

UNIVERSIDAD CENTRAL “MARTA ABREU” DE LAS VILLAS
FACULTAD DE MATEMÁTICA, FÍSICA Y COMPUTACIÓN
DEPARTAMENTO DE CIENCIA DE LA COMPUTACIÓN



DETECCIÓN DE CONGLOMERADOS EN LA SOLUCIÓN DE PROBLEMAS
BIOINFORMÁTICOS Y BIOMÉDICOS

Tesis presentada en opción del grado científico de Doctor en Ciencias Técnicas

Autor: MSc. Laureano Rodríguez Corvea

Santa Clara, 2010

UNIVERSIDAD CENTRAL “MARTA ABREU” DE LAS VILLAS
FACULTAD DE MATEMÁTICA, FÍSICA Y COMPUTACIÓN
DEPARTAMENTO DE CIENCIA DE LA COMPUTACIÓN



DETECCIÓN DE CONGLOMERADOS EN LA SOLUCIÓN DE PROBLEMAS
BIOINFORMÁTICOS Y BIOMÉDICOS

Tesis presentada en opción del grado científico de Doctor en Ciencias Técnicas

Autor: MSc. Laureano Rodríguez Corvea

Tutores: Dra. Gladys Casas Cardoso
Dr. Ricardo Grau Ábalo

Santa Clara, 2010

Agradecimientos

A mis tutores, Gladita y Grau, por su apoyo incondicional, por ser de los buenos entre los buenos. Por estar siempre a mi lado, en especial Gladita, que me alentó cuando el cansancio asomaba y ayudó a levantar luego de cada tropiezo, por demostrarme que es una amiga especial.

A mis compañeros del departamento y del laboratorio de Bioinformática, porque me han ayudado mucho.

En la revisión de la tesis, agradezco a todos los que me apoyaron, por su preocupación a Vicente, María del Carmen, Isis, Morell, Leticia, Mario, Sadiel, Greta, Yailén y en especial a Ramiro por guiarme y estar a mi lado en todos los momentos, entre otros.

A los estudiantes que han investigado a mi lado, por su ayuda incondicional: Elaine, Leidys, Yunier, Lien y Chalala en su tesis de maestría.

A Alicia y Magalys por estar siempre conmigo en los momentos difíciles.

Le agradezco al proceso revolucionario cubano que me ha ayudado a formar y lograr un resultado como este.

A mis profesores, a la universidad Central “Marta Abreu” de Las Villas, y al proyecto de colaboración con las Universidades Flamencas que apoyaron mi formación investigativa.

Síntesis

El trabajo aborda el tema de la detección de conglomerados de un cierto patrón en secuencias. Esta situación tiene una analogía grande con la detección de epidemias en el tiempo, por lo que las técnicas estadísticas y de inteligencia artificial que se usan para resolver ambos problemas son en esencia las mismas.

Entre la gran cantidad de algoritmos reportados en la literatura para detectar conglomerados, se encuentran los métodos Scan. En la presente tesis se exponen sus fundamentos matemáticos y se realiza un estudio de simulación para analizar su capacidad de respuesta. Basado en estos resultados y en la teoría de la lógica borrosa, se proponen novedosos algoritmos: los métodos Scan Borrosos.

El problema de la selección adecuada de los valores para los parámetros se trata también en los métodos propuestos. Se realizan estudios de simulación sobre secuencias pequeñas (de tamaño 100, 300 y 500) y para complementarlo se ejecuta un diseño experimental no paramétrico sobre secuencias más largas (hasta 1 000 000). Finalmente se propone el uso de un algoritmo bioinspirado para encontrar valores adecuados para los parámetros de los métodos estudiados.

Para concluir se muestran varias aplicaciones en el campo de la bioinformática y en dominios epidemiológicos. Todas ellas se reducen en esencia, a detectar conglomerados de un cierto patrón de secuencias.

En los resultados de simulación y en las aplicaciones reales se pone de manifiesto la superioridad de los métodos borrosos.

Summary

This work addresses the detection of clusters of certain pattern inside sequences. This situation has a great analogy with the detection of epidemics in time, that is why the statistical and artificial intelligence techniques used to solve both problems are essentially the same.

Scan methods can be found among the many algorithms reported in literature to detect clusters. In this thesis we present its mathematical foundations and perform a simulation study to analyze its responsiveness. Based on these results and the theory of fuzzy logic, we propose novel algorithms: The Fuzzy Scan methods.

The problem of properly select the values for the parameters is also addressed in the proposed methods. Simulation studies are conducted on small sequences (size 100, 300 and 500) and as a complement, a non-parametric experimental design was executed over longer sequences (up to 1 000 000). Finally, we propose the use of a bioinspired algorithm to find the appropriate values for the parameters of the studied methods.

To conclude, different applications in the bioinformatics field and in the epidemiologic domain are shown. All of them essentially detect clusters of certain pattern inside sequences.

The results of the simulation as well as the results of the real world applications demonstrated the superiority of the fuzzy methods.

TABLA DE CONTENIDOS

INTRODUCCIÓN	1
CAPÍTULO I. LAS TÉCNICAS DE DETECCIÓN DE CONGLOMERADOS Y LA BIOINFORMÁTICA	9
1.1 Técnicas de detección de conglomerados.....	9
1.1.1 El método Scan sobre una línea	10
1.1.2 El método Scan sobre un círculo	12
1.1.3 Algunas consideraciones sobre los métodos Scan.....	13
1.2 Aplicaciones de técnicas de detección de conglomerados en Bioinformática	13
1.2.1 Estudio de secuencias genómicas.....	14
1.2.2 Problemas bioinformáticos que se resuelven mediante técnicas de conglomerados.....	17
1.3 Introducción a la lógica borrosa	20
1.3.1 Funciones de pertenencia.....	23
1.3.2 Borrosificador.....	25
1.3.3 Desborrosificador.....	26
1.4 Diseño de experimentos bifactorial no paramétrico	27
1.5 Algoritmos bioinspirados	30
1.6 Métodos de Monte Carlo.....	33
1.7 Evaluación de los conglomerados como clasificadores.....	35
1.8 Consideraciones finales del capítulo.....	38
CAPÍTULO II. NUEVOS MÉTODOS DE DETECCIÓN DE CONGLOMERADOS. AJUSTE DE SUS PARÁMETROS.....	40
2.1 Generalización de los métodos de detección de conglomerados.....	40
2.1.1 Algoritmo del método Scan Generalizado sobre una línea	42
2.1.2 Algoritmo del método Scan Generalizado sobre un círculo	43
2.2 Estudio con datos simulados	43
2.2.1 Bases de la simulación realizada.....	43
2.2.2 Resultados y discusión	45
2.2.3 Algunas consideraciones del estudio con datos simulados	49
2.3 Los métodos Scan Borrosos.....	50
2.3.1 El método Scan Borroso sobre una línea	50
2.3.2 El método Scan Borroso sobre un círculo.....	55
2.3.3 Estudios de simulación	56

2.3.4 Validar los resultados de la simulación	60
2.3.5 Algunas consideraciones acerca de los métodos Scan Borrosos	62
2.4 El problema del ajuste de los parámetros	62
2.4.1 Diseño experimental bifactorial no paramétrico	63
2.4.2 Algoritmos bioinspirados: optimización basada en enjambre de partículas....	67
2.4.3 Métodos de Monte Carlo combinado con el PSO y los métodos Scan.....	69
2.4.4 Resumen de recomendaciones para la selección de valores adecuados para los parámetros.....	70
2.5 Análisis del comportamiento de los algoritmos	71
2.6 Consideraciones finales del capítulo	73
CAPÍTULO III. APLICACIONES A PROBLEMAS BIOINFORMÁTICOS Y BIOMÉDICOS	74
3.1 Sobre la implementación de los algoritmos	74
3.2 Problemas sobre orígenes de replicación del ADN	76
3.2.1 Concentraciones de palíndromos en los orígenes de replicación del ADN en herpesvirus.....	77
3.2.2 Patrones específicos alrededor de los orígenes de replicación en bacterias .	81
3.3 Problemas sobre alineamiento de secuencias.....	83
3.4 Problemas sobre detección de conglomerados de enfermos	86
3.4.1. Metodología para la aplicación de los métodos Scan en la detección de conglomerados de enfermos.....	87
3.4.2. Análisis y discusión de las enfermedades estudiadas en Cifuentes.....	90
3.4.3. Consideraciones sobre la detección de conglomerados de enfermos.....	98
3.5 Consideraciones finales del capítulo.....	98
CONCLUSIONES Y RECOMENDACIONES	99
REFERENCIAS BIBLIOGRÁFICAS.....	101
Producción científica del autor sobre el tema de la tesis	112
Anexos	115
Anexo 1: ANOVA bifactorial no-paramétrico.....	115
Anexo 2. Scan Lineal Generalizado.....	117
Anexo 3. Scan Circular Generalizado	118
Anexo 4. Scan Lineal Modificado con verdaderos conglomerados creados con el 10% del tamaño total de la secuencia	119

Anexo 5. Scan Circular Modificado con verdaderos conglomerados creados con el 10% del tamaño total de la secuencia	120
Anexo 6. Scan Lineal Borroso.....	121
Anexo 7. Scan Lineal Borroso con verdaderos conglomerados creados con el 10 % del tamaño total de la secuencia	125
Anexo 8. Scan Circular Borroso con verdaderos conglomerados creados con el 10% del tamaño total de la secuencia	126
Anexo 9. Scan Lineal con verdaderos conglomerados creados con el 5% del tamaño total de la secuencia	127
Anexo 10. Scan Circular con verdaderos conglomerados creados con el 5% del tamaño total de la secuencia	128

INTRODUCCIÓN

La secuenciación de genomas ha generado un amplio catálogo de miles de millones de secuencias de bases nucleotídicas de ADN (Ácido desoxirribonucleico), o de aminoácidos, moléculas esenciales de la vida. Una de las dificultades que se afronta en los estudios de Biología Computacional actualmente proviene de la incapacidad de procesar de manera eficiente esa enorme cantidad de datos. Se conocen las secuencias (nucleotídicas o de aminoácidos para los cuales ellas codifican) de más de un millón y medio de proteínas, de más de cien genomas; la estructura tridimensional de más de 20 mil proteínas, etc. Gracias a los experimentos de matrices de ADN o microarreglos (micro arrays) se sabe cuándo y cómo se expresan muchos genes. Todo el conocimiento científico acumulado a lo largo de las últimas décadas se encuentra disperso en más de 12 millones de artículos (Galperin 2007), cifra que continúa en ascenso (Anderson 2008; Bell et al. 2009; Halevy et al. 2009; Romero 2007; Shamsir y Mohamed Hussein 2010).

La disponibilidad de genomas completos de muchas especies, además del humano, el volumen de información ubicado actualmente en las bases de datos públicas, por ejemplo la base de datos GenBank¹ (Benson et al. 2005) entre otros, han generado un cambio de paradigma en las investigaciones biológicas. De una estrategia de extraer el máximo de información a partir de unos pocos datos, se ha pasado a la necesidad de obtener la información esencial a partir de grandes volúmenes de datos. Para poner un ejemplo, cuando se secuencia un genoma se tiene una larga serie de letras (bases nucleotídicas) (Dopazo y Valencia 2002) que constituyen realmente instrucciones y datos complicados. Para avanzar en la comprensión de la información que encierran estos libros de instrucciones se deben encontrar los genes y predecir su función y esto está lejos de ser resuelto para cualquiera de los genomas ya secuenciados.

Por otra parte, los aportes que el desarrollo de las computadoras ha realizado a la ciencia en general son innegables. Las investigaciones médicas y biológicas no constituyen una excepción (Cheng y Baldi 2005). Los primeros análisis computarizados se centraron en el análisis de secuencias, pero contrario a lo esperado, aún en ese

¹ <http://www.ncbi.nlm.nih.gov/Genbank/GenbankOverview.html>

campo persisten problemas no resueltos.

No cabe duda acerca de la necesidad de la revisión y adaptación o modificación, de algoritmos existentes en los campos de la Inteligencia Artificial y de la Estadística Computacional, como una posible solución al problema del análisis de grandes secuencias de ADN. La capacidad para realizar nuevos descubrimientos biológicos en un futuro no muy lejano, depende en gran medida de las habilidades para combinar o transformar algoritmos y lograr mejorar sus soluciones en el presente. El análisis de grandes bases de datos biológicas, que crecen exponencialmente día a día, requiere cada vez más, del surgimiento y la puesta en práctica de novedosas ideas más que de la aplicación esforzada de los métodos tradicionales (Baldi y Brunak 2001; Cheng et al. 2006).

Antecedentes

Los estudios bioinformáticos que se desarrollan en el mundo tienen mucho de experimental, de uso de métodos de prueba y error y son además muy costosos por los materiales y la información que requieren, tanto para la experimentación biológica como para el procesamiento computacional (Baldi y Pollastri 2003). De una forma u otra, muchos de los problemas de bioinformática se reducen, en última instancia, al descubrimiento de ciertas regularidades en las secuencias genómicas.

La detección de conglomerados de una determinada subsecuencia dentro de una secuencia de ADN mayor, que puede ser incluso un genoma completo, es uno de estos problemas (Durbin et al. 2003). Esta situación tiene una gran semejanza con la detección de epidemias en el tiempo por lo que se comenzará comentando ésta, que se ha trabajado anteriormente.

Los epidemiólogos tienen sus propios métodos de detección de epidemias, de hecho, han probado ser eficientes en numerosas ocasiones; les permiten detectar con cierta precisión la aparición de focos infecciosos, pero no son totalmente confiables y en ocasiones conllevan a cometer errores. Los matemáticos están interesados en redefinir y hacer más precisos esos procedimientos mediante el uso de alguna prueba de significación.

Las mayores dificultades surgen cuando los datos tienen una naturaleza anecdótica. No se trata en estos casos de que no puedan aplicarse pruebas estadísticas para

arrojar un resultado, más bien lo que ocurre es que las pruebas utilizadas hasta el momento quedan invalidadas porque los datos pueden estar sesgados o parcializados en algún sentido. La formulación rigurosa de técnicas estadísticas ayuda, entonces, a los epidemiólogos también en un sentido metodológico, con el fin de lograr datos correctos o al menos seguir un esquema o diseño preconcebido. Si ello se logra, aunque el proceso de recolección no sea perfecto, será posible extraer conclusiones más fidedignas en la medida en que se utilice el aparato matemático más amplia y consecuentemente (Casas 2003; Casas et al. 2004).

En la práctica suele ocurrir que la información disponible no es tan satisfactoria y los datos, aunque quizás sugieran una epidemia, no descartan una incidencia puramente al azar. Es en estos casos en los que se debe esperar que algún test de significación estadística ayude al proceso de toma de decisiones (Bailey 1975). En numerosos trabajos se aborda matemáticamente la detección de focos epidémicos buscando “conglomerados”, entendiendo por conglomerado, aglomeración o cluster de enfermos a un exceso de casos diagnosticados con respecto a cierto patrón previamente predefinido.

El mismo problema extrapolado al dominio de la Bioinformática consiste en la aplicación de métodos estadísticos (u otros similares) que busquen conglomerados dentro de secuencias de ADN. La aparición de tales aglomeraciones tiene una importancia bioquímica determinada, que ayudan a enriquecer el conocimiento que se tenga de la secuencia o del genoma analizado.

Las técnicas que detectan focos epidémicos trabajan con fecha ordenadas. Las secuencias de ADN tienen un orden que no puede ser cambiado, pero sus elementos no son fechas sino posiciones en el espacio, en principio lineal, si hablamos de estructura primaria, pero podrían ser bidimensionales o espaciales. De cualquier manera los métodos de detección de conglomerados deben ser modificados para que puedan ser aplicados en contextos bioinformáticos u otros cualesquiera más allá de los estudios epidemiológicos para los que fueron concebidos.

Situación problemática

La existencia de patrones repetitivos en una secuencia de ADN, en un cromosoma o en un gen en particular, ayuda a la interpretación de propiedades biológicas. Los

datos obtenidos a partir de la secuenciación del genoma humano proporcionan un conocimiento de la organización esencial de los genes y de los cromosomas. Muchos científicos creen que la identificación de la dotación genética humana revolucionará el tratamiento y prevención de numerosas enfermedades humanas, ya que penetrará en los procesos bioquímicos básicos que las sustentan².

Lo que se dice para el genoma humano, es de interés también para los genomas de muchas especies, animales o vegetales, o de microorganismos, porque en última instancia todos ellos pueden ser importantes para el hombre. Para ayudar a los investigadores a determinar el sentido de este aluvión de datos, se utilizan, cada vez más, instrumentos informáticos, como sistemas de información y de gestión de bases de datos e interfaces gráficas de usuario, sistemas estadísticos y algoritmos inteligentes, entre muchos otros.

Por otra parte, la ausencia de determinismo en muchos procesos biológicos sugiere inmediatamente el uso de lógica borrosa. La teoría de la lógica borrosa ha constituido toda una revolución en el campo de las matemáticas, (Zadeh 1986; Zadeh 2002; Zadeh 2004). Se han formalizado nuevas disciplinas como la teoría de control borroso, las probabilidades y la estadística borrosa, la optimización borrosa, por mencionar algunas. El cúmulo de aplicaciones también ha crecido de manera notable en los últimos años y sigue en ascenso. La Bioinformática es una ciencia, que aunque nueva, también se ha revolucionado, en los últimos años utiliza y desarrolla muchos métodos computacionales, entre los que se destacan las técnicas de aprendizaje computarizado (Baldi y Brunak 2001).

Todas las técnicas de aprendizajes computarizado, supervisado o no, tienen ventajas y desventajas, en 1997 Wolpert y Macready en el teorema "no free lunch", establecen el principio que ningún sistema de aprendizaje es superior en su desempeño a otro (Wolpert 1996; Wolpert y Macready 1997; Wolpert y Macready 2005). Una tarea siempre interesante es estudiar a fondo sus limitaciones para realizar transformaciones que deriven en algoritmos más eficientes, al menos para problemas específicos.

Las técnicas existentes actualmente para la búsqueda de patrones repetitivos no incluyen técnicas estadísticas clásicas de detección de conglomerados, adaptadas

² Encarta 2009 Microsoft ® Encarta ® 2009. © 1993-2008 Microsoft Corporation. Reservados todos los derechos

convenientemente para el análisis de secuencias biológicas. Tampoco se ha investigado si la adecuación de estas técnicas con elementos de lógica borrosa mejora los resultados, pero es presumible por la mencionada ausencia de determinismo en los datos biológicos. Estas son las primeras interrogantes a responder con la presente investigación.

Otro problema radica en la detección adecuada de los valores de los parámetros que intervienen en el modelo que se utilice. Generalmente los parámetros de los métodos estadísticos los selecciona un investigador experto en el tema. En ocasiones esta tarea resulta ser muy difícil, incluso para un especialista en la temática. Valores incorrectos pueden conducir a resultados erróneos y si se habla de detección de conglomerados, tales errores suelen detectar falsos conglomerados, o no detectar los verdaderos. Hasta qué punto el uso de la lógica difusa puede ayudar en el proceso de selección adecuada de los parámetros es otra pregunta de investigación que trataremos de abordar en el presente trabajo.

Objetivo general

Incorporar elementos de la lógica borrosa a los métodos epidemiológicos clásicos de detección de conglomerados para obtener algoritmos más eficientes que los existentes en el análisis de secuencias y en otros problemas biomédicos.

Este objetivo general se desglosa en los siguientes objetivos específicos:

- Desarrollar nuevos algoritmos de detección de conglomerados que puedan ser aplicados en la solución de problemas bioinformáticos y biomédicos con eficiencia similar o superior a los ya existentes.
- Realizar un estudio de los parámetros para sugerir, dado un problema, valores adecuados para los mismos.
- Realizar la implementación computacional de los métodos propuestos en plataformas de software libre, de modo que se facilite su utilización práctica por la comunidad científica internacional, y a su vez se puedan comparar con las alternativas clásicas.

