental podrá concluirse que entre las celdas consideradas existe al menos un conglomerado.

Para la formulación matemática se definen:

c : número total de celdas.

m : número total de celdas marcadas, $(m < c)$.

y_i : número de celdas adyacentes a la celda i, $i = 1, 2, \ldots, c$, o equivalentemente el número de bordes de la celda y_i.

Para el método de Grimson es importante conocer el promedio de bordes de las celdas (y) y su varianza $(V(y))$; ambas se calculan a partir de las fórmulas clásicas.

El estadístico A de Grimson representa el número observado de pares de celdas marcadas adyacentes. Bajo la hipótesis nula, A tiene distribución normal asintótica con valor esperado y varianza que se calculan según las fórmulas siguientes [10]:

$$E(A) = \frac{\bar{y}m(m-1)}{2(c-1)} \tag{3}$$

$$\begin{aligned} V(A) &= E(A)\left[1 + \frac{2(\bar{y}-1)(m-2)}{c-2} + \frac{(c\bar{y} - 4\bar{y} + 2)(m-2)(m-3)}{2(c-2)(c-3)} - E(A)\right] + \\ & V(y)\left[1 + \frac{m(m-1)(m-2)}{(c-1)(c-2)(c-3)} + \frac{m(m-1)(m-2)(m-3)}{(c-1)(c-2)(c-3)(c-4)}\right]. \end{aligned} \tag{4}$$

Como caso particular, si c es grande y $\frac{m}{c}$ pequeño, A sigue una distribución de Poisson con parámetro $E(A)$ [14].

4.1 Algoritmo para el método Grimson

Algoritmo generalizado del método Grimson.

Paso 1 : Representar en una línea los datos transformados en ceros y unos.

Paso 2 : Dividir la línea del tiempo en intervalos fijos no solapados a los cuales se les llamará celdas.

Paso 3 : Determinar la frecuencia de unos (categoría de interés) de cada celda.

Paso 4 : Las celdas, cuya cantidad de unos exceda el valor esperado de una distribución de Poisson [12] o un umbral determinado a priori, recibirán la categoría de "marcada" .

Paso 5 : Calcular el estadístico A de Grimson contando el número de celdas marcadas adyacentes.

Paso 6 : Obtener la significación bajo la hipótesis nula, [10]

En el anexo aparecen las funciones más importantes que se programaron sobre el paquete Mathematica.

5 Algunas aplicaciones en bioinformática

A continuación se mostrarán algunas aplicaciones de ambos métodos modificados en el campo de la bioinformática

5.1 Ejemplo #1: Detección de repeticiones (repeats)

La repetición de un determinado fragmento nucleotídico dentro de la secuencia de un gen, se asocia en ocasiones con enfermedades congénitas. Este fragmento repetitivo, de acuerdo al lugar donde se encuentre, puede afectar la estructura y la expresión del gen o de la proteína para la que este codifica. Existen algoritmos específicos que detectan estas repeticiones automáticamente [23]. Los métodos Scan y Grimson con valores apropiados para sus parámetros pueden también realizar esta acción satisfactoriamente.

La secuencia a analizar se transformará en una compuesta por ceros y unos. Los unos aparecerán en aquella posiciones en las que se encuentre el fragmento nucleotídico buscado.

En ambos ejemplos se analizaron con diferentes parámetros (sugeridos por especialistas), mostrando en las tablas los resultados obtenidos con: 150 unidades para la ventana móvil y una para el paso en la técnica Scan y cuatro unidades como tamaño de las ventanas disjuntas y el mismo valor para el umbral en el método de Grimson [21]. Con estos parámetros se corrieron los métodos implementados en el paquete Mathematica.

1. **La Distrofia Muscular Oculofaríngea** (**OMDF**) es una enfermedad neuromuscular degenerativa, autosómica dominante, que surge en la región 5' del gen **PABPN1** por la expansión del triplete $(\mathbf{GCG})_n$. En los individuos sanos aparece 6 veces y entre 8-13 en los enfermos.

   En la secuencia la presencia del trinucleótido **GCG** se identifica por un uno y el resto de los nucleótidos por un cero, tomando parámetros adecuados, el método Scan y Grinsom solo reporta significación estadística cuando las personas son enfermas [20].