Para el cumplimiento de estos objetivos se trazaron las siguientes:

Tareas de investigación

1. Confeccionar el marco teórico relacionado con la teoría de las técnicas de detección de conglomerados y sus aplicaciones. Revisar detalladamente la fundamentación matemática de los métodos a modificar y los elementos esenciales de la teoría de la lógica borrosa y otros temas matemáticos que ayudarán a formalizar la nueva propuesta.
2. Desarrollar y formalizar nuevos algoritmos de detección de conglomerados.
3. Implementar las nuevas contribuciones en un paquete utilizando lenguaje de software libre como Java.
4. Validar su superioridad.
5. Realizar un estudio de los parámetros de los métodos con el fin de brindar sugerencias efectivas de sus posibles valores para maximizar su efectividad.
6. Mostrar y evaluar los resultados de la aplicación en problemas tales como:
 - a. Detección de orígenes de replicación.
 - b. Concentración de gaps (huecos) en el alineamiento de secuencias.
 - c. Detección de focos de enfermedades.

Novedad Científica

La novedad científica y el consecuente valor teórico del presente trabajo se resumen en los siguientes puntos:

1. Se desarrollan y formalizan nuevos algoritmos para la detección de conglomerados en secuencias lineales, así como en secuencias circulares, tales como los genomas mitocondriales.
2. Se establecen reglas para determinar los valores adecuados para los parámetros de los métodos desarrollados.
3. Se muestran nuevos enfoques para afrontar problemas aún no resueltos cabalmente en Bioinformática, relacionados por ejemplo, con los orígenes de réplicas, y la concentración de pares de bases con importancia biológica. Se ilustra

además la generalidad de los enfoques para dar solución a otros problemas de la ciencia, como por ejemplo detección de epidemias de personas enfermas.

La novedad está avalada por las publicaciones que se describen al final de la tesis.

Valor práctico

La disponibilidad de la implementación de los nuevos algoritmos en plataformas de software libre, facilita su uso inmediato y generalizado por la comunidad científica bioinformática, pero además, posibilita su comparación con otros algoritmos previamente desarrollados o por desarrollar para la solución de problemas similares, tanto en bioinformática como en otras áreas de aplicación.

Hipótesis de investigación

Después de la revisión de la literatura y el desarrollo consecuente del marco teórico se formularon las siguientes hipótesis de investigación:

Combinando elementos de la lógica borrosa con métodos epidemiológicos clásicos se pueden definir nuevos algoritmos de detección de conglomerados que tienen una eficiencia similar o superior a los descritos en la literatura.

Con ayuda de la simulación, de métodos de diseño de experimentos bifactoriales no paramétricos y de métodos de optimización bioinspirados, se pueden formular reglas de ayuda en la adecuada selección de los valores de los parámetros de las técnicas de detección de conglomerados estudiadas.

Estructura de la tesis

El trabajo se presenta esencialmente en tres capítulos a partir de la presente Introducción.

El Capítulo I se dedica a la elaboración del marco teórico desde el punto de vista de las tendencias actuales en el desarrollo y evaluación de los conglomerados. Se muestran algunas aplicaciones interesantes de estas técnicas, especialmente en el campo de la Bioinformática.

En el Capítulo II se propone y formaliza matemáticamente la generalización de los métodos Scan y se realiza un intenso estudio de simulación. Se modifican los métodos epidemiológicos clásicos de detección de conglomerados, introduciéndole elementos de lógica borrosa para mejorar su desempeño. Se realiza además un estudio de la influencia de los valores de los parámetros con ayuda del diseño factorial no paramétrico y de un método de optimización bioinspirado. Por último se presenta un análisis de la complejidad temporal de los algoritmos propuestos.

El Capítulo III está dedicado a mostrar el comportamiento de los nuevos algoritmos en dos problemas bioinformáticos y en la predicción de focos de enfermos como ejemplo de aplicación en otras ramas.

Finalmente, se formulan las conclusiones y recomendaciones, y se muestran las referencias bibliográficas y anexos con detalles complementarios.

CAPÍTULO I. LAS TÉCNICAS DE DETECCIÓN DE CONGLOMERADOS Y LA BIOINFORMÁTICA

El presente capítulo se dedica a sustentar teóricamente el tema de la tesis, por lo que se analizan aquellos enfoques y antecedentes relacionados con las técnicas de detección de conglomerados y su aplicación, por ejemplo a problemas bioinformáticos de análisis de secuencias. Se provee un marco de referencia teórico de diferentes aspectos utilizados para dar solución a algunas de las limitaciones de los algoritmos, para lograr mejores soluciones a los problemas desarrollados, y se exponen los conceptos relacionados con la Bioinformática en función de las aplicaciones que se desarrollan. Se analizan los problemas actuales existentes en esta temática y la posible aplicación en problemas de la vida real.

1.1 Técnicas de detección de conglomerados

Según la literatura especializada de epidemiología se denomina conglomerado o cluster a un exceso de casos de enfermos diagnosticados superior a lo esperado en un área geográfica determinada (conglomerado espacial), en un período de tiempo limitado (conglomerado temporal), o considerando ambos dominios (conglomerado espacio-temporal). La detección de los conglomerados de enfermos es un problema epidemiológico en el que se ha venido trabajando desde hace relativamente poco tiempo (Jacquez et al. 1996a). Las primeras publicaciones al respecto aparecieron en 1964 por Knox (1964) y a partir de esa fecha han tenido un incremento exponencial (Jacquez y Waller 1996; Jacquez et al. 1996b).

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 (Jain et al. 1999).

Tampoco existen técnicas globales que puedan aplicarse a todas las situaciones, por eso hay gran diversidad de métodos con la misma finalidad. En un estudio preliminar de las técnicas de detección de conglomerado, se eligió una de las más populares y sobre ella se trabajó: el método Scan (Naus 1965) porque trabaja sobre una línea, en

principio temporal, pero que puede extenderse al sentido espacial (Rodríguez et al. 2008b).

1.1.1 El método Scan sobre una línea

Los métodos Scan en general se utilizaron inicialmente para detectar aglomeraciones dentro de períodos de tiempo consecutivos, pues puede suceder que un conglomerado temporal se extienda por dos o más intervalos. (Jacquez et al. 1996b). Todos los casos diagnosticados deben estar ordenados cronológicamente de acuerdo con la fecha de los primeros síntomas o de diagnóstico de la enfermedad, de muerte o cualquier otro evento de salud que se considere.

Sean $X_1, X_2, \dots, X_n$ variables aleatorias independientes e idénticamente distribuidas que denotan las fechas de ocurrencias de n eventos en el intervalo $[0, T]$. Se quiere probar la hipótesis nula de que los eventos están uniformemente distribuidos contra la alternativa de que existe un conglomerado dentro de algún subintervalo de $[0, T]$ (Nagarwilla 1996).

Se define en el método, un intervalo o una ventana, de tamaño fijo de acuerdo con la duración esperada de la epidemia. Para evitar subjetividad, esto debe hacerse con criterios epidemiológicos antes de inspeccionar los datos recolectados. La ventana seleccionada se desplaza a lo largo de la línea del tiempo y se determinan en cada caso, la cantidad de enfermos asociados a ella, (Aldrich y Wanzer 1993).

Para la formulación más precisa, sean:

t : amplitud de la ventana.

T : período de tiempo total que se analiza.

$L = T/t$: fracción que representa el período de tiempo total que se analiza con relación al ancho de la ventana.

n : cantidad de enfermos diagnosticados en T .

λ : número esperado de casos por unidad de tiempo en un proceso de Poisson.

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

Hipotéticamente el estadístico: $\eta = \max_{0 \leq y \leq T-t} \{w_{y, y+t}\} \in Z^+$ representa el mayor número de casos que aparecen en una ventana cuando se mueve continuamente a lo largo del tiempo. En la práctica, la ventana $[y, y+t)$ se mueve discretamente a partir de una sucesión de puntos equidistantes $y_1, y_2, \dots, y_k$ que cubren todo el período de análisis de amplitud T . Se denomina paso del Scan o paso del desplazamiento a $\Delta y = y_k - y_{k-1}$.

Realmente, el estadístico anterior se estima por su versión discreta:

$$\eta' = \max_{1 \leq i \leq kt} \{w_{y_i, y_{i+t}}\}$$

La idea del método es que si existe un conglomerado el número máximo de casos hallados en una ventana debe ser grande con respecto a los demás valores. El test estadístico depende de varios de los parámetros explicados con anterioridad 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 Naus (1982):

$$p = P^*(\eta, \lambda L, 1/L) = 1 - Q^*(\eta, \lambda L, 1/L) \quad (1.1)$$

donde Q^* puede ser aproximado para cualquier $L > 2$ a partir de sus valores con $L = 2$ y $L = 3$.

$$Q^*(\eta, \lambda L, 1/L) \approx Q^*(\eta, 2\lambda, 1/2) [Q^*(\eta, 3\lambda, 1/3) / Q^*(\eta, 2\lambda, 1/2)]^{L-2} \quad (1.2)$$

La aproximación (1.2) es fácilmente calculable usando una microcomputadora personal. El cálculo exacto de $Q^*(\eta, 2\lambda, 1/2)$ y $Q^*(\eta, 3\lambda, 1/3)$ se basa en un teorema demostrado también por Naus (1982) y cuya esencia se resume aquí:

Para $\eta > 2$, $p_i = e^{-\lambda} \lambda^i / i!$, $F_\eta = \sum_{i=0}^{\eta} p_i$, $\lambda > 0$, se tiene que:

$$Q^*(\eta, 2\lambda, 1/2) = F_{\eta-1}^2 - (\eta-1)p_\eta p_{\eta-2} - (\eta-1-\lambda)p_\eta F_{\eta-3} \quad (1.3)$$

$$Q^*(\eta, 3\lambda, 1/3) = F_{\eta-1}^3 - A_1 + A_2 + A_3 - A_4 \quad (1.4)$$

donde:

$$A_1 = 2p_\eta F_{\eta-1} ((w-1)F_{\eta-2} - \lambda F_{\eta-3}) \quad (1.5)$$

$$A_2 = 0.5 p_{\eta}^2 ((\eta - 1)(\eta - 2) F_{\eta-3} - 2(\eta - 2) \lambda F_{\eta-4} + \lambda^2 F_{\eta-5}) \quad (1.6)$$

$$A_3 = \sum_{r=1}^{\eta-1} p_{2\eta-r} F_{r-1}^2 \quad (1.7)$$

$$A_4 = \sum_{r=2}^{\eta-1} p_{2\eta-r} p_r ((r-1) F_{r-2} - \lambda F_{r-3}) \quad (1.8)$$

con $F_i = 0$ para todo $i < 0$.

La aproximación (1.2) puede calcularse para valores no enteros de L . Esto la diferencia de otras expresiones matemáticas que se usaban con estos fines anteriormente. Además de ser menos restrictiva, varios autores demuestran que (1.2) es mucho más precisa (Glaz 1993; Naus 1982; Sahu et al. 1993).

1.1.2 El método Scan sobre un círculo

Este método es una variación del anterior y se utiliza para enfermedades que tengan un comportamiento estacional. Los datos se encuentran ordenados cronológicamente a lo largo de la línea del tiempo y el círculo se forma uniendo la última fecha con la primera. En epidemiología tiene mucho sentido, para estudiar conglomerados de enfermedades que pueden tener un carácter periódico.

La ventana se desplaza sobre el círculo y se determina en cada una, la cantidad de enfermos 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 “los del siguiente período”. En el caso bioinformático ello tiene mucho sentido en el estudio de genomas circulares, por ejemplo, los genomas mitocondriales (Mott y Berger 2007; YU et al. 2004).

Según Naus (1982) la probabilidad de observar η o más casos en un intervalo o ventana de tamaño fijo en el caso circular se estima por:

$$p = P_c^*(\eta, \lambda L, 1/L) = 1 - Q_c^*(\eta, \lambda L, 1/L) \quad (1.9)$$

donde ahora:

$$Q_c^*(\eta, \lambda L, 1/L) \approx Q^*(\eta, 4\lambda, 1/4) [Q^*(\eta, 3\lambda, 1/3)]^{L-2} [Q^*(\eta, 2\lambda, 1/2)]^{L-1} \quad (1.10)$$

Para hallar $Q^*(\eta, 4\lambda, 1/4)$ se utiliza $L=4$ en (1.2). Después de simplificar se obtiene:

$$Q^*(\eta, 4\lambda, 1/4) \approx [Q^*(\eta, 3\lambda, 1/3)]^2 / Q^*(\eta, 2\lambda, 1/2) \quad (1.10)$$

Luego $Q^*(\eta, 4\lambda, 1/4)$ queda en función de $Q^*(\eta, 2\lambda, 1/2)$ y de $Q^*(\eta, 3\lambda, 1/3)$, valores que se calculan utilizando las funciones (1.3) y (1.4) respectivamente.

1.1.3 Algunas consideraciones sobre los métodos Scan

Como se ha visto, la probabilidad p hallada para un conjunto particular de casos, depende del ancho de la ventana y del paso del Scan seleccionados por el investigador. Resulta imposible determinar los valores ideales para cada enfermedad, por lo que se recomienda realizar varias repeticiones del método utilizando amplitudes diferentes (Aldrich y Wanzer 1993; Kulldorff 2001).

Algunos autores han tratado de modificar el método Scan de diferentes formas. Por ejemplo, el método no es válido cuando los factores de riesgos de población varían. Martín (1981) siguió una estrategia de generalización que resuelve este problema.

Se han realizado esfuerzos para aumentar el dominio de aplicación del Scan a dos y a tres dimensiones. Con dos dimensiones se pueden detectar conglomerados geográficos, (Kulldorff 1997; Kulldorff 1999; Kulldorff 2001; Kulldorff et al. 2007) mientras que con tres la detección puede ocurrir en el espacio-tiempo. (Kulldorff 1998).

Además se han publicado otras variantes, como es el caso de la versión nombrada r-Scan que trabaja con casos y controles y que utiliza la distribución de Bernoulli para el cálculo de su significación (Dembo y Karlin 1992; Glaz et al. 1994).

En este trabajo se realiza una generalización del método Scan en sus dos variantes: lineal y circular, para encontrar conglomerados no sólo de enfermos, sino de cualquier categoría de interés en cualquier rama de la ciencia.

1.2 Aplicaciones de técnicas de detección de conglomerados en Bioinformática

La detección de cierta sucesión inhabitual de un mismo suceso a lo largo del tiempo (conglomerado) está presente en numerosos problemas de la ciencia, no sólo en epidemiología. Podrían conjeturarse ejemplos relacionados con ocurrencia de

accidentes del tránsito, de huracanes, de huelgas, de otros eventos de salud, educacionales, económicos o sociales entre muchos otros. Los métodos de detección de conglomerados han sido ya aplicados con éxito en diversas áreas, donde el problema está relacionado con la fecha de aparición del evento, no sólo para la detección de epidemias o pandemias de enfermos, también por ejemplo, en centrales telefónicas con el fin de poder absorber clusters de llamadas simultáneas, o en el control de calidad examinando conglomerados de objetos defectuosos en una cadena de producción, entre otros (Langrand 2005). La Bioinformática no se exceptúa como campo de aplicación, pues en muchas situaciones se necesita conocer si existe agrupamiento de una base, conjunto de bases, aminoácidos (en general subsecuencias específicas) (Pupo et al. 2006), en una secuencia genómica más larga.

Como se ha planteado anteriormente, el desarrollo alcanzado por las ciencias biológicas ha permitido la acumulación de mucha información experimental disponible en grandes bases de datos. La secuenciación del ADN (Benson et al. 2005; Consortium 2004), produjo un crecimiento exponencial de las descripciones lineales de proteínas y moléculas de ADN y ARN (ácido ribonucleico) y planteó los problemas informáticos de interés biológico: el almacenamiento y manejo eficiente de la información y la extracción de información útil para en última instancia, comprender las relaciones entre los genes, las proteínas, la funcionabilidad, la vida y la salud. La Bioinformática constituye el campo de conocimientos multidisciplinario entre la biología, la informática, estadística y la matemática que debe abordar este problema. En ella surge en particular, la necesidad de desarrollar herramientas útiles que ayuden a comprender el flujo de información desde los genes a las estructuras moleculares, a sus funciones bioquímicas, a su importancia biológica, y finalmente, a su influencia sobre las enfermedades y la salud, para en definitiva, mejorar la vida.

1.2.1 Estudio de secuencias genómicas

Hoy en día hay muchas especies de las cuales ya se ha obtenido su genoma completo y estos son relativamente fáciles de acceder a través de diferentes páginas de la WEB. Dentro de las más importantes internacionalmente, están: GenBank, EMBL³, PIR-

³ <http://www.ebi.ac.uk/embl/index.html>

International Protein Sequence Database (PSD)⁴, SwissProt⁵ y DNA DataBank of Japan (DDBJ)⁶, de donde se pueden descargar los ficheros en formato FASTA. Se trata de ficheros de texto que contienen entre otras informaciones una largas secuencias formadas por combinaciones de 4 letras, correspondientes a las iniciales de los 4 nucleótidos presentes en el ADN (A = Adenina, C = Citosina, G = Guanina y T = Timina). En el caso de secuencias de ARN, esta última se sustituye por U = Uracilo). En una primera observación, estas secuencias parecen generadas aleatoriamente, pero no es así, pueden encontrarse patrones de repetición o de ausencia que aporten información valiosísima sobre su contenido, como por ejemplo, la localización de los genes, que son secuencias únicas en un cromosoma.

El modo más confiable de determinar la estructura de una molécula biológica grande o las funciones de la misma es por la experimentación directa, pero debido al costo y tiempo requerido para procesar este gran cúmulo de información, es necesario automatizar el análisis in silico de las secuencias de aminoácidos o de bases nucleotídicas que codifican para ellos. Esto requiere un conocimiento amplio de la biología celular y del organismo, ya que se debe organizar, clasificar y analizar la riqueza inmensa de los datos de la secuencia. Esto es más que una tarea abstracta de análisis, ya que detrás de las bases nucleotídicas o aminoácidos está la complejidad total de la biología molecular. Por ello hay que crear métodos robustos, escalables y confiables, que sean en principio capaces de capturar esa complejidad, integrando fuentes de diversas informaciones biológicas en limpios, generales y manejables modelos, para el análisis de secuencias.

La mayor parte de los problemas en el análisis computacional de secuencias son esencialmente estadísticos. Fuerzas estocásticas evolutivas actúan sobre el genoma distinguiendo semejanzas o diferencias significativas entre secuencias que divergen entre un caos de mutaciones aleatorias, la selección natural, y el flujo genético, presentan las señales específicas al problema. Muchos de los métodos más poderosos de análisis disponibles usan entre otras técnicas, la teoría de las probabilidades. Entre los llamados modelos probabilísticos, pueden citarse en particular, los modelos ocultos

⁴ <http://pir.georgetown.edu/>

⁵ <http://www.expasy.ch/sprot/>

⁶ <http://www.nig.ac.jp/home.html>

de Markov (HMMs) que proporcionan una estructura general para el análisis estadístico de una amplia variedad de problemas de análisis de secuencias, pero hay realmente una gama no estrecha de modelos grafo-probabilísticos para resolver tareas de este tipo (Janssens et al. 2005).

Aunque el análisis de secuencias genómicas depende del problema a dar solución, es importante destacar que la comparación de diversas secuencias utilizando alineamientos⁷ es la tarea de mayor madurez y aplicabilidad en Bioinformática. No sólo es necesaria la comparación de dos secuencias (pairwise alignment) sino la de múltiples (multialignment). El problema del alineamiento de secuencias, es esencialmente un problema matemático de programación dinámica (Giegerich 2000), y para salvar su complejidad computacional no polinómica, se usan técnicas heurísticas que defienden diferentes formas de hacer las comparaciones (distancias) entre secuencias. Entre los algoritmos más populares para hacer multialineamiento de secuencias, se encuentran los denominados Needleman-Wunsch, Smith-Waterman, BLAST (Basic Local Alignment Search Tool)⁸ y FASTA (FAST-All, (EBI 1999))⁹.