   Obsérvese que, en individuos sanos, el resultado es medianamente significativo en el método Scan ($p = 0.052$) y cuando el individuo está enfermo (7 repeticiones en lo adelante), el test de Scan detecta diferencias significativas. El método de Grimson por su parte, resulta ser más radical, pues ante 6 repeticiones su estadígrafo es 0 y por tanto el resultado no es significativo y cuando la secuencia corresponde a individuos enfermos el resultado es siempre altamente significativo.

   Los resultados expuestos muestran que ambos métodos son capaces de detectar repeticiones anormales de ciertos nucleótidos en secuencias de proteínas.

2. **La Epilepsia Progresiva Miclónica** de tipo Unverricht-Lundborg (**EPM1**) es una enfermedad congénita autosómica recesiva. La causa de esta enfermedad es una mutación en el gen (**CSTB**) que codifica un inhibidor de la cistein proteasa, la cual consiste en la repetición anormal (35-70 veces) del dodecámero (**CCCCGCCCCGCG**) que se encuentra repetido de dos a tres veces en la región 5´ del gen en los individuos sanos [1].

Secuencias	Scan Lineal		Grimson	
	Estadígrafo	p	Estadígrafo	p
$(GCG)_6$	6	0.052	0	0.05
$(GCG)_7$	7	0.019	1	0.00
$(GCG)_8$	8	0.007	1	0.00
$(GCG)_6(GCA)_1(GCG)_1$	8	0.007	1	0.00
$(GCG)_9$	9	0.002	1	0.00
$(GCG)_6(GCA)_1(GCG)_2$	9	0.002	1	0.00
$(GCG)_{10}$	10	0.000	2	0.00
$(GCG)_{11}$	11	0.000	2	0.00
$(GCG)_6(GCA)_1(GCG)_4$	11	0.000	2	0.00
$(GCG)_{12}$	12	0.000	2	0.00
$(GCG)_6(GCA)_3(GCG)_3$	12	0.000	2	0.00
$(GCG)_{13}$	13	0.000	3	0.00

Tabla 1: Resultados de la aplicación de los métodos Scan y Grimson a datos de OMDF.

La secuencia es codificada por un uno los dodecámero y por cero el resto de los nucleótidos, se observan que las técnicas estudiadas dan significación cuando las personas están enfermas siempre que los parámetros sean adecuados.

Secuencias	Scan Lineal		Grimson	
	Estadígrafo	p	Estadígrafo	p
$(CCCCGCCCCGCG)_{\ 2}$	2	0.1332	0	0.5
$(CCCCGCCCCGCG)_{\ 3}$	3	0.0671	0	0.5
$(CCCCGCCCCGCG)_{\ 35}$	35	0.0000	7	0.0
$(CCCCGCCCCGCG)_{\ 70}$	70	0.0000	15	0.0

Tabla 2: Resultados de la aplicación de los métodos Scan y Grimson a datos de EPM1.

En este ejemplo se aprecian resultados similares al ejemplo anterior. Cuando la repetición que se observa es de dos o tres (primera y segunda filas de la Tabla 2) ambos métodos arrojan resultados no significativos. Con repeticiones de 35 y más los resultados son altamente significativos, lo que evidencia la presencia de una elevada repetición de la cadena buscada dentro de la secuencia.

5.2 Ejemplo #2: Alineamientos en el ADN

Con el auge de la biología molecular, las bases de datos de secuencias de ADN crecieron de forma exponencial. Los científicos investigan constantemente similitudes entre diferentes secuencias que pudieran sugerirles igual funcionalidad biológica.

La comparación se puede hacer partiendo de secuencias ortólogas (el mismo gen en

especies diferentes) o parálogas (genes en una misma especie que son resultado de la duplicación). Un problema típico es cuando dos secuencias han sido alineadas siguiendo cierto criterio y entonces el investigador pone en evidencia una coincidencia anormalmente larga de dos sub-secuencias ubicadas en la misma posición relativa dentro de la secuencia.

En este caso se sigue el siguiente procedimiento:

1. Se alinea la secuencia Cytochrome c oxidase subunit 1 [*Arabidopsis thaliana*] con cbb3-type cytochrome oxidase component FixN [*Rhizobium leguminosarum bv. viciae*], que pertenecen a familias diferentes.

2. Se alinean la secuencias Cytochrome c oxidase subunit 1 [*Arabidopsis thaliana*] con Cytochrome c oxidase polypeptide I, que pertenece a la misma familia.