No obstante, los alineamientos aún no son perfectos y se siguen buscando algoritmos con mayores niveles de selectividad y confianza (Lambert et al. 2003), profundizándose cada vez más en la evolución histórica de las moléculas biológicas, sus estructuras tridimensionales, y otros rasgos que obligan la evolución de la secuencia primaria. Como se ha comentado, dentro de los alineamientos, el que más atención recibe hoy en día, es el de comparación múltiple, que es frecuentemente utilizado para caracterizar familias de proteínas, para predecir plegamientos (estructuras secundarias y terciarias) y su funcionalidad. Estas aplicaciones son un elemento clave que interesa por ejemplo, a las empresas diseñadoras de fármacos, porque les facilita in silico la efectividad de nuevas drogas posibles (Marrero-Ponce et al. 2006; Notredame 2002; Pérez et al. 2006; Rivera-Borroto et al. 2008).

⁷ Alineamiento: Dos o más secuencias supuestamente similares ordenadas entre las partes que realmente juegan el mismo rol, introduciendo, si es necesario en las secuencias, "gaps" para lograr desplazamientos adecuados a la derecha o la izquierda de zonas reconocibles.

⁸ BLAST se utiliza para buscar regiones similares entre secuencias biológicas.

⁹ FASTA permite hacer una comparación rápida de proteínas o nucleótidos.

Existe otra amplia gama de problemas que pueden resolverse buscando patrones específicos en la secuencia de ADN, como son por ejemplo codones de inicio y terminación, patrones de secuencias en puntos de splicing, zonas de promotores, regiones no traducidas (UTP) entre otros (Boutros 2006; Wang et al. 2004). La detección de estos patrones determina la existencia o no de alguna función general o específica del genoma, y se realiza con ayuda de herramientas algorítmicas y computacionales.

Entre las técnicas más exitosas hoy en día se utilizan las cadenas ocultas de Markov (Baldi y Brunak 2001; Delvin 2006; Durbin et al. 2003; Prinzie y Vanden 2007), las redes neuronales (Bonet et al. 2007; Bonet et al. 2008; Chávez et al. 2007b; Chávez et al. 2008b; Rodríguez y Bonet 2007) las máquinas de vectores de soporte (Support Vector Machines (SVM) (Jaronski et al. 2005; Rodríguez et al. 2006; Rodríguez et al. 2007a; Vanhulsel et al. 2009) y hasta otras herramientas que no son exactamente de aprendizaje supervisado o no, por ejemplo de aprendizaje reforzado (Peeters et al. 2008).

Los métodos de detección de conglomerados por su parte, no constituyen una excepción en las aplicaciones bioinformáticas como se mostrará en el siguiente epígrafe. Es en este contexto donde desarrolla la presente tesis.

1.2.2 Problemas bioinformáticos que se resuelven mediante técnicas de conglomerados

Las técnicas de detección de conglomerados son valiosas en estudios bioinformáticos, siempre que sea necesario comprobar que un patrón determinado se encuentre repetitivamente en una secuencia. Técnicas de detección se aplican en la actualidad en los campos de la genética, la genómica y los sistemas biológicos entre otros. Ejemplos concretos lo constituye la detección de orígenes de replicación, de genes, o de anomalías repetitivas en las secuencias que caracterizan algunas enfermedades genéticas (Iliende et al. 2007). En el trabajo se mencionan algunos ejemplos que la literatura recoge en esta área.

En (Masse et al. 1992; Reisman et al. 1985; Weller et al. 1985) se reportaron altas

concentraciones de palíndromos¹⁰ en la proximidad de los orígenes de las repeticiones de herpes virus. Por otro lado si se conoce las localizaciones de las réplicas originales de los virus se puede reforzar el desarrollo de agentes antivirales bloqueando la repetición de ADN viral o interviniendo en el proceso de infección.

Basados en estos hechos en (Leung et al. 2005) se realiza un análisis de una colección de genomas de 16 herpesvirus. Se identifican las regiones que contienen conglomerados significativos de palíndromos y se comparan con las posiciones **conocidas** de los orígenes de las replicación. En este momento sólo se conocían orígenes de diez herpes virus.

Para el estudio se procede de la siguiente forma:

- Se escoge una cota superior de la longitud de los palíndromos de cada uno de los herpes virus utilizando la distancia de Wasserstein entre el proceso de palíndromos y el proceso de Poisson. Se procede entonces a buscar los palíndromos de cada uno de los herpes virus estudiados.
- Formada la secuencia de cada herpes virus según sus palíndromos se procede a calcular los conglomerados significativos de palíndromos utilizando el método r-Scan, que a continuación se describe brevemente.
 - o Modelo basado en la distribución de Bernoulli (Kulldorff 1997) para datos de eventos individuales o por individuos (1 y 0 para identificar casos y controles).
 - o Dada la secuencia $x_1, x_2, \dots, x_n$ con m casos y $n-m$ controles, entonces $X(i)$ representa la posición del caso i en la secuencia anterior.
 - o Sea $S_i = X(i+1) - X(i)$. Para $i = 1, \dots, m-1$
 - o Para un entero fijo $r \in [1, m-1]$ y $i = 1, \dots, m-1$ entonces:

$$Ar = \min(Ar(i)) \quad \text{donde} \quad Ar_{(i)} = \sum_{j=1}^{i+r-1} S_j \quad (\text{Dembo y Karlin 1992})$$
 - o Para calcular la significación de Ar Glaz (1994) propone la siguiente aproximación:

¹⁰ Los palíndromos son palabras simétricas de ADN en el sentido que ellos pueden leerse exactamente igual que leyendo las secuencia complementarias en la dirección inversa

$$P(Ar \leq w) \approx 1 - \text{Exp}\left\{-(m-r)\pi(1-p+p^r(r+p-rp))\right\}$$

donde:

$$\pi = Q_1$$

$$p = 1 - \frac{Q_2}{Q_1}$$

- $Q_1 = \sum_{j=r}^m B(j; m, w)$
- $Q_2 = \sum_{j=r}^m (-1)^{r+j} B(j; m, w)$
- $B(j; m, w) = \binom{m}{j} w^j (1-w)^{m-j}$

- El método r-Scan está implementado en el software SaTScan, que es un programa desarrollado para analizar datos de eventos de salud en tiempo, espacio y espacio-tiempo utilizando el estadístico Scan (Martínez-Piedra et al. 2004). Utiliza dos tipos de modelos diferentes: el tradicional basado en Poisson, y otro basado en Bernoulli (r-Scan).

En el capítulo III se muestra un estudio comparativo de estos resultados con otros obtenidos a partir de los métodos propuestos en esta tesis.

Otras investigaciones sobre palíndromos en secuencias de ADN describen diferentes efectos biológicos que producen los mismos, tales como:

- La distribución no aleatoria de palíndromos en el cáncer (Cromie et al. 2000; Leach 2005; Tanaka et al. 2005; Vasconcelos et al. 2000).
- La longitud de los palíndromos influye en la inestabilidad genética y en la estructura reparativa del ADN (Leach 2005; Neiman et al. 2008).
- Los palíndromos pueden servir como el factor de transcripción, por ejemplo concentración de los mismos en los intrones y déficit en los exones (Lu et al. 2007), entre otros.

Una aplicación diferente en este campo, es la localización de las llamadas “islas” **C_pG** frecuentemente se escribe C_pG para distinguir el par de bases C-G en ambas hélices del ADN (Durbin et al. 2003). El dinucleótido menos frecuente en muchos genomas es CG, aun cuando se tenga en cuenta las probabilidades, independientes de las de C y la G. La razón para esto, es que la Citosina es fácilmente metilada cuando precede a

Guanina y el resultado del metilo - Citosina tiene una tendencia a mutar en Timina Figura 1.1 (Delvin 2006). Por razones biológicamente importantes el proceso de metilación se inhibe en cadenas pequeñas del genoma, como es por ejemplo alrededor de los promotores o 'en el principio' de las regiones de muchos genes con el objetivo de intervenir entre otros en el proceso de replicación y de transcripción de los genes de muchas especies (Durbin et al. 2003). En fin, a estas áreas se les llama **islas** C_pG (Bird 1987), y en ellas el dinucleótido CG aparece frecuentemente. Un problema importante es definir y ubicar las islas C_pG en un texto genómico amplio (Durbin et al. 2003).

Muchos autores han usado islas C_pG como marcadores genéticos para identificar: - sitios de rupturas y réplicas del ADN (Ponger y Mouchiroud 2002; Prioleau 2009), - para reconocer algunas enfermedades tales como el cáncer de próstata (Irizarry et al. 2008; Kron et al. 2009), síndrome Xq frágil (SXF) (Iliende et al. 2007), etc., - empleo potencial terapéutico en osteoarthritis. (Ezura et al. 2009), para mencionar algunas.

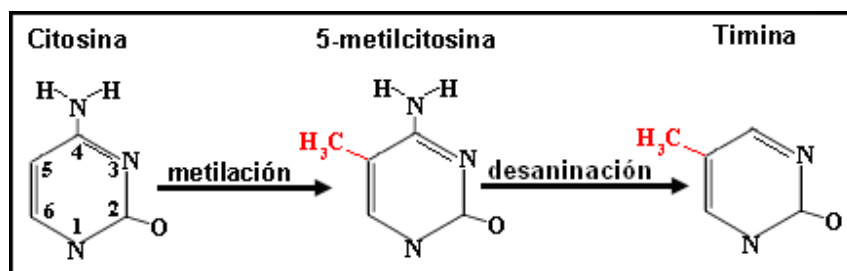


Figura 1.1: Proceso de mutación de la Citosina en Timina

Como se ha visto, el problema de la detección de conglomerados tiene gran importancia y una alta aplicabilidad en dominios bioinformáticos.

1.3 Introducción a la lógica borrosa

Dos de los aspectos que contaminan normalmente la información en cualquier área del saber, son la imprecisión que tiene en su expresión y la incertidumbre que puede provocar la fuente que la proporciona. Ciertas personas tienen suficiente habilidad para tomar decisiones correctas a partir de un conjunto de datos que vienen expresados de forma vaga o imprecisa (borrosos) casi siempre utilizando adjetivos o adverbios como mucho, poco, alto, bajo, normal, muy, entre otros. Tales personas pueden controlar eficientemente un proceso tecnológico (en un central azucarero el tradicional puntista que controla el proceso de cristalización del azúcar), diagnosticar enfermedades o una

enfermedad a partir de síndromes y síntomas (el médico clínico), o tomar una decisión acertada en una determinada empresa e institución. El ser humano se desenvuelve con extraordinaria facilidad a la hora de manejar este tipo de información; sin embargo, cuesta trabajo explicar qué procedimientos sigue para ello (Calviño 2003).

Para hacerle frente a la información imprecisa han surgido diferentes teorías matemáticas: teorías como la de la clásica probabilidad (Feller 1971), como la de la evidencia (Yager 2008), o como la de los Factores de Certeza (Shortliffe y Buchanan 1975). Estas teorías han despertado un creciente interés en la investigación científica. La herramienta por excelencia para modelar fenómenos en los que rige el principio de simultaneidad gradual es la Teoría de los Subconjuntos Borrosos, cuya base son las lógicas multivalentes desarrolladas en las primeras tres décadas del siglo XX (Lukasiewicz 1910). El concepto de conjunto borroso (que caracteriza de manera apropiada la imprecisión en la información) fue introducido en la década del 60 por Lofti A. Zadeh quien se considera el padre de la lógica borrosa (Zadeh 1973; Zadeh 1975).

En términos más rigurosos, la teoría de conjuntos borrosos parte de la teoría clásica de conjuntos, añadiendo una función de pertenencia al conjunto, definida como un número real entre 0 y 1. Así, se introduce el concepto de conjunto o subconjunto borroso asociado a un determinado valor lingüístico, definido por una palabra, adjetivo o etiqueta lingüística A , es decir podemos definir a un subconjunto borroso como $A = \{(x, \mu_A(x)) \mid x \in X\}$ siendo la función de pertenencia:

$$\mu_A : X \rightarrow [0,1]$$

$$x \in X \rightarrow \mu_A(x) \in [0,1]$$

donde 0 indica la no pertenencia al conjunto A y 1 la pertenencia absoluta. Existe una degradación del nivel de pertenencia de forma que si $\mu_A(x) = 0.9$, el nivel de pertenencia del elemento x es muy elevado, y si $\mu_A(x) = 0.1$ el nivel de pertenencia de x es muy bajo. Así, la función de pertenencia puede ser interpretada como el grado en que un elemento particular que se considera, cumple con las especificaciones que definen a los elementos del conjunto en cuestión y no debe interpretarse como la probabilidad de pertenencia. Si la probabilidad de que un elemento x pertenece al conjunto A es de 0.9 y se afirma que x pertenece al conjunto A , tenemos un 90% de probabilidad de acertar, pero el elemento intrínsecamente pertenece o no pertenece al conjunto A . Cuando se dice que la función de pertenencia de x es 0.9 se quiere decir

que cumple en nuestro criterio con el 90% de las características que definen los elementos del conjunto A. En resumen, la probabilidad indica incertidumbre estadística mientras que la función de pertenencia indica vaguedad y subjetividad.

En realidad, esta diferencia entre probabilidad y pertenencia tiene sólo un sentido interpretativo, pero no conceptual desde el punto de vista matemático. La pertenencia es, en última instancia, la “probabilidad, o verosimilitud” de que el objeto se ajuste a la interpretación del conjunto borroso A. Teóricamente puede ser demostrado, que sobre la base de un conjunto simple de axiomas (llamados Axiomas de Cox & Jaynes), que tienen un sentido común, y que en particular se satisfacen racionalmente por las funciones de pertenencia, ellas resultan, salvo una constante multiplicativa, una función de probabilidad.

Concretamente sea $\pi(X|I)$ un número que denota en cualquier sentido la “plausibilidad, creencia o certidumbre” de X, condicionada a la información I, digamos por ejemplo, la plausibilidad de X = “fiebre alta” considerando que I = “temperatura 38° C”.

Los tres axiomas de Cox & Jaynes, establecen, modesta, o mínimamente, que:

1. La función de plausibilidad o certidumbre, debe ser “transitiva”; específicamente, si X es más plausible que Y, e Y es más plausible que Z, *entonces* X debe ser más plausible que Z, o formalmente:

$$\pi(X|I) > \pi(Y|I) \text{ y } \pi(Y|I) > \pi(Z|I) \text{ implica } \pi(X|I) > \pi(Z|I)$$

2. Debe existir una función F que hable de la “no plausibilidad de X”, en términos de la “plausibilidad de X”

$$\pi(\sim X|I) = F(\pi(X|I))$$

3. Debe existir una función G que hable de la plausibilidad de hechos concomitantes, producto de su interacción

$$\pi(X, Y|I) = G(\pi(X|I), \pi(Y|X, I))$$

Con estas condiciones, existe $k > 0$, tal que $P(X|I) = k \pi(X|I)$ está en $[0, 1]$ y P satisface los axiomas de probabilidad (Cox 1946). Aquí resulta $F(x) = 1 - x$, $G(x, y) = xy$. Además, la propiedad de simetría $P(X, Y|I) = P(Y, X|I)$ del axioma 3 conduce al conocido Teorema de Bayes y así, el razonamiento probabilístico bayesiano se convierte en la única forma

consistente de hacer inferencias y deducciones (Baldi y Brunak 2001). No se “trata” entonces de “otra matemática”.

Desde la aparición de la lógica borrosa, son incontables las aplicaciones que se han hecho de ella en el mundo de la investigación en general y en particular en las matemáticas. Estas aplicaciones de forma general tienden a seguir el esquema de la figura 1.2. Algunas de las variables de entradas necesitan suavizarse, tal es el caso de la variable x_1 , mientras que otras no, variable x_2 . Con estos datos se realizan ciertas operaciones, descritas bajo el nombre de “caja negra”, y finalmente se necesita obtener un valor duro por lo que es necesario realizar el proceso inverso a la “Borrosificador”, llamado en la figura 1.2 como “Desborrosificador”, terminología utilizada en (Martín del Brío y Sánchez 2005). No obstante, quizás la principal aplicación actual sean los sistemas de control borroso, que utilizan sus expresiones para formular reglas orientadas al control de sistemas (Brubaker y Cedric 1992). Dichos sistemas de control borroso pueden considerarse una extensión de los sistemas expertos, pero superando los problemas prácticos que éstos presentan en el razonamiento en tiempo real, causados por la explosión exponencial de las necesidades de cálculo requeridas para el análisis lógico completo de las amplias bases de reglas que manejan. Un ejemplo relevante de los sistemas borrosos es el frenado automático de los trenes en el Metro de la ciudad japonesa de Sendai inaugurado el 15 de julio de 1987 (Martín del Brío y Sánchez 2005).

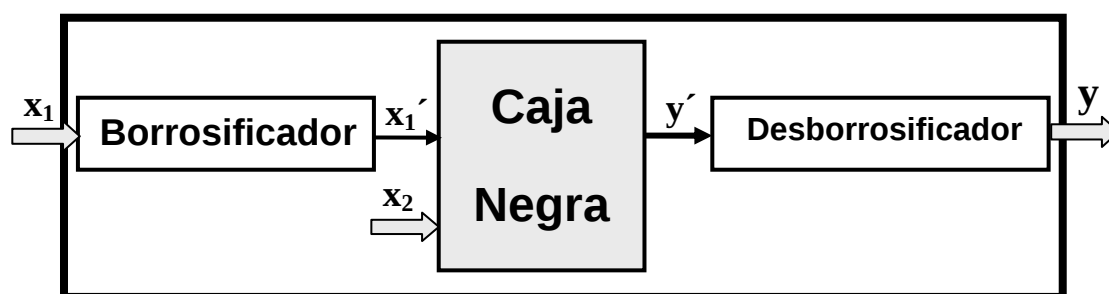


Figura 1.2 Funcionamiento de un sistema de control borroso

1.3.1 Funciones de pertenencia

La función de inclusión o pertenencia (membership function) de un conjunto borroso consiste en un conjunto de pares ordenados $A = \{(u, \mu_A(u)) \mid u \in U\}$ si la variable es

discreta, o una función continua si no lo es. Como ya se ha comentado, el valor de $\mu_A(u)$ indica el grado en que el valor u de la variable U está incluida en el concepto representado por la etiqueta A . Para la definición de estas funciones de pertenencia se utilizan convencionalmente ciertas familias de formas estándar, en las que se encuentran:

Función de pertenencia triangular

Se define por sus límites inferior a y superior b , y el valor modal m , tal que $a < m < b$.

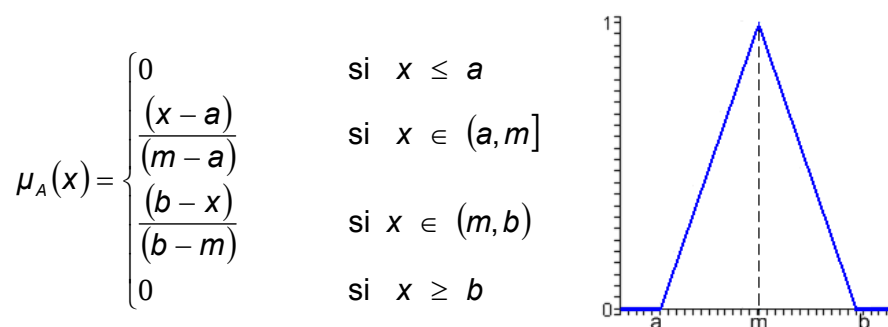


Figura 1.3 Función de pertenencia triangular.

Función de pertenencia trapezoidal

Definida por sus límites inferior a y superior d , y los límites de su soporte, b y c , inferior y superior respectivamente.