Las secuencias están dadas en formato FASTA como se muestra a continuación:

> **gi**|13449404|ref|NP_085587.1| cytochrome c oxidase subunit 1 [*Arabidopsis thaliana*]
mknlvrwlfstnhkdigtlyfifgaiagvmgtcfsvlirmelarpgdqilggnhqlynvlitahaflmiffmvmpamiggf
gnwfvpiligapdmafprlnnisfwllppsllllssalvevgsgtgwtvypplsgitshsggavdlaifslhlsgvssilgsinfi
ttifnmrgpgmtmhrlplfvwsvlvtafllllslpvlagaitmlltdrnfnttffdpagggdpilyqhlfwffghpevyililpg
fgiishivstfsgkpvfgylgmvyamisigvlgflvwahhmftvgldvdtrayftaatmiiavptgikifswiatmwggsiq
yktpmlfavgfiflftiggltgivlansgldialhdtyyvvahfhyvlsmgavfalfagfyywvgkifgrtypetlgqihfwitf
fgvnltffpmhflglsgmprripdypdayagwnalssfgsyisvvgiccffvvvtitlssgnnkrcapspwalelnsttlew
mvqsppafhtfgelpaiketksyvk

> **gi**|461786|sp|P33517|COX1_RHOSH Cytochrome c oxidase polypeptide I [Cytochrome AA3 subunit 1]
madaaihghehdrrgfftrwfmstnhkdigvlylftgglvglisvaftvymrmelmapgvqfmcaehlesglvkgffqsl
wpsavenctpnghlwnvmitghgilmmffvvipalfggfgnyfmplhigapdmafprmnnlsywlyvagtslavasl
fapggngqlgsgigwvlypplstsesgystdlaifavhlsgassilgainmittflnmrapgmtmhkvplfawsifvtawli
llalpvlagaitmlltdrnfgttffqpsgggdpvlyqhilwffghpevyiivlpafgivshviatfakkpifgylpmvyamva
igvlgfvvwahhmytaglsltqqsyfmmatmviavptgikifswiatmwggsielktpmlwalgfflflftvggvtgivls
qasvdryyhdtyyvvahfhyvmslgavfgifagstsggigkmsgrqypewagklhfwmmfvganltffpqhflgrqgm
prryidypeafatwnfvsslgaflsfasflfflgvifyslsgarvtannywnehadtlewtltspppehtfeqlpkrederapa
h

> **gi**|2114418|gb|AAB58264.1|cbb3-type cytochrome oxidase component FixN [*Rhizobium leguminosarum bv. viciae*]
iatvfwgvvgflvgviialqlafpdlniapylnfgrlrpvhtsavifafggnalimtsfyvvqrtcrarlfggnlawfvfwgyq
lfivmaatgyvlgitqgreyaepewyvdlwltivwvaylavylgtilkrkephiyvanwfylsfivtiamlhvvnnlavpa
sflgsksysvssgvqdaltqwwyghnavgffltagflgmmyyfvpkqanrpvysyrlsiihfwalifmyiwagphhlhy
talpdwaqtlgmvfsimlwmpswggminglmtlsgawdkirtdpiirmmivaiafygmstfegpmmsvktvnsls
hytewtighvhsgalgwvgmitfgaiyyltpklwgrerlyslrmvnwhfwlatfgivvyaavlwvagiqqglmwreyn
sqgflvysfaetvaamfpyyvlravggtlylagglvmawnvfmtirghlrdeaaipttfvpqaqpae

Para realizar el alineamiento se utiliza el procedimiento FASTA disponible on-line en el sitio `http://fasta.bioch.virginia.edu/fasta/lalign.htm`. Cada uno de los

alineamientos se realizó con el esquema de parámetros sugeridos por los especialistas que se muestra en la Tabla 3.

Matriz de peso	Penalización del	
	Gap	Gap Extendido (Ext.)
Pam250	-16	-4
Pam120	-22	-4
Blosum62	-16	-4
Blosum62	-22	-4

Tabla 3: Parámetros utilizados para realizar el alineamiento.

Se cambian por unos todos los aminoácidos que hayan tenido homología y por cero cualquier otro caso. Cada alineamiento se analizó con diferentes parámetros (sugeridos por los especialistas) en ambas técnicas. A continuación, la Tabla 4 muestra los resultados obtenidos con los siguientes parámetros:

- Para el método Scan: 150 unidades para la ventana móvil y una unidad como paso.
- Para el método Grimson, 15 unidades como tamaño de las ventanas disjuntas y cinco unidades para el umbral.

Alineamiento			Scan Lineal		Grimson	
Matrix	Gap	Ext.	Estadígrafo	p	Estadígrafo	p
Alineamiento **a)**						
Blosum62	-16	-4	148	0.528	35	0.999
Blosum62	-22	-4	148	0.535	35	0.999
Pam120	-22	-4	146	0.539	39	0.999
Pam250	-16	-4	148	0.575	35	0.999
Alineamiento **b)**						
Blosum62	-16	-4	83	0.000	13	0.000
Blosum62	-22	-4	83	0.000	13	0.000
Pam120	-22	-4	13	0.000	1	0.000
Pam250	-16	-4	96	0.000	16	0.000

Tabla 4: Resultados de la aplicación de los métodos Scan y Grimson sobre aminoácidos homólogos.

Con el alineamiento **a)** no ocurre lo mismo. Ambas bacterias pertenecen a la misma familia, por lo que existe un gran número de aminoácidos que son homólogos. Tales aminoácidos se distribuyen con cierta uniformidad por toda la secuencia analizada, y ello provoca que resulte prácticamente imposible reconocer un conglomerado. Es por ello que ninguno de los métodos lo detecta. En este caso se podría invertir el estudio y tratar de

Alineamiento			Scan Lineal		Grimson	
Matrix	Gap	Ext.	Estadígrafo	p	Estadígrafo	p
Alineamiento **a)**						
Blosum62	-16	-4	39	0.000	3	0.000
Blosum62	-22	-4	40	0.000	3	0.000
Pam120	-22	-4	42	0.000	2	0.000
Pam250	-16	-4	37	0.000	5	0.000

Tabla 5: Resultados de la aplicación de los métodos Scan y Grimson sobre aminoácidos no homólogos.

encontrar una sub-secuencia "consecutiva" de aminoácidos no similares. En este caso la categoría de interés son los aminoácidos no homólogos, ellos serán representados por uno y cualquier otra combinación será representada por ceros. La Tabla 5 muestra los resultados obtenidos:

Como puede apreciarse, ambas técnicas coincidieron en arrojar resultados altamente significativos para todos los casos considerados.

5.3 Ejemplo #3: Detección de conglomerados de sitios DAM

A la sucesión consecutiva de los cuatro pares de bases **GACT** en el ADN se le llaman sitios **DAM**. Estos sitios tienen una importancia especial en el proceso de reparación o réplica del ADN. Se desea conocer si existen conglomerados de sitios **DAM** en el ADN de *Escherichia Coli*, [22] y [17].

El ADN de la *Escherichia Coli* es circular, por lo que en esta situación se debe utilizar sólo la variante circular del método Scan. Se cuenta con una secuencia de 4.7 millones de pares de bases. La longitud de la ventana que se consideró fue de 250 bases. Se encontraron ocho sitios **DAM** dentro de la secuencia total, la pregunta que se quiere responder es si esos ocho sitios forman un conglomerado o no.

Se aplicó el método Scan sobre un círculo. El promedio de sitios **DAM** dentro de la cadena fue de 1.1. Al realizar los cálculos se obtuvo una $p = 0.66$ por lo que puede concluirse que no existe una aglomeración.

Si en vez de ocho, se hubiesen reportado diez sitios, entonces el resultado sí sería significativo: $p = 0.01$. Ello coincide con resultado obtenidos por otros autores [17].

6 Conclusiones

Los resultados aquí expuestos constituyen un estudio detallado de los métodos Scan y Grimson utilizados para la detección de conglomerados. Basado en ello, se enuncian las siguientes conclusiones:

- Se generalizaron los métodos Scan y Grimson para la detección de conglomerados. Ellos trabajan ahora sobre una secuencia de ceros y unos, creada convenientemente

en función del problema. Ambas técnicas intentan detectar aglomeraciones de unos dentro de la secuencia considerada.

- Los métodos modificados aumentan notablemente el espectro de aplicaciones posibles. A modo de ejemplo se presentan programas en el Mathematica que permiten varias aplicaciones a problemas clásicos de bioinformática.

Referencias

[1] Alakurtti, K.; Weber, E.; Rinne, R.; Theil, G.; Lindhout, D.; Salmikangas, P.; Saukko, P.; Lahtinen, U. (2005) "Loss of lysosomal association of cystatin B proteins representing progressive myoclonus epilepsy, EPM1, mutations", *Hum Genet* **13**: 208–215.

[2] Altschul, S.F. (1996) "Local alignment statistics", *Meth. Enzymol* **274**: 460–480.