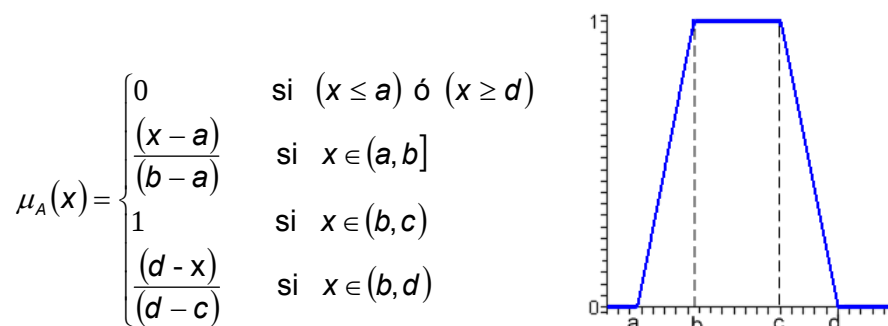


Figura 1.4 Función de pertenencia trapezoidal.

Función de pertenencia Gaussiana

Definida por su valor medio m y el valor $k > 0$. Es la típica campana de Gauss. Cuanto mayor es el valor de k , más estrecha es la campana:

$$\mu_A(x) = e^{-k(x-m)^2}$$

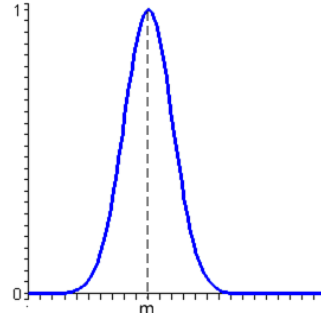


Figura 1.5 Función de pertenencia Gausiana.

Función de pertenencia S

La función S está definida por sus límites inferior a y superior b , y el valor m , o punto de inflexión tal que $a < m < b$. El valor típico es: $m = (a+b) / 2$. El crecimiento es más lento cuanto mayor sea la distancia $a - b$.

$$\mu_A(x) = \begin{cases} 0 & \text{si } x \leq a \\ 2 \left(\frac{(x-a)}{(b-a)} \right)^2 & \text{si } x \in (a, m] \\ 1 - 2 \left(\frac{(x-b)}{(b-a)} \right)^2 & \text{si } x \in (m, b) \\ 1 & \text{si } x \geq b \end{cases}$$

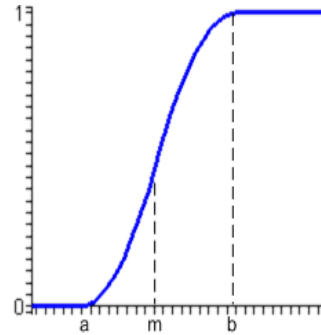


Figura 1.6 Función de pertenencia S.

1.3.2 Borrosificador

Un borrosificador establece una relación entre los puntos, $x = (x_1, x_2, \dots, x_n)$, de entrada no borrosos del sistema, y su correspondiente conjunto borroso A en U (las variables procedentes del exterior serán, en general, valores no borrosos y habrá que “borrosificarlas”¹¹ previamente). Se pueden utilizar diversas estrategias de borrosificación: (Martín del Brío y Sánchez 2005).

¹¹ Borroso, como fuzzy, en inglés es un adjetivo. En la literatura en inglés sobre lógica difusa, lo han convertido en un verbo: “to fuzzy” en el sentido de convertir una variable no borrosa a borrosa. Aquí se hace lo mismo en español cuando se habla de borrosificar.

Borrosificador Singleton

Es el método de borrosificación más utilizado, principalmente en los sistemas de control, y consiste en considerar los propios valores discretos como conjuntos borrosos. De otra forma, para cada valor de entrada x se define un conjunto A' que lo soporta, con función de pertenencia $\mu_A(x')$, de modo que:

$$\mu_A(x') = \begin{cases} 1 & \text{si } x' = x \\ 0 & \text{si } x' \neq x \end{cases} \quad \forall x' \in U$$

Borrosificador no Singleton

Este método utiliza la función exponencial siguiente: $\mu_A(x') = a * \exp[-(x' - x/\sigma)^2]$, función con forma de campana, centrada en el valor x de entrada, de ancho σ y amplitud a .

1.3.3 Desborrosificador

Un desborrosificador es una función que transforma un conjunto borroso en el conjunto V , es decir la salida del dispositivo de inferencia borrosa la convierte en un valor no borroso, $y \in V$ (Martín del Brío y Sánchez 2005). Para esta tarea se utilizan diversos métodos tales como:

Desborrosificador por máximo

Consiste en $y' = \arg.\sup_{y \in V} (\mu_{B'}(x))$ es decir, y' es el punto de V en que $\mu_{B'}(x)$ alcanza su valor máximo, donde $\mu_{B'}(x)$ es el conjunto de los grados de pertenencia de todas las etiquetas analizadas en el problema.

Desborrosificador por medida de centros de salidas:

$$y = \frac{\sum_{l=1}^M y_l \left(\mu_{B'} \left(\begin{pmatrix} -l \\ y \end{pmatrix} \right) \right)}{\sum_{l=1}^M \left(\mu_{B'} \left(\begin{pmatrix} -l \\ y \end{pmatrix} \right) \right)}$$

y^l representa el centro del conjunto borroso G^l , es decir, el punto en V donde $\mu_{G^l}(y)$ alcanza su valor máximo y $\mu_{B^l}(y) = \sup_{x \in U} [\mu_{F_1^1 x \dots x F_n^1 \rightarrow G^1}(x, y) * \mu_{A^l}(x)]$.

Desborrosificador por centro de área

$$y = \frac{\sum_{l=1}^M M^l \left(\mu_{B^l} \left(\begin{pmatrix} -l \\ y \end{pmatrix} \right) \right)}{\sum_{l=1}^M A^l \left(\mu_{B^l} \left(\begin{pmatrix} -l \\ y \end{pmatrix} \right) \right)}$$

M^l es el momento (entorno al eje y del universo de discurso de la salida V) de la función de inclusión del conjunto borroso G^l , A^l es el área, y $\mu_{B^l}(y)$ se define como: $\mu_{B^l}(y) = \sup_{x \in U} [\mu_{F_1^1 x \dots x F_n^1 \rightarrow G^1}(x, y) * \mu_{A^l}(x)]$.

Hasta aquí se han presentado las funciones de pertenencia y los métodos de desborrosificación de uso más frecuente reportados en la literatura. Algunas de ellas se utilizarán con posterioridad formando parte de la contribución propuesta.

1.4 Diseño de experimentos bifactorial no paramétrico

Los métodos no paramétricos constituyen una rama de la estadística que estudia los datos cuya distribución no se ajusta a los llamados criterios paramétricos. La utilización de estas técnicas se hace recomendable cuando no se puede asumir que los datos se ajusten a una distribución conocida, cuando el nivel de medida empleado no sea, como mínimo, de intervalo. Tal es el caso de las investigaciones en el campo de la Bioinformática.

Un experimento factorial es aquel en el que se estudian simultáneamente varios factores de modo que los tratamientos se forman por todas las posibles combinaciones de los niveles de los factores. Un experimento factorial no constituye un nuevo diseño experimental, sino un diseño para la formación de los tratamientos. Los experimentos factoriales pueden ser conducidos bajo los lineamientos de cualquier diseño experimental tal como el diseño complementario al azar (DCA), diseño de bloques al azar (DBCA) o diseño cuadrado latino (DCL) (Montgomery 2008).

Los experimentos factoriales son ampliamente utilizados y son de gran valor en el trabajo exploratorio cuando se sabe poco sobre los niveles óptimos de los factores o ni siquiera qué factores son importantes.

Existen paquetes estadísticos para realizar los diseños de experimentos clásicos: DCA, DBCA, DCL, diseños factoriales y muchos otros (Hinkelmann y Kempthorne 2005; Hinkelmann y Kempthorne 2008).

Análisis bifactorial no paramétrico

Existe un fundamento teórico de cómo realizar un análisis en el caso de diseños equilibrado. La idea esencial fundamentada por R.R. Sokal and F. J. Rohlf, (1995) fue elaborar un análisis de varianza bifactorial no paramétrico, ranqueando la variable dependiente, como lo hace el test de Kruskal-Wallis. Se utilizan las sumas de cuadrados de la variable dependiente ranqueada y se recalculan grados de libertad de cada factor y su interacción para ofrecer finalmente una significación de cada efecto (Shad y Madden 2004). Si algún factor tiene más de dos niveles, se pueden utilizar test de rangos a posteriori clásicos, que se basa fundamentalmente en rangos para obtener subconjuntos homogéneos, por ejemplo, el test de Dunnett C, válido incluso ante falta de homogeneidad de varianzas.

Algoritmo para un análisis bifactorial no paramétrico:

1. Ranquear la variable dependiente.
2. Aplicar el análisis de varianza sobre la variable dependiente ranqueada, para obtener la suma de cuadrados (SC) por cada factor y su interacción, así como sus grados de libertad.
3. Calcular el CMT (Cuadrado Medio Total)

$$CMT = \frac{abr(abr + 1)}{Total}$$

Donde,

a: es el número de niveles del primer factor.

b: es el número de niveles del segundo factor.

r: es el número de réplicas de cada combinación.

Total: Total de casos analizados.

4. Calcular el estadígrafo H para cada factor y la interacción

$$H = \frac{SC(\text{correspondiente})}{CMT}$$

5. Calcular la significación de cada H utilizando la distribución de chi-cuadrado, teniendo presente los grados de libertad del factor o de la interacción analizada. (La variable H tiene distribución chi-cuadrado).

Dicha fundamentación teórica, implica desde el punto de vista práctico, que:

1. Se puede utilizar el paquete estadístico SPSS¹² o cualquier otro para:
 - ✓ Hacer el análisis descriptivo de datos, por ejemplo, a través de cubos OLAP¹³, que indiquen los posibles resultados a obtener, y que finalmente permitirán interpretar los resultados obtenidos.
 - ✓ Ranquear la variable dependiente, al estilo de como lo haría el test de Kruskal-Wallis.
 - ✓ Aplicar el análisis de varianza sobre la variable dependiente ya ranqueada, para obtener la suma de cuadrados, y de paso obtener los test de rangos sobre cualquier factor con más de dos niveles.
2. Utilizar después el paquete *Mathematica*¹⁴ (podría ser incluso el Excel) para implementar como tal, el test bifactorial no paramétrico, y en particular:
 - ✓ Usar las sumas de cuadrados de los rangos y sus grados de libertad obtenidas en el paso anterior.
 - ✓ Recalcular con el *Mathematica*, el valor de CMT , los valores de H y las diferencias honestamente significativas, desde el punto de vista no

¹² Statistical Package for the Social Sciences (SPSS) paquete de programas estadístico muy usado en las ciencias sociales y las empresas de investigación de mercado.

¹³ OnLine Analytical Processing (OLAP), realiza una disposición de los datos en vectores para permitir un análisis rápido de los mismos.

¹⁴ Programa de propósito general utilizado en áreas científicas, de ingeniería, matemáticas y áreas computacionales, también puede ser utilizado como un sistema de álgebra computacional.

paramétrico, debidas a efectos principales y/o su interacción, pero acorde a la nueva teoría.

3. Con los resultados del *Mathematica* poder regresar a las salidas del SPSS para:
 - ✓ Interpretar los resultados generales. Ello se logra con las estadísticas descriptivas proporcionadas por el OLAP del SPSS en la primera parte. Un cubo OLAP debe ser re-visualizado de manera que evidencie la influencia de uno u otro factor y su posible interacción.
 - ✓ Interpretar los resultados de los tests de comparaciones múltiples del SPSS sobre factores con más de dos niveles. Ello se puede visualizar a través de las estadísticas descriptivas de conformación de grupos homogéneos que proporciona el propio ANOVA.

Para no ser tan engorroso el trabajo práctico con el procedimiento anterior se ha programado completamente el mismo en el paquete *Mathematica* con tres funciones simples utilizando el contexto de ANOVA dentro del paquete de *Mathematica* que permite realizar el análisis de varianza a la variable ranqueada (Anexo 1).

1.5 Algoritmos bioinspirados

En la actualidad los modelos bioinspirados se muestran eficientes en la solución de problemas de optimización prácticos de diversas áreas. Dentro de los algoritmos bioinspirados usados para la selección de rasgos, la inteligencia de enjambres (Swarm Intelligence, SI) ha sido objeto de estudio, investigación y de mucha aplicación por su simplicidad y robustez. En particular se puede mencionar el uso de estas técnicas en la búsqueda de la estructura de una red bayesiana (Chávez et al. 2007a; Chávez et al. 2008a).

La metaheurística PSO, (Particle Swarm Optimization), fue desarrollada por Kennedy y Eberhart (Kennedy 1997; Kennedy y Eberhart 1995a; Kennedy et al. 1998) y está inspirada en el comportamiento social observado en grupos de individuos tales como bandadas de pájaros, enjambres de insectos o bancos de peces. Un enjambre se define como una colección estructurada de organismos (agentes) que interactúan. La inteligencia no está en los individuos sino en la acción de todo el colectivo. Tal comportamiento social se basa en la transmisión del éxito de cada individuo a los

demás del grupo, lo cual resulta un proceso “sinérgico” que permite a los individuos satisfacer de la mejor manera posible sus necesidades más inmediatas, tales como la localización de alimentos o de un lugar de cobijo. Cada organismo (partícula) se trata como un punto en un espacio N dimensional el cual ajusta su propio “vuelo” de acuerdo a su propia experiencia y la experiencia del resto de la banda. La banda (swarm) “vuela” por el espacio de búsqueda localizando regiones o partículas prometedoras (Kennedy y Eberhart 1995b; Kennedy et al. 1998).

En general el PSO se puede emplear en la solución de problemas complejos de optimización global y presentan características muy interesantes tales como:

- Tiene potente capacidad de exploración.
- Su proceso de búsqueda gradual aproxima las soluciones óptimas.
- Sencillo de entender e implementar.
- Bajo costo computacional en términos de memoria y tiempo.

Fundamentos generales del Algoritmo

Sean:

R^N – R espacio de búsqueda designado, N : cantidad de dimensiones que cuenta dicho espacio.

$\mathbf{x}_i^k = (x_{i1}^k, x_{i2}^k, \dots, x_{iN}^k)$ – Posición de la i -ésima partícula en R^N de la iteración k .

$\mathbf{v}_i^k = (v_{i1}^k, v_{i2}^k, \dots, v_{iN}^k)$ – Velocidad de la i -ésima partícula en R^N de la iteración k .

$\mathbf{p}_i = (p_{i1}, p_{i2}, \dots, p_{iN})$ – Mejor posición de la i -ésima partícula en R^N de las k iteraciones.

$\mathbf{p}_g = (p'_1, p'_2, \dots, p'_N)$ – Mejor posición del grupo (Mejor partícula entre las k iteraciones).

f_i^k – Valor de la función objetivo evaluada en $\mathbf{x}_i^k$.

f_i^{best} – Mejor valor de la función objetivo evaluada en la i -ésima partícula de las k iteraciones.

f_g^{best} – Mejor valor de la función objetivo evaluada en el grupo.

V_{max} – Velocidad máxima que puede alcanzar una partícula, entonces $V_{min} = -V_{max}$ es la velocidad mínima que puede tener una partícula.

ω – Coeficiente de inercia: valor aleatorio en el rango $[0.5, 1]$.

c_1, c_2 – Parámetros sociales y cognoscitivos.

r_1, r_2 – Números aleatorios entre $[0, 1]$.

A continuación se describen los pasos del algoritmo:

Paso 1: Inicializar una población de p partículas.

- a. Darle valores a las variables v_{max}, c_1, c_2 .
- b. Inicializar la población de las partículas $\mathbf{x}_i^0 \in \mathbf{D}$ en $\mathbf{R}^n$ para $i = 1, \dots, p$.
- c. Inicializar la velocidad de las partículas $-v_{max} \leq \mathbf{v}_i^0 \leq v_{max}$ para $i = 1, \dots, p$.
- d. $k = 1$

Paso 2: Optimizar.

- e. Calcular los valores de f_i^k
- f. Si $f_i^k \leq f_i^{best}$ entonces $f_i^{best} = f_i^k, \quad \mathbf{p}_i = \mathbf{x}_i^k$
 Si $f_i^k \leq f_g^{best}$ entonces $f_g^{best} = f_i^k, \quad \mathbf{p}_g = \mathbf{x}_i^k$
- g. Si se cumple la condición de parada entonces ir a 3.
- h. Actualizar la velocidad de las partículas como sigue

$$\mathbf{v}_{id}^{k+1} = \omega * \mathbf{v}_{id}^k + c_1 r_1 (\mathbf{p}_{id} - \mathbf{x}_{id}^k) + c_2 r_2 (\mathbf{p}_{gd} - \mathbf{x}_{id}^k) \quad \text{para } d = 1, \dots, N$$
- i. Actualizar la posición de las partículas como sigue

$$\mathbf{x}_{id}^{k+1} = \mathbf{x}_{id}^k + \mathbf{v}_{id}^k \quad \text{para } d = 1, \dots, N$$
- j. Incrementar k .
- k. Ir a 2(a).

Paso 3: Terminar.

La velocidad es una función que está compuesta por tres sumandos. El primero es la velocidad anterior de la partícula, conociéndose a esta parte como inercia. El segundo sumando es la diferencia entre la mejor posición encontrada por la partícula con la actual posición, esta es la parte cognitiva que representa el aprendizaje de su propia experiencia. El último sumando es la diferencia entre la mejor posición alcanzada por un vecino, con la posición actual de la partícula y es la parte social, que representa el aprendizaje del grupo (Kennedy et al. 2001; Wang et al. 2007). El coeficiente de inercia

ω regula el impacto de la velocidad para valores grandes, significa que las partículas deben cambiar su velocidad instantáneamente y moverse lejos de su posición según su conocimiento, o sea se favorece la exploración global (global search), mientras que para valores pequeños la partícula no hará cambios bruscos, es decir la inercia sugiere continuar el camino original, aún cuando se conozca el mejor estado (fitness), favoreciendo la exploración local (local search).

La selección de los parámetros ω , c_1 y c_2 tienen impacto en la velocidad de convergencia y la velocidad del algoritmo para encontrar el óptimo. Se recomienda que c_1 y c_2 no tomen necesariamente el mismo valor sino, que se generen aleatoriamente con distribución uniforme en el intervalo $[0, 2]$. En (Beielstein et al. 2002) se recomienda que la suma de estos valores sea menor o igual a 4. El trabajo de Beielstein et al. resulta interesante pues hace un análisis de los parámetros del algoritmo PSO mediante técnicas de diseños experimentales (Mahamed et al. 2005). Para obtener una mayor información acerca de la influencia de estos parámetros en la efectividad del algoritmo PSO ver (Beielstein et al. 2002; Kennedy et al. 2001; Shi y Eberhart 1998).

1.6 Métodos de Monte Carlo

Los métodos de Monte Carlo son un conjunto de algoritmos computacionales que basan sus resultados en el uso de un muestreo aleatorio con reposición (Buckley y Jowers 2007). Se utiliza con frecuencia para simular el comportamiento de sistemas físicos o matemáticos complejos. Debido a su uso intensivo, a partir de la generación de números aleatorios (o pseudo-aleatorios), los métodos de Monte Carlo se utilizan para realizar sus cálculos con ayuda de microcomputadoras. Algunas de sus aplicaciones son las siguientes:

- Criptografía
- Densidad y flujo de tráfico
- Diseño de reactores nucleares
- Ecología
- Econometría
- Física de materiales

- Sistemas de colas

La invención del método de Monte Carlo se asigna a Stan Ulam y a John Von Neumann. En 1946, Ulam explicó cómo se le ocurrió la idea mientras jugaba un solitario durante una enfermedad en 1946. A principios de 1947 Von Neumann envió una carta a Los Álamos en la que expuso de modo influyente tal vez el primer informe por escrito del método de Monte Carlo.

El método fue llamado así por ser el principado de Mónaco, “la capital del juego de azar”, al tomar una ruleta como un generador simple de números aleatorios. El uso real de los métodos de Monte Carlo como una herramienta de investigación, viene a la luz con el diseño de la bomba atómica durante la Segunda Guerra Mundial.