[3] Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, L. J. (1997) "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", *Nucl. Acids Res.* **25**: 3389–3402.

[4] Bailey, N. T. J. (1975). *The mathematical theory of infectious diseases and its applications.* Charles Griffin & Company LTD, Second Edition.

[5] Casas, G; Grau, R. (2001) "Validación de dos métodos de detección de conglomerados temporales usando un modelo de epidemia simple", *Investigación Ooperacional* **6**(2): 175–187.

[6] Casas, G. (2003) *Técnicas de Detección de Conglomerados Incluyendo Factores Adicionales.* Tesis de Doctorado, Universidad Central de Las Villas, Cuba.

[7] Casas, G.; Grau, R.; Cardoso, G. (2004) "Introducción de factores de riesgo en los métodos de Knox y Grimson para el estudio de conglomerados espaciotemporales", *Revista de Matemática: Teoría y Aplicaciones* **11**(1): 69–80.

[8] Gentle, J.E.; Hardle, W.; Mori, Y. (2004) *Handbook of Computacional Statistics.* Springer, Heidelberg.

[9] Grimson, R. (1993) "Disease clusters, exact distributions of maxima and p-values", *Statistics in Medicine* **12**: 1773–1794.

[10] Grimson, R.; Rose, R. (1991) "A versatile test for clustering and a proximity analysis of neurons", *Meth. Inform. in Med.* **30**: 299–303.

[11] Grimson, R. (1994) "Disease cluster test based on the maximum occupancy frequency", *Proceedings of the section on Epidemiology, American Statistical Association*: 64–69.

[12] Jacquez, G.; Waller, L. (1964) "Disease cluster statistics for imprecise space-time locations" , *Statistics in Medicine* **15**: 873–85.

[13] Jacquez, G. (1996) "The analysis of disease clusters, part I: state of the art" , *Infection Control and Hospital Epidemiology* **17**(5): 319–327.

[14] Jacquez, G. (1996) "The analysis of disease clusters, part II: introduction to techniques", *Infection Control and Hospital Epidemiology* **17**(6): 385–397.

[15] Jain, A.K.; Murty, M.N.; Flynn, P.J. (1999) "Data clustering: a review", *ACM Computing Surveys* **31**(3): 264–323.

[16] Knox, E. (1964) "The detection of spece-time interactions" , *Appl. Statist.* **13**: 25–29.

[17] Langrand, C. (2005) "Scan Statistics: definición y ejemplos", *Seminario ANY 2005*. Universidad Politécnica de Cataluya. España.

[18] Nagarwilla, N. (1996) "A Scan statistic with a variable window", *Stat. in Med.* **15**: 845–850.

[19] Nauss, J. I. (1982) "Approximations for distributions of Scan statistics", *Journal of the American Statistical Association* **77**: 177–183.

[20] Robinson, D. O.; Hammans, S. R. (2005) "Oculopharyngeal muscular dystrophy (OPMD): analysis of the PABPN1 gene expansion sequence in 86 patients reveals 13 different expansion types and further evidence for unequal recombination as the mutational mechanism", *Hum Genet* **116**(4): 267–271.

[21] Rodríguez, L.; Casas, G.; Grau, R.; Pupo, M. (2006) "Scan Statistics. Bioinformatics Applications", *First International WorkShop on Bioinformatics Cuba-Flanders'2006*, Universidad Central de Las Villas, Santa Clara, Febrero.

[22] Rodríguez, L.; Casas, G.; Grau, R. (2006) "Aplicación de los métodos Scan en Bioinformática" *Uciencia 2006. II Conferencia Científica*, UCI, La Habana, Julio.

[23] Volfovsky, N., Haas, B.J. (2001) "A clustering method for repeat analysis in DNA sequences", *Genome Biology* **2**(8).

7 Anexo: Implementación de las cláusulas fundamentales de los métodos en el paquete Mathematica

7.1 Método Scan sobre una línea

```
Fnn[m_,n_]:=Module[If[n<0,p:=0,p:=CDF[PoissonDistribution[m],n]];Return[N[p,10]]]
Psi[m_,i_]:=Module[{},p:=PDF[PoissonDistribution[m],i];Return[N[p,10]]]
A1[m_,n_]:=2 Psi[m,n] Fnn[m,n-1] ((n-1) Fnn[m,n-2]-m Fnn[m,n-3])
A2[m_,n_]:=0.5Psi[m,n] 2 ((n-1)(n-2)Fnn[m,n-3]-2(n-2)mFnn[m,n-4]+m2 Fnn[m,n-5])
A3[m_,n_]:= Psi[m, 2n-r] Fnn[m, r-1] 2
A4[m_,n_]:= Psi[m, 2n-r] Psi[m, r] ((r-1) Fnn[m, r-2] – m Fnn[m, r-3])
Q2[m_,n_]:= Fnn[m, n-1] 2 –(n-1) Psi[m, n] Psi[m, n-2] – (n-1–m) Psi[m, n] Fnn[m, n-3]
Q3[m_,n_]:= Fnn[m, n-1] 3 – A1[m, n] + A2[m, n] + A3[m, n] – A4[m, n]
Q[m_,n_,L_]:= Q2[m, n] (Q3[m, n] / Q2[m, n])L - 2
Pfinal[Wmax_, λL_,L_]:= 1 - Q[λL, Wmax, L]     (*Calcula probabilidad del estadígrafo*)
(*                    Procedimiento fundamental del Scan Lineal                    *)
ScanValidation[sec_, t_, Paso_]:= CompoundExpression[
    Win=Partition[sec, t, Paso];                (*Determina ventanas según ancho y paso*)
    W=Map[Function[lis, Plus@@lis], Win];       (*Cantidad de unos de cada ventana*)
    Wmax=Max[W];                                (*Calcula el estadígrafo del SCAN*)
    λL=Mean[W];                                 (*Promedio de unos por ventana*)
    T=Length[sec];
    L=T/t;
```

```
    signif=N[Pfinal[Wmax, λL, L],10];          (*Cálculo de la significación según Nauss*)
    Return[signif]
]
```

7.2 Método Scan sobre un círculo

```
Q4[m_,n_]:= (Q3[m, n] 2 / Q2[m, n])
Q[m_,n_,L_]:= Q4[m, n] Q3[m, n] L-2 Q2[m, n] L – 1
Pfinal[Wmax_, λL_,L_]:= 1 - Q[λL, Wmax, L]    (*Calcula probabilidad del estadígrafo*)
(*              Procedimiento fundamental del Scan Circular                    *)
ScanValidation[sec_, t_,Paso_]:=CompoundExpression[
    secc=Join[sec,Take[sec, t-1]];                        (*Convierte la lista en circular*)
    Win=Partition[secc, t, Paso];              (*Determina ventanas según ancho y paso*)
    W=Map[Function[lis, Plus@@lis], Win];       (*Cantidad de unos de cada ventana*)
    Wmax=Max[W];                                  (*Calcula el estadígrafo del SCAN*)
    λL=Mean[W];                                    (*Promedio de unos por ventana*)
    T=Length[sec];
    L=T/t;
    signif=N[Pfinal[Wmax, λL, L],10];          (*Cálculo de la significación según Nauss*)
    Return[signif]
]
```

7.3 Método Grimson

```
Sig[EspA_,VarA_,A_]:=
            Module[{},nd:=CDF[NormalDistribution[EspA,VarA],A];Return[N[nd,10]]]
iter[k_,a_] = k! / a!
GrimsonValidation[sec_,AnchoW_,Umbral_]:=
CompoundExpression[
    Win=Partition[sec,AnchoW,AnchoW-1];          (*Listas según ancho de la ventana*)
    W=Map[Function[lis,Plus@@lis], Win];          (*Suma los unos de cada ventana*)
    c=Length[W];                                        (*Total de celdas o ventanas*)
    temp= Select[W,#1>=Umbral&];                (*# celdas  mayores que un umbral*)
    m=Length[temp];                            (*Total de celdas mayores que un umbral*)
    A=0;
    Table[If[((W[[i]]>=Umbral) && (Win[[i+1]]>=Umbral)),A++],{i,2,c-1}];
    Medy = 2(c – 1) /c;
    Vary = 2(c - 2) / c ^2;
    EspA = (Medy * m * (m - 1)) / (2*(c-1));
    VarA = EspA + (1 + 2(Medy-1) (m-2) / (c–2) + (c*Medy–4 Medy+2)(m–2)(m–3)
           /(2(c-2)(c-3)) – EspA) + Vary (iter[m,3] / iter[c-1,2] - iter[m,4]/ iter[c-1,3]);
    Signif = 1 - Sig[EspA,VarA, A];          (*Cálculo de la significación según Grimson *)
    Return[signif]
]
```