De manera general, el método de Monte Carlo, también conocido como “Simulación de Monte Carlo” da solución a una gran variedad de problemas matemáticos haciendo experimentos con muestreos estadísticos en una computadora. Es aplicable no sólo a problemas estocásticos, sino también determinísticos.

Generalmente en estadística los modelos aleatorios se usan para simular fenómenos que poseen algún componente aleatorio y por ello el método de Monte Carlo aparece frecuentemente. Ejemplos típicos son la mejor aproximación de la significación de los test no paramétricos, generando aleatoriamente muchas tablas aleatorias con distribución similar a los de una muestra real y repitiendo el test para todas las muestras, proponiendo como significación la media de las obtenidas, añadiendo un intervalo de confianza para ella. Pero como se ha dicho, el método puede utilizarse en problemas que no tienen un componente aleatorio explícito en estos casos un parámetro determinista del problema se expresa como una distribución aleatoria y se simula dicha distribución. Un ejemplo clásico es su uso para el cálculo eficiente de integrales impropias o múltiples con altas dimensiones. Otro ejemplo interesante es el famoso problema de las Agujas de Buffon¹⁵ (Pertusa 2003).

Así, las técnicas de Monte Carlo tienen el objetivo de generar un suceso aleatorio o pseudo-aleatorio para estudiar el comportamiento del modelo o problema tratado. Se

¹⁵ Naturalista y matemático del siglo XVIII Georges-Louis Leclerc, Conde de Buffon, descubrió un ingenioso método para la estimación de π basado en el lanzamiento al azar de agujas sobre un tablero, esto permite calcular la longitud de un objeto.

considera que el método de Monte Carlo es una herramienta de investigación basada fundamentalmente en la técnica de muestreo artificial, empleada para operar numéricamente sistemas complejos que tengan componentes aleatorios o determinísticos, manteniendo tanto la entrada como la salida un cierto grado de incertidumbre. Cuando la generación de números aleatorios es relativamente reducida, los resultados obtenidos en la simulación pueden ser muy sensibles a las condiciones iniciales. Se usa con frecuencia los métodos Quasi-Monte Carlo, los cuales consisten en acotar la generación de los números aleatorios.

Se insiste en que generalmente, en estadística los modelos aleatorios se usan para simular fenómenos que poseen algún componente aleatorio. Pero en el método de Monte Carlo, por otro lado, el objeto de la investigación es el modelo en sí mismo, y se usa un suceso aleatorio o pseudo-aleatorio para estudiar el modelo.

Principios básicos del método de Monte Carlo

El fundamento del método hay que buscarlo en el teorema del Límite Central de la teoría de probabilidades, donde el valor medio de una variable aleatoria ξ , puede estimarse por el valor medio de N valores resultantes del sorteo de la variable, el cual se distribuye aproximadamente normal, cuya varianza es $\frac{\sigma(\xi)}{\sqrt{N}}$.

En general, los valores de la variable ξ se obtienen partiendo de un sorteo de la variable aleatoria γ equiprobable en el intervalo $(0, 1)$, es decir generando números aleatorios en dicho intervalo mediante las diversas técnicas existentes al respecto, la relación entre los valores de γ y de ξ , viene dado por:

$$\gamma = \int_a^{\xi} p(x) dx \text{ siendo } p(x) \text{ la densidad de probabilidad correspondiente a la variable}$$

aleatoria ξ , definida en el intervalo (a,b) . Resulta difícil expresar analíticamente la función $\xi = f^{-1}(\gamma)$, a partir de la ecuación anterior, por lo que se recurre a procedimientos numéricos.

1.7 Evaluación de los conglomerados como clasificadores

Cuando los conglomerados se utilizan como un modelo clasificador, como se hará en el presente trabajo, se hace necesario evaluar su desempeño, al igual que se realiza la evaluación en cualquier problema de clasificación supervisada. Para ello se utilizan criterios¹⁶ tales como: por ciento de clasificaciones correctas, diferentes medidas del error, el índice de Kappa (Brender et al. 1994), medida F (Van-Rijsbergen 1979), y funcionales de calidad y error (Donald et al. 1994; Ruiz-Shulcloper y Abidi 2002). La capacidad del modelo para representar confiablemente el sistema real, se relaciona esencialmente con su exactitud (accuracy) (Daalen 1992) (Daalen 1992). No existe un modelo clasificador mejor que otro de manera general; para cada problema nuevo es necesario determinar con cuál se pueden obtener mejores resultados, y es por esto que han surgido varias medidas como las mencionadas anteriormente, para evaluar la clasificación y comparar los modelos empleados para un problema determinado. Las medidas más conocidas para evaluar la clasificación están basadas en la matriz de confusión que se obtiene cuando se prueba el clasificador en el conjunto de datos del entrenamiento.

Matriz de Confusión		Clase verdadera		Total fila
		Pos	Neg	
Clase Predicha	pos	<i>VP</i>	<i>FP</i>	<i>P*</i>
	neg	<i>FN</i>	<i>VN</i>	<i>N*</i>
Total columna		<i>P</i>	<i>N</i>	<i>Total</i>

Figura 1.7. Matriz de confusión de un problema de dos clases.

En la Figura 1.7 se muestra la matriz de confusión de un problema de dos clases, donde Pos/pos es la clase positiva y Neg/neg la clase negativa. Las siglas *VP* y *VN* representan los elementos bien clasificados de la clase positiva y negativa respectivamente y *FP* y *FN* identifican los elementos negativos y positivos mal clasificados respectivamente. Basados en estas medidas, se calcula el error, la exactitud, la razón de *VP* (*rVP*) o sensibilidad, la razón de *FP* (*rFP*), la precisión y la especificidad, que se muestran en las expresiones de la Tabla 1.1.

¹⁶ Indistintamente se utilizan los términos criterio o medida para hacer referencia a los aspectos cuantitativos o cualitativos a considerar en la evaluación.

Tabla 1.1. Métricas de evaluación estándar.

Nombre	Medida
Exactitud	$\frac{VP + VN}{P + N}$
rVP o sensibilidad	$\frac{VP}{P}$
rVN o especificidad	$\frac{VN}{N}$
rFP	$\frac{FP}{N}$
rFN	$\frac{FN}{P}$
Precisión	$\frac{VP}{VP + FP}$
Medida F	$\frac{2}{\frac{1}{precision} + \frac{1}{sensibilidad}}$
Correlación de Matthews	$mcc = \frac{VP * VN - FP * FN}{\sqrt{(VP + FN)(VN + FP)(VP + FP)(VN + FN)}}$

Otra forma de evaluar el rendimiento de un clasificador es mediante las curvas ROC (Receiver Operating Characteristics graphs, curvas de características de operación del receptor) (Fawcett 2004). En esta curva se representa el valor de razón de VP contra la razón de FP , mediante la variación del umbral de decisión. Se denomina umbral de decisión a aquel que decide si una instancia x , a partir del vector de salida del clasificador, pertenece o no a cada una de las clases. Usualmente, en el caso de dos clases se toma como umbral por defecto 0.5 ; pero esto no es siempre lo más conveniente. Se usa el área bajo esta curva, denominada AUC (Area Under the Curve, área bajo la curva ROC) como un indicador de la calidad del clasificador. En tanto dicha área esté más cercana a la unidad, el comportamiento del clasificador está más cercano al clasificador perfecto (aquel que lograría 100% de VP con un 0% de FP).

En (Larrañaga et al. 2005) se hace una comparación de diferentes paradigmas de

clasificación supervisada en Bioinformática: bayesianos, estadísticos, inductivos y de IA. Resulta interesante el uso de las curvas ROC para la comparación, así como análisis de la razón de error basado en la matriz de confusión (Fawcett 2004).

Existen otros tipos de gráficos que permiten comparar clasificadores, por ejemplo las curvas precision-recall pueden ser particularmente útiles cuando las clases son desbalanceadas porque a diferencias de las curvas ROC ellas si son sensibles a la distribución de las clases. En el artículo fundamental de Fawcett se comenta brevemente este tema pero además hay otros artículos en que profundiza en la relación entre las curvas ROC y precision-recall, por ejemplo (Davis y Goadrich 2006)¹⁷. Por la experiencia anterior de uso en Bioinformática, comentadas en el párrafo precedente, se decidió trabajar entonces con las curvas ROC.

1.8 Consideraciones finales del capítulo

En el presente capítulo se enuncia la definición de las técnicas de detección de conglomerados. Se presentan los fundamentos matemáticos del método Scan en sus dos variantes: lineal y circular y se discuten algunas consideraciones relacionadas con la influencia de los valores de los parámetros en la capacidad de detección de conglomerados, así como su aplicación al análisis de secuencias genómicas y de otros problemas bioinformáticos.

Para la elaboración del marco teórico se tuvieron en cuenta además, numerosos aspectos de estadística y de inteligencia artificial que se utilizan más adelante para fundamentar la propuesta de la contribución. Entre ellos pueden mencionarse elementos de la lógica borrosa, la teoría de los diseños de experimentos no paramétricos, los algoritmos bioinspirados, en particular el PSO y también las técnicas de simulación de Monte Carlo.

Los problemas de detección de conglomerados de algún patrón de secuencias en bioinformática son generalmente problemas complejos que requieren de un largo proceso de análisis y procesamiento. Por ello se hace necesario generalizar los

¹⁷ ACM International Conference Proceeding

algoritmos utilizados en epidemiología e implementarlos en plataformas de software libre para que puedan ser usados por la comunidad científica.

En el próximo capítulo se proponen la generalización de las dos variantes del método Scan, la utilización de diferentes técnicas en la detección de los parámetros del método deben favorecer la obtención de mejores niveles de exactitud y precisión del mismo.

CAPÍTULO II. NUEVOS MÉTODOS DE DETECCIÓN DE CONGLOMERADOS. AJUSTE DE SUS PARÁMETROS

Entre los métodos de detección de conglomerados más populares en Higiene y Epidemiología están el de Grimson y el Scan en sus dos variantes: lineal y circular. Ellos se caracterizan porque tienen como datos de entrada una variable relacionada con las fechas del suceso que se analiza, que se ordena cronológicamente y se realiza el análisis correspondiente para determinar la existencia de conglomerados en el tiempo.

Existen muchas ramas de la ciencia donde los datos analizados no están relacionados con fechas, pero que los mismos tienen un orden que debe respetarse y resulta importante conocer si existen conglomerados de algunas de sus categorías respetando el orden establecido. Se hace necesario entonces modificar los métodos anteriores para ampliar su rango de aplicación. Por ejemplo, en el campo de la Bioinformática se estudian conglomerados de ciertas subcadenas de nucleótidos en el ADN de ciertas especies. La localización de tales conglomerados es de interés porque puede brindar información genética. Algunas veces, la existencia como tal de esos conglomerados pueden informar sobre diferentes alteraciones biológicas importantes, orígenes de replicación, enfermedades, entre otros.

2.1 Generalización de los métodos de detección de conglomerados

Como se ha mencionado hay varias razones que han propiciado la idea de estudio de conglomerados de una categoría de interés, no relacionada con el tiempo; pero en estos casos es necesario (o al menos es suficiente para lograr la generalización más inmediata) que los nuevos datos estén ordenados por algún criterio. Por ejemplo si se trabaja con secuencias de bases que representan algún gen completo, o una porción de éste, sería correcto asumir que tal juego de datos ya está ordenado en el orden que aparecen los nucleótidos en la estructura primaria.

Definición 1:

Un conglomerado o cluster de la categoría de interés es un exceso de dicha categoría, respecto a su valor esperado.

Por tanto se transforma dicha secuencia en una secuencia 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 u otro evento que se considere. El valor cero se asociará a todas las demás categorías, (Langrand 2005). Los datos transformados se representan en una línea, donde los valores son equidistantes. 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 porción de la secuencia del gen **Ataxin 2** y que dentro de ella resulta de interés determinar si existen conglomerados de la subsecuencia **cag** y de esta forma inferir una Ataxia Espino-cerebelar. La transformación de la secuencia original en una dicotómica se realiza como se muestra en la Figura 2.1:

Secuencia:	... tcgctgaagccc cag cag cag cag cag cag cag cag cag cag ...
Transformación:	... 000000000000 1 1 1 1 1 1 1 1 1 1 ...

Figura 2.1. Ejemplo de la conversión de una porción de la secuencia de un gen

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

De manera general pueden definirse 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 la secuencia de la categoría de interés.

El método generalizado define un intervalo o ventana de tamaño fijo que se mueve, con un determinado paso, por el eje de longitud, es decir 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.

Para la formulación matemática es necesario definir los siguientes aspectos de manera análoga a como se hace en epidemiología, que se describen en el epígrafe 1.1.1.

Definición 2: Sean:

t : amplitud de la ventana móvil.

T : longitud de la secuencia analizada.

$L = \frac{T}{t}$: fracción que representa la longitud total que se analiza con relación al ancho de la ventana.

n : cantidad de la categoría de interés (unos) presentes en la secuencia.

λ : número esperado de la categoría de interés por unidad de espacio en un proceso de Poisson.

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

$\eta = \max_{0 \leq y \leq T-t} \{w_{y, y+t}\} \in \mathbb{Z}^+$: estadígrafo del test.

2.1.1 Algoritmo del método Scan Generalizado sobre una línea

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

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

Paso 2: Definir una ventana móvil de longitud fija y un paso (cantidad de elementos). Calcular *cantidad* de unos en la ventana, inicializar *máximo* y *acumular* la suma.

Paso 3: Utilizando el paso desplazar la ventana a lo largo de la línea de longitud y calcular en cada caso: *cantidad* de unos asociados, guardar el *máximo*, *acumular* la suma.

Paso 4: Calcular promedio y fracción mínima de ventanas a formar.

Paso 5: Calcular la probabilidad del test utilizando la fórmula propuesta en Naus (1982).

En el Anexo 2 se muestran la programación de funciones más importantes sobre el paquete *Mathematica*. La función **ScanValidation** determina los parámetros necesarios que necesita las demás funciones para el cálculo de las fórmulas propuestas en el Paso 5.

2.1.2 Algoritmo del método Scan Generalizado sobre un círculo

Este método constituye una variación del anterior. 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 en esencia es el mismo que el 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 análogos del último intervalo considerado con los del principio, como si los mismos están relacionados por algún motivo, por ejemplo en el caso de bioinformática ocurre con los genomas mitocondriales que son circulares.

Dado los detalles anteriores, el algoritmo varía en el paso cinco en el cálculo de la probabilidad de observar η o más casos en un intervalo o ventana de tamaño fijo ya que Q no se estima de igual forma, en general se utiliza las formulas definidas en (1.9) y (1.10) (Naus 1982).

En el Anexo 3 aparecen las mismas funciones que se programaron en el Anexo 2, resaltando las diferentes instrucciones necesarias para el análisis circular.

2.2 Estudio con datos simulados

Para probar prácticamente la generalización de los métodos Scan se realiza un experimento bastante amplio con datos simulados: se generaron verdaderos y falsos conglomerados utilizando secuencias de ceros y unos, generados aleatoriamente con la distribución de Bernoulli. Se define así juegos de datos de verdaderos y falsos conglomerados de 1000 secuencias de igual tamaño. Las diferencias fundamentales entre estos además de ser de verdaderos y falsos conglomerados es que cada uno de ellos se caracteriza por ser de conjuntos de secuencias de diferentes longitudes, pues es conocido que los métodos de detección de conglomerados no responden de la misma forma ante poblaciones de diferentes tamaños (Casas et al. 2004).

2.2.1 Bases de la simulación realizada

Para generar verdaderos conglomerados se siguen los siguientes pasos:

- Un quinto del tamaño de la secuencia (20%) es generada con una probabilidad

grande de presencia de unos. La probabilidad con que se genera el conglomerado y su tamaño pueden modificarse por el investigador.

- El resto de la población es generada con una probabilidad pequeña de presencia de unos.
- En el conjunto de menor probabilidad de unos se determina un número aleatorio entre uno y la longitud de esta subsecuencia más uno, insertándose en esta posición el conjunto de mayor cantidad de unos. De esta forma se obtiene una secuencia de ceros y unos que tiene al menos un conglomerado.

Ejemplo 1: tamaño de la secuencia igual a 40

1^{ero}. Con una probabilidad 0.95 de presencia de unos se generan 8 valores:

1 1 1 1 1 1 1 1

2^{do}. Se genera el resto de la población con probabilidad 0.09 de presencia de unos (32 valores):

0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0

3^{ero}. Se genera aleatoriamente un valor entre 1 y 33: **17**

4^{to}. El conjunto con verdaderos conglomerados se inserta en la posición 17:

0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0

Como puede apreciarse a simple vista en la secuencia generada existe un conglomerado de unos.

Para generar secuencia con falsos conglomerados se utiliza la distribución de Bernoulli con una probabilidad de 0.5 (mayor entropía) de presencia de unos.

Ejemplo 2: tamaño de la secuencia igual a 20

0 0 1 0 1 1 0 1 0 0 1 0 1 0 1 1 0 0 1 0 1 0 0 1 0 1 0 0 1 1 0 0 0 1 0 0 1 0 0 1

Se simularon juegos de datos con tamaños de secuencias iguales a 100, 300 y 500 elementos. Los juegos de datos con verdaderos conglomerados y falsos conglomerados se generaron de la forma explicada, con 1000 secuencias cada uno. Para determinar si existe o no un conglomerado de unos se analiza el nivel de significación de los métodos de la siguiente forma:

- ✓ $p \leq 0.05$ se detecta conglomerado, (significación).

- ✓ $0.05 < p$ no detecta conglomerado, (no significativa).

El método Scan Generalizado se aplica a cada juego de datos generado. Los parámetros se varían de la siguiente forma:

- Paso: 1%, 15% y 25%.
- Ventana móvil: varía desde el valor más pequeño posible: (paso) hasta el valor mayor posible: 100%.

Debe aclararse que el porcentaje está calculado en base al tamaño de la secuencia que se procesa.

Los resultados de la aplicación del método se pueden graficar en dos dimensiones. El eje de las abscisas representa el valor de la ventana móvil en por ciento, mientras que el eje de las ordenadas contabiliza la frecuencia absoluta de la existencia o no de conglomerados según el caso, ver Figuras 2.2 y 2.3. Las dos curvas que se muestran tienen la siguiente interpretación:

1. Curva significativa (línea continua de color negro), representa la frecuencia absoluta de la detección de conglomerados de cada una de las ventanas.
2. Curva no significativa (línea de color menos intenso), representa la no detección de conglomerados, es decir para cada ventana es 1000 menos la frecuencia absoluta de la ventana móvil analizada.

2.2.2 Resultados y discusión

Los resultados son analizados en las secuencias con verdaderos y falsos conglomerados de secuencias de tamaño 100, 300 y 500 elementos, con ambas variantes del Método Scan.

2.2.2.1 Secuencias con verdaderos conglomerados

En ambas variantes del método Scan Generalizado no se detectan conglomerados para valores pequeños y grandes de las ventanas móviles, esto implica que la curva significativa (línea continua de color negro) tenga valores nulos al inicio y final de las mismas, por esta misma razón la curva no significativa (línea de color menos intenso) comienza y termina con valores máximos. Debe señalarse que las curvas significativas

y no significativas tienen comportamientos opuestos, relacionados con su monotonía. (Figura 2.2 y 2.3). Observe además que la primera curva tiene un máximo y la segunda tiene un mínimo, que son de tipo meseta en dependencia de la evidencia del conglomerado en cada juego de datos; en los Anexos 4 y 5 se muestran los gráficos donde los conglomerados fueron creados con diferentes longitudes en correspondencia al tamaño total de la secuencia. Lo explicado anteriormente es válido para cualquier paso analizado, con la excepción de que para pasos grandes no hay ventanas pequeñas y por tanto comienza detectando conglomerados.

A continuación se muestra los gráficos correspondientes al aplicar cada juego de datos con los métodos Scan Generalizado.

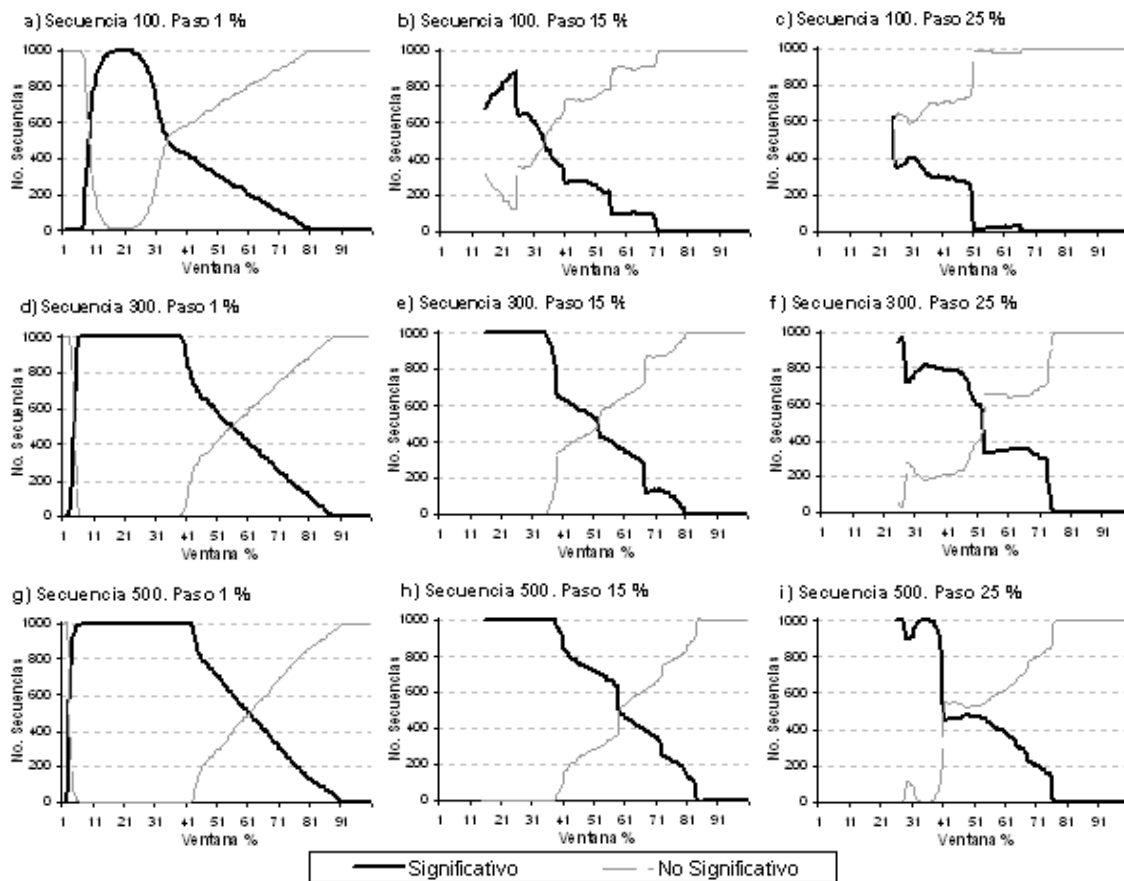


Figura 2.2 Scan Generalizado sobre una línea en población de secuencias de tamaño 100, 300 y 500 elementos con verdaderos conglomerados creados con el 20% del tamaño total de la secuencia.

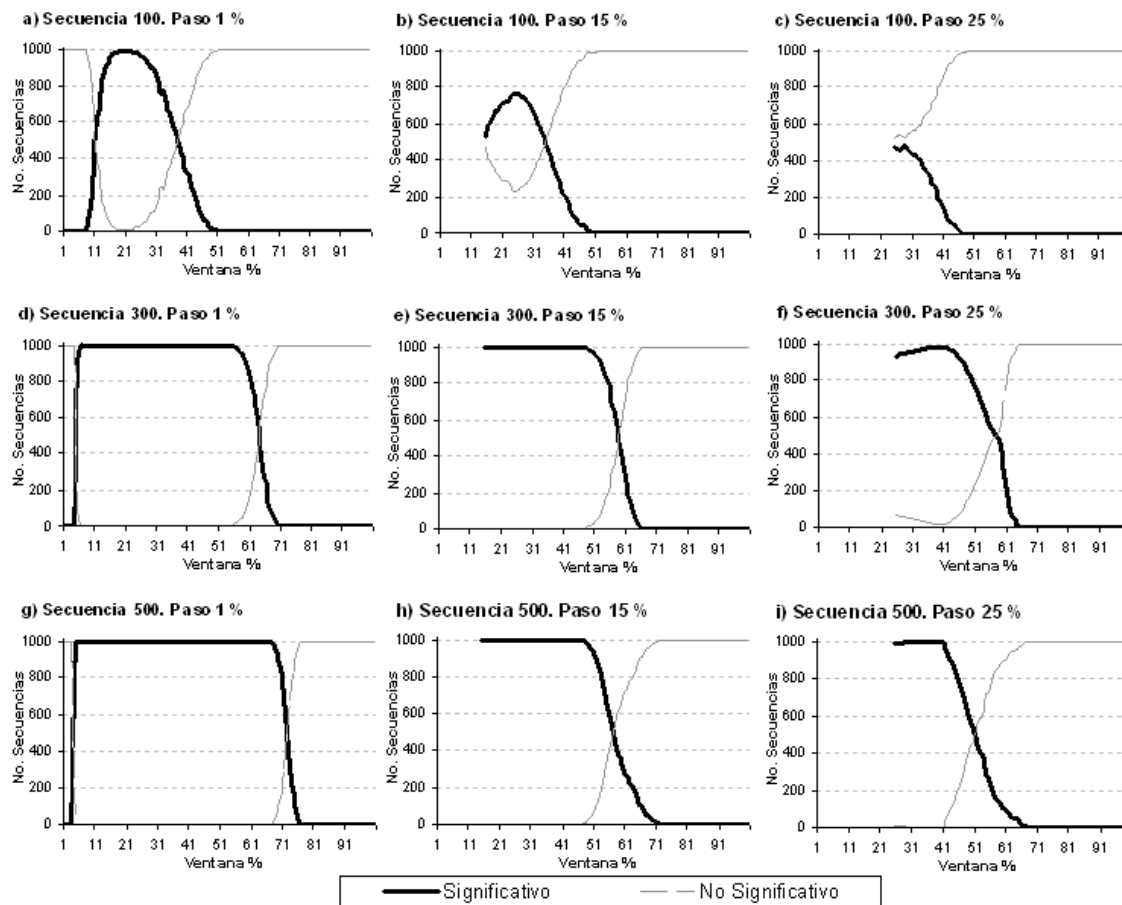


Figura 2.3 Scan Generalizado sobre un círculo en población de secuencias de tamaño 100, 300 y 500 elementos de bases con verdaderos conglomerados creados con el 20% del tamaño total de la secuencia.

El comportamiento de la significación con respecto al tamaño de las ventanas móviles t y el paso de cada juego de datos con verdaderos conglomerados de ambos métodos generalizados se resumió en la Tabla 2.1.

En secuencias de tamaño 100, con verdaderos conglomerados y paso de un 1% se detectan conglomerados a partir de las ventanas móviles con tamaños entre 7% y 82% con respecto al tamaño total de la secuencia. La significación de conglomerados en más del 80% de los casos analizados se logra en las ventanas móviles de tamaños de 11% al 29%. La significación en más del 95% de los casos se logra en un intervalo más estrecho: 14% al 26%. No existe un tamaño de ventana en el que todos los casos sean significativos.

La demás juegos de datos con verdaderos conglomerados en ambos métodos para

cada uno de los pasos se explican de forma similar a las anteriores. Los siguientes rasgos se cumplen en ambos métodos, en las poblaciones con verdaderos conglomerados:

- En las secuencias de igual tamaño teniendo en cuenta el paso los intervalos de ventanas móviles de mayor rango de significación son subconjuntos de los intervalos de menor rango de significación.
- En las secuencia de igual tamaño a medida que el paso aumenta el intervalo de significación de la ventana móvil es subconjunto del paso anterior, para un rango fijo.
- En un paso fijo los intervalos de significación de las secuencias de menos tamaño, son subconjuntos de las bases de datos de mayor tamaño para cada rango.

Tabla 2.1: Rango significativo de las ventanas móviles dado en por ciento en cada población con verdaderos conglomerados, creados con el 20% del tamaño total de la población.

Tamaño de la secuencia	Paso	Significación del Scan Generalizado en rango de las ventanas móviles							
		Scan sobre una línea				Scan sobre un círculo			
		1% o más	80% o más	95% o más	100%	1% o más	80% o más	95% o más	100%
100	1%	[7-82]	[11-29]	[14-26]	---	[8-54]	[13-31]	[16-26]	---
	15%	[15-70]	[21-25]	---	---	[15-51]	---	---	---
	25%	[25-66]	---	---	---	[25-49]	---	---	---
300	1%	[4-89]	[5-41]	[6-40]	[7-38]	[4-71]	[5-61]	[6-57]	[7-50]
	15%	[15-80]	[15-38]	[15-36]	[15-34]	[15-66]	[15-53]	[15-50]	[15-42]
	25%	[25-74]	[25-27] [33-37]	[26-27]	---	[25-64]	[25-46]	[29-41]	---
500	1%	[3-91]	[4-46]	[5-43]	[6-42]	[3-77]	[4-70]	[4-68]	[4-65]
	15%	[15-83]	[15-42]	[15-39]	[15-38]	[15-77]	[15-58]	[15-55]	[15-45]
	25%	[25-76]	[25-40]	[25-28] [32-38]	[25-28]	[25-75]	[25-50]	[25-46]	[36-40]

En general ambos métodos generalizados se comportan de forma similar, se debe destacar que en secuencias con verdaderos conglomerados al aumentar el paso con que se mueve la ventana móvil, en ambas variantes del método, el resultado se debilita, siendo los cambios más bruscos en la variante lineal, observe las Figuras 2.2, 2.3 y la Tabla 2.1 (Anexos 4 y 5). En la Tabla 2.2¹⁸ se calculan el desempeño de cada variante generalizada (suavizado 0) a través de la curvas ROC, donde se comprueba lo planteado anteriormente.

2.2.2.2 Secuencias con falsos conglomerados

Al aplicar ambos métodos del Scan Generalizado a los juegos de datos con falsos conglomerados no se obtienen ningún caso significativo para todas las posibles ventanas móviles de cada juego de datos, esto implica que la curva significativa sea una línea que coincida con el eje que representa el tamaño de la ventana móvil ($y=0$), mientras que la curva relacionada con la no significación sea una línea paralela al eje que representa el tamaño de la ventana móvil y a una distancia de 1000 unidades de este ($y=1000$). Esto ocurre para todas las secuencias con diferentes pasos, por tal motivo no es necesario graficar las mismas.

2.2.3 Algunas consideraciones del estudio con datos simulados

Se deben resaltar los siguientes aspectos en el método Scan Generalizado en sus dos variantes:

- o El tamaño de la ventana móvil influye en los resultados cuando hay verdaderos conglomerados, y se puede señalar que:
 - o Los métodos no son capaces de detectar conglomerados para valores extremos de la ventana móvil, es decir valores muy pequeños (cerca de uno) o valores muy grandes (cerca del tamaño de la secuencia).
 - o Cuando los conglomerados son evidentes¹⁹ aumenta el intervalo del tamaño de la ventana móvil que detecta conglomerado.

¹⁸ Ver epígrafe 2.3.4

¹⁹ Cantidad de unos cercanos en la secuencia binaria es alta

- o Con pasos de tamaño pequeños en por ciento de la longitud total de la secuencia se logran mejores resultados que para pasos mayores.
- o En el caso de falsos conglomerados, el método Scan Generalizado en sus dos variantes resulta ser muy efectivo, independiente al tamaño de la secuencia considerada, el tamaño de la ventana móvil y el paso analizado.

2.3 Los métodos Scan Borrosos

En el epígrafe anterior se demostró que la respuesta del método Scan Generalizado en sus dos variantes con secuencias de verdaderos conglomerados depende fundamentalmente del tamaño de la ventana móvil y del paso. Estos parámetros, de forma general, son difíciles de precisar para encontrar si realmente la secuencia posee algún conglomerado, en el segundo parámetro podemos encontrar la mayor precisión del mismo con paso igual a uno, pero el parámetro ventana móvil tendrá un rango de valores para los cuales encontrará conglomerados si estos existen, estos rangos pueden ampliarse si los datos alrededor del estadígrafo favorecen la formación de conglomerados. Por lo que se propone modificar la ventana de tamaño fijo por otra que al aplicarle una función de pertenencia deje idénticamente la ventana de tamaño fijo pero sus extremos queden pesados por la presencia o no de categoría de interés. De esta forma, se suavizan los extremos y surge el concepto de “ventana móvil borrosa” (Rodríguez et al. 2009). Este método se utilizará en principio sobre secuencias binarias.

2.3.1 El método Scan Borroso sobre una línea

La función de pertenencia le asigna un peso menor que uno a la categoría de interés que se encuentran añadidas en los extremos de la ventana móvil, determinado por la siguiente función de inclusión:

$$\mu_w(k) = \begin{cases} \frac{(i-k+g+1)}{(g+1)} & i = k-g, \dots, g \\ 1 & i = k, \dots, k+t-1 \\ \frac{(k+t+g-i)}{(g+1)} & i = k+t, \dots, k+t+g-1 \end{cases} \quad (2.1)$$

donde:

- ✓ t : longitud de la ventana fija.
- ✓ k : variable que toma valores desde uno hasta $(T - t) / \text{paso} + 1$.
- ✓ g : cantidad de elementos en ambos extremos de la nueva ventana. A esta parte se le llamará “suavizado”.

La nueva ventana se define de la forma siguiente:

$$w'_k = \sum_{i=k-g}^{k+t+g-1} \mu_w(k) * S_i \quad (2.2)$$

donde:

- ✓ $S_1, S_2, \dots, S_n$: secuencia binaria para i desde 1 hasta n .
 - Si $i < 1$ entonces $S_i = 0$
 - Si $i > n$ entonces $S_i = 0$

La formulación matemática del test es esencialmente la misma: el método “escanea” los datos usando una ventana móvil borrosa. Pero ahora, se busca el peso máximo de la categoría de interés reportado en una ventana, por lo tanto este valor puede ser real, lo que lo diferencia del método Scan Generalizado que siempre era un número entero. La Figura 2.4 muestra una representación gráfica de ambas ventanas.

Método Scan ($t=5$)		
	Clásico	Borroso ($g=1$)
Secuencia:	0 1 1 1 0 1 0 1 0 0 0 1 0 1	0 1 1 1 0 1 0 1 0 0 0 1 0 1
Ventana:	3	.5 + 3 + 0
Estadígrafo:	$\eta_{\max} = 3$	$\eta^*_{\max} = 3.5$

Figura 2.4 Ventanas clásica y borrosa en el método Scan sobre una línea.

El valor del estadígrafo se calcula de la siguiente forma:

$$\eta^*_{\max} = \max_{0 \leq k \leq T-t} \{ w'_k \} \in \mathbb{R}^+ \quad (2.3)$$

Se observa en el epígrafe 2.1.1 que el valor de la significación del método Scan sobre una línea se basa en distribuciones de Poisson. Esta distribución está definida para variables aleatorias discretas, entonces para continuar utilizando las fórmulas de Naus (1982) en el cálculo de la significación hay que buscar variantes para calcular la probabilidad puntual ($P[x = \eta]$) y acumulada ($P[x \leq \eta]$) del nuevo estadígrafo real (η^*). Considere a λ como el parámetro de la Distribución de Poisson²⁰, se proponen tres formas diferentes para calcular la significación.

1. Aproximar el valor real al valor entero más próximo. Las distribuciones de probabilidad y de distribución acumulada de Poisson se utilizan de forma similar en las expresiones que están en el epígrafe 1.1.1, donde la probabilidad puntual y acumulada se calculan de la forma: ($P[x = \text{redondeo}(\eta)]$) y ($P[x \leq \text{redondeo}(\eta)]$). De aquí se deduce que la propagación del error pudiera no ser tan pequeña. Se refiere a este método como aproximación borrosa 1, ver Figura 2.5.

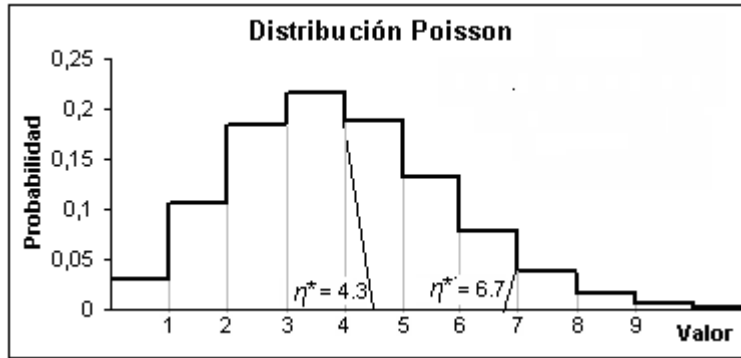


Figura 2.5: Función de probabilidad de Poisson ajustando la aproximación el estadígrafo al valor entero más próximo, (aproximación borrosa 1).

2. Aproximar el valor real usando una combinación de dos distribuciones: Poisson hasta el valor entero inferior y uniforme²¹ para estimar la parte decimal, se refiere a este método como aproximación borrosa 2, ver Figura 2.6:

²⁰ Distribución de Poisson $(f_{(k, \lambda)} = \frac{e^{-\lambda} \lambda^k}{k!} \quad k = \{0, 1, 2, \dots\})$

²¹ Distribución Uniforme utilizada $f_{(x)} = \begin{cases} \frac{e^{-\lambda} \lambda^{n+1}}{(n+1)!} & n < x < \frac{e^{-\lambda} (n+1)!}{\lambda^{n+1}} + n \\ 0 & \text{en los demás casos} \end{cases}$

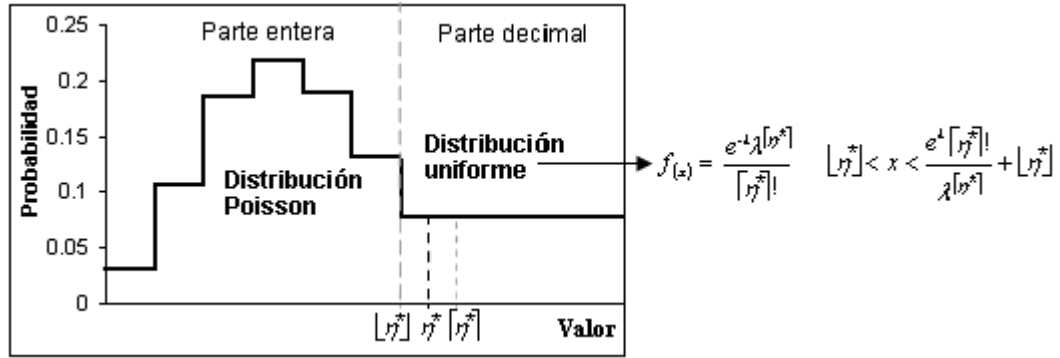


Figura 2.6: Función de probabilidad de Poisson ajustando el estadígrafo a dos distribuciones: Poisson y uniforme, (aproximación borrosa 2).

Las fórmulas originales de Naus (1982), necesitan ser modificadas de la manera siguiente:

- Probabilidad acumulada:

$$P[x \leq \eta^*] = \sum_{n=0}^{\lfloor \eta^* \rfloor} \frac{e^{-\lambda} \lambda^{\lfloor \eta^* \rfloor}}{[\eta^*]!} + \text{parte_decimal}(\eta^*) * \frac{e^{-\lambda} \lambda^{\lceil \eta^* \rceil}}{[\eta^*]!} \quad (2.2)$$

- Probabilidad en un punto, se usa el siguiente factor de corrección porque la fracción se ubica en la distribución continua y la formulas de Naus requieren un valor diferente de cero.

$$P[x = \eta^*] = \frac{e^{-\lambda} \lambda^{\lfloor \eta^* \rfloor}}{[\eta^*]!} + \text{parte_decimal}(\eta^*) * \left(\frac{e^{-\lambda} \lambda^{\lfloor \eta^* \rfloor}}{[\eta^*]!} - \frac{e^{-\lambda} \lambda^{\lceil \eta^* \rceil}}{[\eta^*]!} \right) \quad (2.3)$$

- Se ajusta la formula de A_3 .

$$A_3 = \sum_{r=1+\text{parte_decimal}(\eta^*)}^{\eta^*-1} P[x = 2*\eta^* - r] * P[x \leq r-1]^2 \quad (2.4)$$

- Se ajusta la formula de A_4 .

$$A_4 = \sum_{r=2+\text{parte_decimal}(\eta^*)}^{\eta^*-1} P[x = 2*\eta^* - r] * P[x = r] * ((r-1)P[x \leq r-2] - \lambda[x \leq r-3]) \quad (2.5)$$

3. Aproximar el valor real utilizando funciones de interpolación. La interpolación es un método matemático de “construcción” de nuevos datos a partir de los ya existentes. En nuestro caso, los datos ya existentes se corresponden con las funciones de probabilidad y de distribución de Poisson respectivamente. Se utilizó un polinomio de interpolación de grado 4, por ser sencillo y es el grado

implícito de la función de interpolación del paquete *Matemática*, se refiere a este método como aproximación borrosa 3, ver Figura 2.7.

Para calcular las fórmulas originales de Naus (1982) es necesario modificar:

- Probabilidad puntual: $P[x = \eta^*] = P[\text{int_prob}(\eta^*)]$ (2.6)

- Probabilidad acumulada: $P[x \leq \eta^*] = P[\text{int_acum}(\eta^*)]$ (2.7)

- Se calcula A_3 y A_4 utilizando las formulas 2.4 y 2.5 respectivamente, con la variante que al calcular la probabilidad puntual y acumulada se utiliza las formulas 2.6 y 2.7 según el caso.

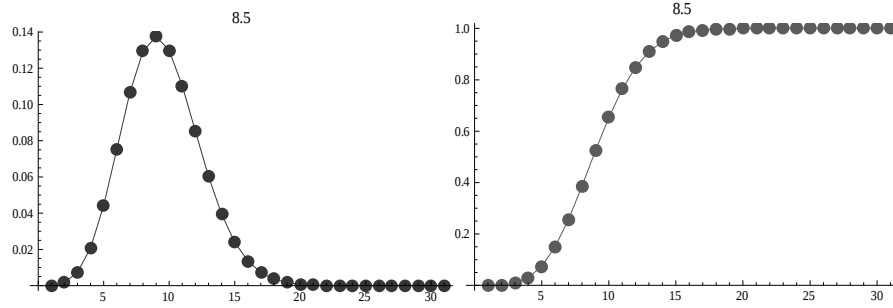


Figura 2.7: Funciones de interpolación para las funciones de probabilidad y de distribución de Poisson, (aproximación borrosa 3).

Finalmente la respuesta del método se “particiona” en dos conjuntos borrosos con las etiquetas: “significativo” y “no significativo”, siendo adecuado en este caso utilizar una función de pertenencia S monótona decreciente y creciente respectivamente para ambos conjuntos borroso, por similitud a los conceptos estadísticos se definen de la forma siguiente:

No significativo:

$$S_{(u, 0.05, 0.0625, 0.075)} = \begin{cases} 0 & u \leq 0.05 \\ 2 * \left(\frac{u - 0.05}{0.025} \right)^2 & 0.05 < u < 0.0625 \\ 1 - 2 * \left(\frac{u - 0.075}{0.025} \right)^2 & 0.0625 \leq u < 0.075 \\ 1 & u \geq 0.075 \end{cases} \quad (2.5)$$

Significativo:

$$S'_{(u, 0.05, 0.0625, 0.075)} = \begin{cases} 1 & u \leq 0.05 \\ 1 - 2 * \left(\frac{u - 0.05}{0.025} \right)^2 & 0.05 < u < 0.0625 \\ 2 * \left(\frac{u - 0.075}{0.025} \right)^2 & 0.0625 \leq u < 0.075 \\ 0 & u \geq 0.075 \end{cases} \quad (2.6)$$

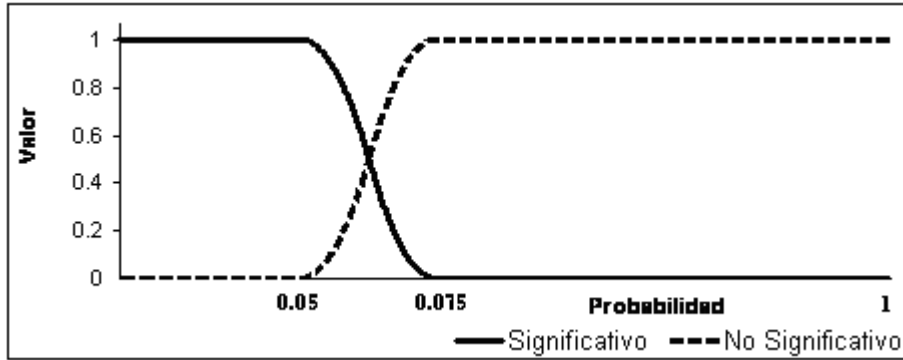


Figura 2.8 Funciones de pertenencia borrosas: “significativo” y “no significativo”.

Se aplica el método del máximo para eliminar el término borroso y obtener una respuesta dura (Martín del Brío y Sánchez 2005).

2.3.2 El método Scan Borroso sobre un círculo

El Scan Borroso sobre un círculo, se obtiene de una forma similar a su equivalente lineal. La ventana móvil se “suaviza” con la misma función de pertenencia definida en la fórmula (2.1) con la variante de que la variable k toma valores desde 1 hasta T , para poder realizar el análisis circular sobre la secuencia. La nueva ventana se define con la fórmula (2.2), interpretándose de la siguiente forma:

- ✓ k : variable que toma valores desde uno hasta T / paso .
- ✓ $S_1, S_2, \dots, S_n, S_{n+1}, S_{n+2}, \dots, S_{n+t-1}$: secuencia formada por:
 - $S_1, S_2, \dots, S_n$: secuencia binaria para i desde 1 hasta n .
 - $S_{n+j} = S_j$: para $j = 1$ hasta $t - 1$

- Si $i < 1$ entonces $S_i = S_{n-i}$
- Si $i > n + t - 1$ entonces $S_i = S_{i-n}$

La formulación matemática del test es esencialmente la misma: la ventana se desplaza por la secuencia circular y contabiliza el peso de la cantidad de categoría de interés en cada ventana suavizada, por tal razón el peso máximo reportado en una ventana no es necesariamente un valor entero, sino real.

De la misma forma que para el método Scan sobre una línea, se definen tres formas diferentes de calcular la significación del test:

- Aproximar el valor real al valor entero más próximo.
- Aproximar el valor real usando una combinación de dos distribuciones: Poisson hasta el valor entero inferior y uniforme para estimar la parte decimal.
- Aproximar el valor real utilizando funciones de interpolación.

La explicación de estas formas es básicamente la misma que se expuso en el epígrafe anterior.

Del mismo modo, la respuesta del método se “particiona” en dos conjuntos borrosos con las etiquetas: “significativo” y “no significativo”. Cada uno de ellos tiene una función de pertenencia S como muestra la Figura 2.8.

Se desborrosifica aplicando el método del máximo para eliminar el término borroso y obtener una respuesta dura. En el Anexo 6 aparecen las funciones más importantes del Scan Borroso sobre una línea que se programaron sobre el paquete *Mathematica*.

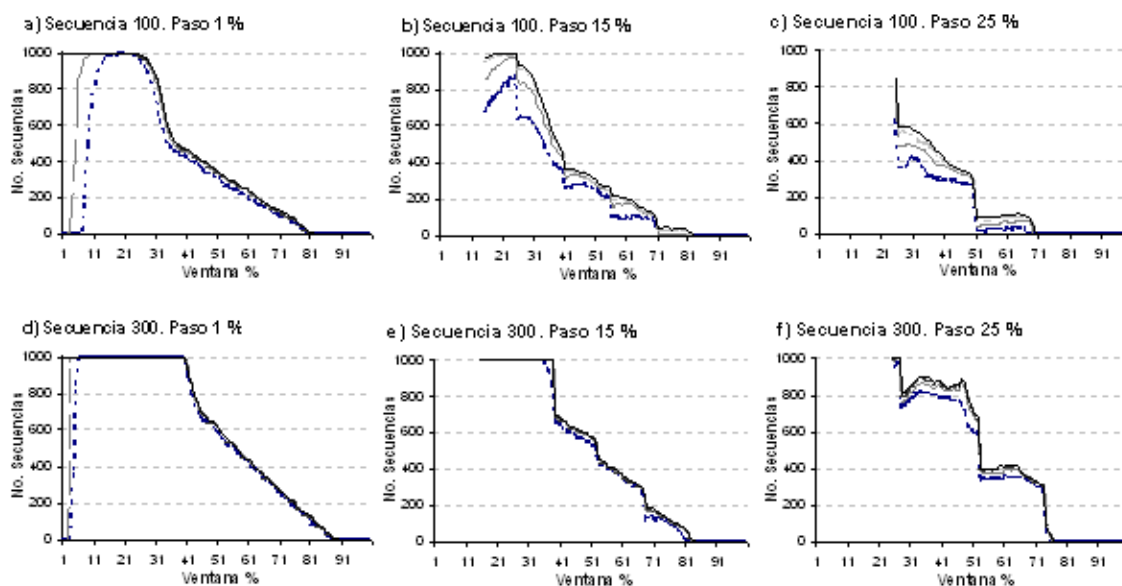
2.3.3 Estudios de simulación

En el epígrafe 2.2 se realizó una intensa simulación de datos para probar prácticamente la generalización de los métodos Scan, con estos mismos juegos de datos de verdaderos y falsos conglomerados se prueban los métodos Scan Borrosos en sus dos variantes. Los conjuntos borrosos significativo y no significativo tienen comportamientos opuestos, razón por la cual se trazan las curvas que representa al conjunto borroso significativo con el objetivo de lograr sencillez en los gráficos.

Los resultados obtenidos para las tres formas de calcular la significación son similares, lo que se muestra en la Tabla 2.2 de los resultados del área bajo la curva ROC de cada uno ellos para cada juego de datos. Por ello se decide mostrar sólo los gráficos de los resultados utilizando la forma de interpolación para calcular la significación con ventana móvil suavizada cero (Scan Generalizado), dos, cuatro y cinco, para la discusión de los resultados separamos los juegos de datos en verdaderos y falsos conglomerados de ambas variantes de los métodos Scan Borroso.

2.3.3.1 Secuencias con verdaderos conglomerados

En ambas variantes del método Scan Borroso no se detectan conglomerados para valores grandes de las ventanas móviles, esto implica que todas las curvas tengan valores nulos al final de las mismas, por esta misma razón la curva del conjunto no significativa termina con valores máximos. Además las curvas del conjunto significativo tienen un máximo para cada uno de los juegos de datos de diferentes tamaños de secuencia, válido para cualquier paso analizado. Los resultados obtenidos para los datos con conglomerados en ambos métodos borrosos se muestran en las Figuras 2.9 y 2.10. En los Anexos 7, 8, 9 y 10 se muestran gráficos con conglomerados creados con diferentes porcentos del tamaño total de la secuencia.



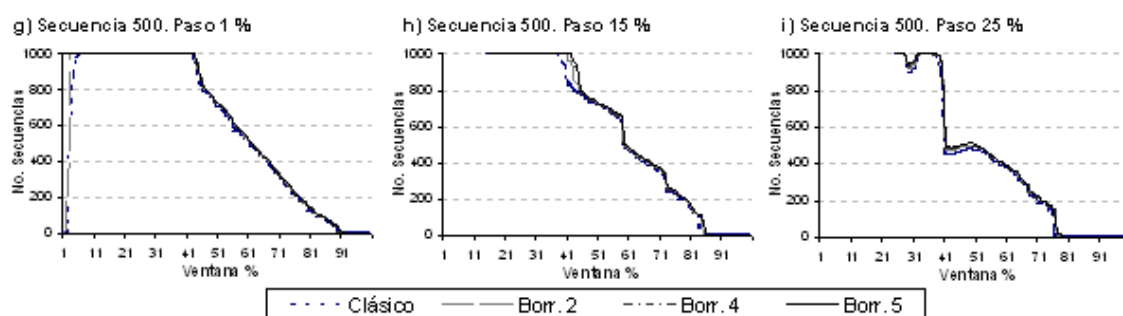


Figura 2.9 Scan Borroso sobre una línea en secuencias de tamaño 100, 300 y 500 elementos con verdaderos conglomerados creados con el 20% del tamaño total de la secuencia.

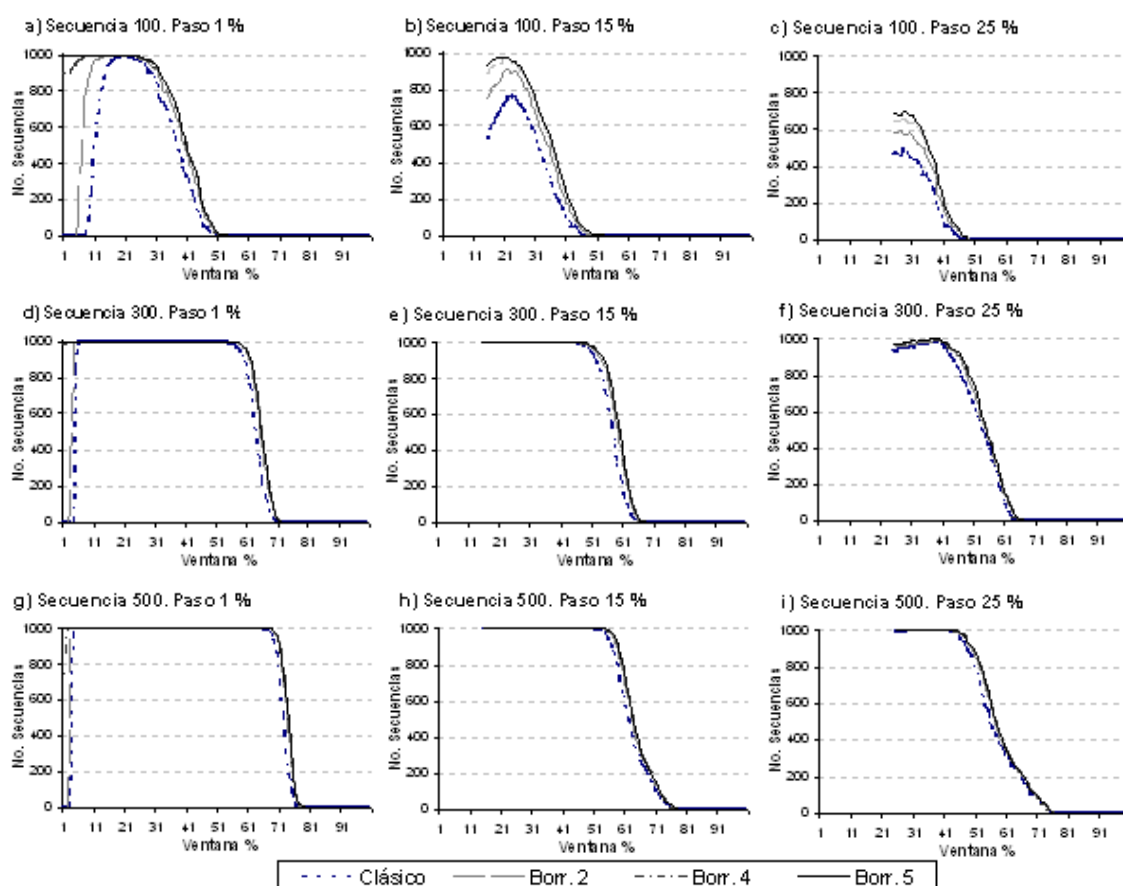


Figura 2.10 Scan Borroso sobre una círculo en secuencias de tamaño 100, 300 y 500 elementos con verdaderos conglomerados creados con el 20% del tamaño total de la secuencia.

Las curvas que representan al conjunto significativo en cada una de los juegos de datos en los diferentes pasos analizados se caracterizan por:

- Una ventana móvil donde la frecuencia absoluta del conglomerado es mayor que las demás, es decir, existe un máximo, el cual tiende a hacer máximo meseta a medida que aumenta la suavidad de la ventana y/o la evidencia de los conglomerados en la secuencias.
- Las curvas con ventana móvil de mayor suavidad tienen mayor frecuencia de secuencias que pertenecen al conjunto borroso significativo que las curvas con ventanas de menor suavidad fundamentalmente para los valores de la ventana móvil pequeño.
- Para valores de la ventana móvil mayores de donde se encuentra el máximo, las curvas de diferentes suavizado tienen comportamiento similares, relativamente algo superior a la curva de suavizado cero (curva determinada por el método Scan Generalizado).
- En el método Scan Borroso sobre una línea, las curvas del conjunto significativos tienen un comportamiento más brusco a medida que aumenta el paso.

Las curvas del conjunto borroso de no significación tiene un comportamiento opuesto a las curvas del conjunto borroso de Significación.

2.3.3.2 Secuencias con falsos conglomerados

En las Figuras 2.11 y 2.12 se observan los resultados de los métodos Scan Borroso en los juegos de datos de la secuencias de falsos conglomerados con secuencias de diferentes tamaños (100, 300 y 500). En todos los casos las curvas que representa la frecuencia absoluta de las ventanas móviles que representa al conjunto borroso significativo se caracterizan por:

- Para suavizado menor o igual a tres son rectas que tienden a confundirse con el eje de las abscisas ($y=0$).
- Para incertidumbre mayor que tres, la frecuencia de las ventanas móviles de tamaño pequeño aumenta a medida que aumenta el valor del parámetro de suavizado (son curvas decrecientes que convergen rápidamente a la recta $y=0$).

Para los casos particulares donde el paso es 15 ó 25% las ventanas móvil comienzan en dichos valores, por lo tanto para estos casos los métodos Scan Borroso con falsos conglomerados tiende a detectar correctamente a la mayoría de los casos, por tal

razón todas las curvas de significación correspondiente a los diferentes suavizados tienden a la recta $y = 0$. Motivo por lo cual sólo se trazan las curvas de significación correspondiente al paso 1% de las diferentes poblaciones.

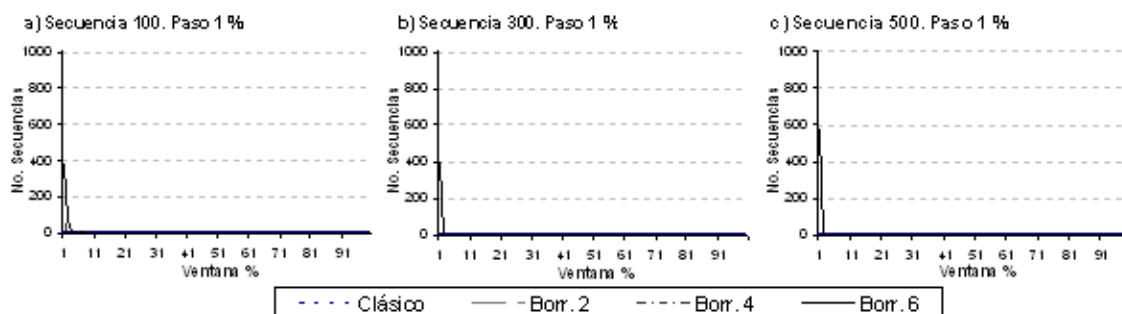


Figura 2.11: Scan Borroso sobre una línea con de falsos conglomerados en secuencias de tamaño 100, 300 y 500 elementos para paso 1%.

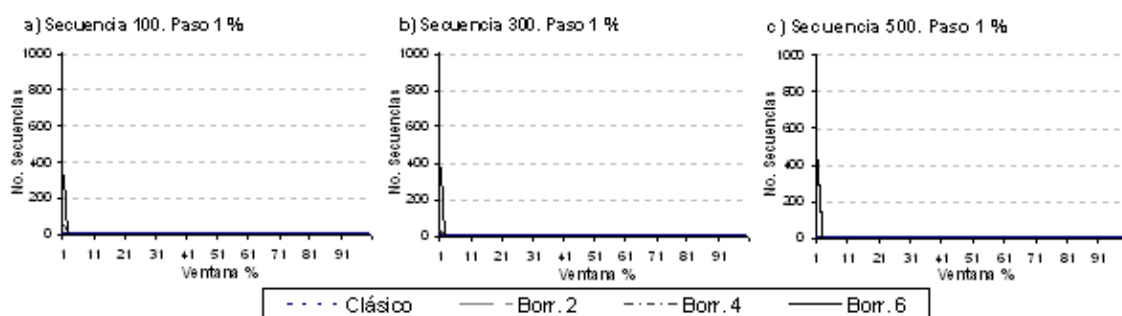


Figura 2.12: Scan Borroso sobre un círculo con falsos conglomerados en secuencias de tamaño 100, 300 y 500 elementos para paso 1%.

Las curvas que representa al conjunto borroso No Significativo en cada una de los juegos de datos con falsos conglomerados tiene comportamiento opuesto a las curvas significativas, es decir son rectas que se confunden con $y=1000$, excepto para el paso 1% para suavizados mayores a tres son curvas crecientes que convergen rápidamente a la recta $y=1000$.

2.3.4 Validar los resultados de la simulación

La detección de conglomerados usando las técnicas del Scan puede considerarse un problema de clasificación. Dada una secuencia de longitud n habrá que determinar si existe o no conglomerados en dependencia de los parámetros utilizado en el método Scan. En particular se calculan las curvas ROC de los juegos de datos con diferentes tamaños de secuencias en todos los métodos mostrados en los epígrafes anteriores,

añadiéndose las tres formas del cálculo del Scan Borroso (Aproximado, Distribución de Poisson y Uniforme e Interpolación de polinomio), las cuales se muestran en un resumen con respecto al suavizado en la Tabla 2.2.

Tabla 2.2: Área por debajo de la curva ROC en secuencias de tamaño 100, 300, 500. Usando las tres variantes para el cálculo de la significación.

Secuencia de tamaño	Paso	Método Scan						
		Suavizado	Sobre una línea			Sobre un círculo		
			Aprox.	Poisson Uniforme	Polinóm.	Aprox.	Poisson Uniforme	Polinóm.
100	1%	0	0.880	0.880	0.880	0.735	0.735	0.735
		2	0.905	0.905	0.915	0.765	0.765	0.770
		4	0.908	0.914	0.912	0.778	0.780	0.777
		5	0.888	0.901	0.892	0.772	0.777	0.774
	15%	0	0.831	0.831	0.831	0.733	0.733	0.733
		2	0.901	0.895	0.895	0.750	0.744	0.744
		4	0.901	0.901	0.901	0.750	0.744	0.744
		5	0.885	0.883	0.883	0.750	0.744	0.750
	25%	0	0.776	0.776	0.776	0.697	0.697	0.697
		2	0.789	0.789	0.789	0.717	0.711	0.711
		4	0.796	0.796	0.796	0.717	0.711	0.711
		5	0.791	0.793	0.793	0.717	0.711	0.711
300	1%	0	0.930	0.930	0.930	0.840	0.840	0.840
		2	0.940	0.940	0.940	0.855	0.855	0.855
		4	0.947	0.949	0.947	0.865	0.865	0.863
		5	0.939	0.945	0.940	0.863	0.863	0.863
	15%	0	0.884	0.884	0.884	0.826	0.826	0.826
		2	0.895	0.895	0.895	0.831	0.831	0.831
		4	0.901	0.901	0.901	0.831	0.831	0.831
		5	0.900	0.900	0.900	0.831	0.831	0.831
	25%	0	0.829	0.829	0.829	0.776	0.776	0.776
		2	0.836	0.836	0.836	0.783	0.783	0.783
		4	0.842	0.842	0.842	0.783	0.783	0.783
		5	0.840	0.840	0.842	0.783	0.783	0.783
500	1%	0	0.945	0.945	0.945	0.875	0.875	0.875
		2	0.950	0.950	0.950	0.880	0.880	0.880
		4	0.955	0.955	0.954	0.890	0.890	0.889
		5	0.950	0.952	0.949	0.889	0.890	0.888
	15%	0	0.901	0.901	0.901	0.866	0.866	0.866
		2	0.907	0.907	0.907	0.866	0.872	0.872
		4	0.919	0.919	0.919	0.872	0.872	0.872
		5	0.918	0.918	0.918	0.872	0.872	0.872
	25%	0	0.842	0.842	0.842	0.836	0.836	0.836
		2	0.849	0.849	0.849	0.842	0.842	0.842
		4	0.855	0.855	0.855	0.842	0.842	0.842
		5	0.855	0.855	0.855	0.842	0.842	0.842

Nota: Suavizado 0 es equivalente a los métodos Scan Generalizado.

En la Tabla 2.2 los siguientes rasgos se cumplen en ambos métodos y en cada una de los diferentes tamaños de secuencias:

- En los métodos del Scan Generalizado al aumentar el paso en una población disminuye su desempeño como clasificador, es decir disminuye el área por debajo de la curva ROC.
- En un juego de datos de secuencia de tamaño fijo y en un mismo paso al aumentar el suavizado mejora su desempeño hasta un suavizado determinado o mantiene su desempeño para cualquier suavizado en dependencia del paso.
- En un juego de datos de secuencia de tamaño fijo y con un suavizado determinado disminuye su desempeño al aumentar el paso.
- Con un paso y un suavizado fijo aumenta su desempeño al aumentar el tamaño de las secuencias de los juegos de datos.

Es de destacar que en todos los casos, cualquiera de las variantes borrosas analizadas, muestra resultados más favorables que la versión clásica correspondiente a ella. Este es un hecho importante porque muestra la superioridad de los métodos borrosos con respecto al lineal clásico (Rodríguez et al. 2007b).

2.3.5 Algunas consideraciones acerca de los métodos Scan Borrosos

La detección del parámetro óptimo para el tamaño de la ventana, o al menos la detección de un parámetro adecuado, sigue siendo un problema no resuelto. En algunas aplicaciones epidemiológicas, en las que se conoce bien el comportamiento de una determinada enfermedad, la selección del ancho de la ventana puede no ser un problema tan grave. Sin embargo, en la mayoría de los estudios bioinformáticos, esta selección a priori no resulta ser tan sencilla. La selección de parámetros no adecuados, puede conllevar a falsas conclusiones.

En el siguiente epígrafe, se explican los fundamentos de un algoritmo de optimización que pretende ayudar a resolver el problema anterior.

2.4 El problema del ajuste de los parámetros

Se han desarrollado numerosos experimentos de simulación en los que se le presentan a los métodos Scan secuencias binarias con verdaderos y falsos conglomerados. Tales

estudios demuestran que los métodos Scan de forma general responden muy bien ante falsos conglomerados. La respuesta de no existencia de conglomerados en esas secuencias es correcta casi en el 100% de los casos, con independencia de los valores de los parámetros utilizados, sólo se incluye falsos positivos para ventanas móvil de longitud muy pequeña cuando el grado de suavizamiento es alto.

Las dificultades surgen al analizar secuencias en las que exista al menos una aglomeración, donde el método Scan Borroso supera al método clásico, pero falla cuando se consideran tamaños de ventanas grandes. Se conoce el comportamiento de los parámetros en secuencias relativamente pequeñas, por lo que es necesario realizar un análisis de diseño experimental bifactorial no paramétrico para analizar si los parámetros se comportan de forma similar cuando las secuencias son extremadamente grandes, que son los casos frecuentes en Bioinformática. Es lógico que si la longitud de secuencias binarias es extremadamente extensa y realmente posee al menos un conglomerado se hace difícil encontrar los parámetros capaces de obtener dicho resultados, para ayudar al investigador se ha ideado utilizar un algoritmo bioinspirado que facilite dicha tarea.

2.4.1 Diseño experimental bifactorial no paramétrico

En los epígrafes anteriores se demostró el comportamiento de los métodos Scan de forma general para secuencias pequeñas (100, 300 y 500), para analizar el comportamiento en poblaciones con secuencias grandes se diseña un experimento bifactorial no paramétrico para ambas variantes de los métodos Scan Generalizado y Scan Borroso, debido a la superioridad del borroso sobre el clásico. Se simularon juegos de datos con secuencias de tamaño 10 000, 100 000 y 1 000 000 de elementos con falsos y verdaderos conglomerados de igual forma que la descrita en el epígrafe 2.2.1, pero como las secuencias son muy grandes los conglomerados se crean con el cinco por ciento de la población total.

El método Scan clasifica si en una secuencia existe al menos un conglomerado de la categoría de interés, por lo que interesa es medir la influencia que produce los parámetros tamaño de la ventana móvil y paso en su desempeño. Por tal razón la información analizada es la exactitud (accuracy) obtenida utilizando el conjunto de verdaderos y falsos conglomerados de cada una de las poblaciones. Con el objetivo de

generalizar los resultados en las distintas poblaciones el tamaño de la ventana y el paso se trabajan de igual forma en por ciento con relación al tamaño de la población.

Con los resultados obtenidos hasta este epígrafe, en los métodos Scan en cualquiera de sus variantes las curvas de desempeño están por encima o alrededor del 50% de elementos bien clasificados, fundamentalmente cuando el paso es pequeño las curvas de desempeño del clasificador con respecto al parámetro ventana móvil tiene un comportamiento cuadrático para la primera mitad de la población, para la segunda mitad de la población el desempeño es pequeño y va decreciendo hasta ser equivalente al 50% a medida que la ventana se acerca al final de la secuencia (Rodríguez et al. 2007b).

Se realizan varios experimentos con los factores tamaño de ventana y paso, con el objetivo de verificar como influyen los factores en cada experimento por separado (Díaz et al. 2009). El factor paso influye en el valor de comienzo del factor ventana móvil, por lo que los niveles de los factores de cada experimento son detallados en la Tabla 2.3.

Tabla 2.3: Niveles de los factores en cada experimento factorial realizados.

Experimento	Niveles de los Factores		Tipo de experimento
	Paso	Ventana Móvil	
Primero	1% y 2%	6%, 25% y 50%.	2 x 3
Segundo	1% y 15%	25% y 50%.	2 ²
Tercero	1% y 25%	25% y 50%.	2 ²

Como se ha demostrado el parámetro suavizado puede influir en los resultados, por lo que es controlado en experimentos para suavizado 0 (Scan Generalizado) y suavizado (Scan Borroso). Cada uno de estos experimentos tiene tres réplicas cada una con probabilidades diferentes de presencia de la categoría de interés en el conglomerado (probabilidad de 0.9, 0.7 y 0.5).

En la figura 2.13 y figura 2.14 se ilustra respectivamente que el método Scan Generalizado y Scan Borroso en ambas variantes, tienen un comportamiento similar en su desempeño en todos sus juegos de datos teniendo en cuenta que:

- El factor ventana móvil aumenta su respuesta para el primer experimento al cambiar sus valores del 6% al 25% y disminuye su respuesta al variar sus valores del 25% al 50% en todos los experimentos.
- El factor paso en el primer experimento (paso con niveles iguales a 1 y 2) tiende a mantener la respuesta al variar de un nivel al otro, en los experimentos restantes este factor disminuye la respuesta de significación al pasar del nivel bajo al alto. A medida que el paso aumenta la respuesta disminuye más rápidamente.

Scan sobre una línea

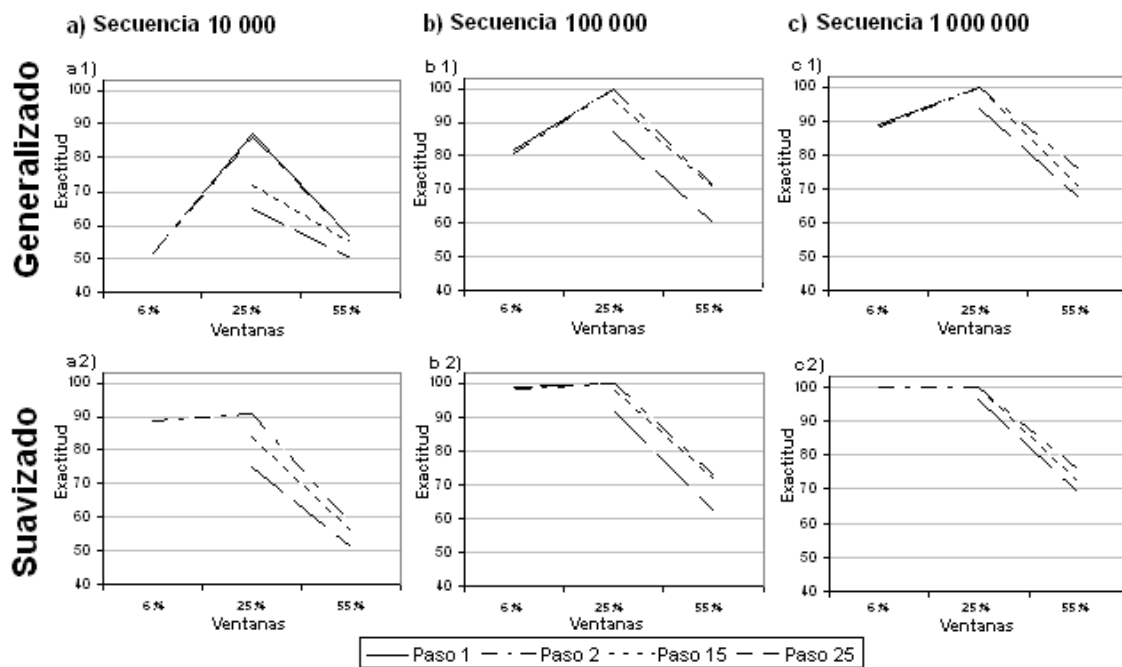
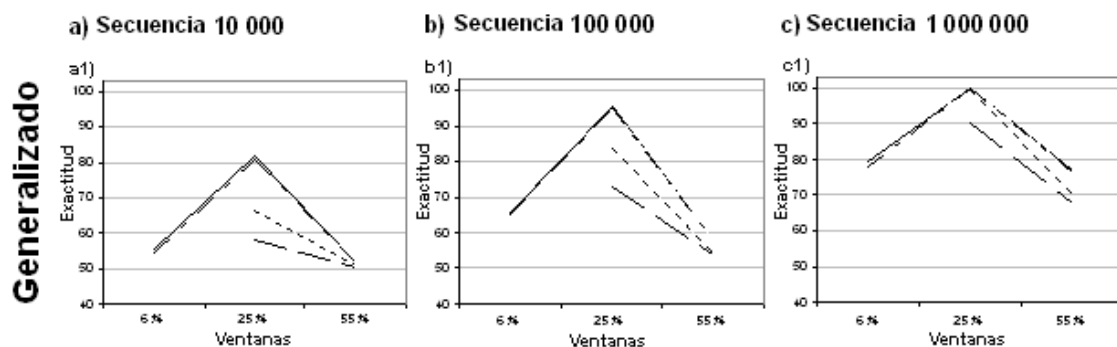


Figura 2.13: Gráfico del factor paso contra el factor ventana móvil en el Scan sobre una línea.

Scan sobre una círculo



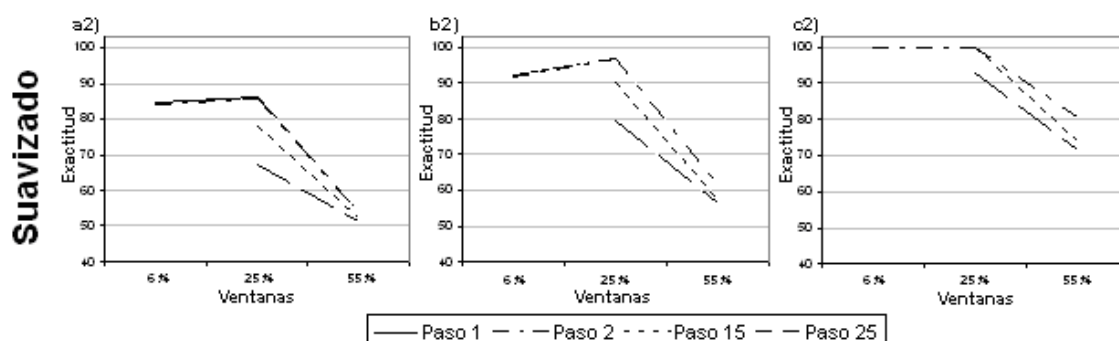


Figura 2.14: Gráfico del factor paso contra el factor ventana móvil en el Scan sobre un círculo.

En el ambas variantes del Scan para cada población la variante suavizada obtiene mejores resultados que la variante clásica, como para todos los niveles del factor ventana móvil la variante borrosa obtiene mejores resultados que la variante clásica, destacándose que el nivel inferior de la ventana en la variante suavizada es la que obtiene un notable aumento de los resultados comparados con los restantes niveles, estos resultados concuerdan con los planteados en (Rodríguez et al. 2008a; Rodríguez et al. 2008c; Rodríguez et al. 2009).

Tabla 2.4: Significación del análisis bifactorial no paramétrico.

Tamaño de la secuencia	Exper.	Scan sobre una línea						Scan sobre un círculo					
		Borroso = 0			Borroso = 4			Borroso = 0			Borroso = 4		
		Vent.	Paso	VxP	Vent.	Paso	VxP	Vent.	Paso	VxP	Vent.	Paso	VxP
10 000	1 ^{ero.}	.001	.757	.998	.003	.860	.992	.001	.724	.998	.003	.825	.986
	2 ^{do.}	.002	.566	.964	.004	.354	.949	.008	.310	.909	.007	.331	.909
	3 ^{ero.}	.004	.047	.843	.014	.077	.924	.019	.038	.447	.010	.145	.834
100 000	1 ^{ero.}	.001	.757	.964	.003	.895	.994	.009	.825	.998	.009	.860	.998
	2 ^{do.}	.005	.233	.612	.005	.354	.685	.012	.480	.998	.012	.310	.883
	3 ^{ero.}	.014	.019	.485	.015	.024	.676	.041	.045	.622	.022	.077	.612
1 000 000	1 ^{ero.}	.001	.860	.992	.003	.965	.986	.003	.930	.992	.010	.930	.998
	2 ^{do.}	.008	.171	.522	.006	.200	.736	.006	.233	.849	.024	.145	.823
	3 ^{ero.}	.025	.015	.349	.019	.024	.504	.040	.024	.587	.085	.038	.504

En la Tabla 2.4 se presenta la significación de los factores ventana, paso y la interacción de ellos en cada uno de los experimentos de las poblaciones de diferentes tamaño de secuencia, se concluye que el total de casos bien clasificados es afectado

en todos sus experimentos significativamente por el factor ventana con una confiabilidad de 90%. Mientras que el factor paso sólo afecta el tercer experimento (paso con niveles iguales a 1 y 25) de todas las poblaciones significativamente con una confiabilidad de 90%; por lo que se corrobora que a medida que el paso crece afecta desfavorablemente el desempeño del clasificador. La interacción de los factores no afecta significativamente a ningún experimento.

Consideraciones generales del diseño experimental bifactorial no paramétrico

Las variantes clásica y borrosa del método Scan se caracteriza por:

- Afectar las respuestas al variar el tamaño de la ventana.
- Respuestas pobre o nula para las poblaciones con verdaderos conglomerados para valores grandes del factor ventana.
- Mejores resultados para ventanas de tamaños cercanos al 25% de la población.
- La variante borrosa aumenta considerablemente su respuesta para valores pequeño del factor ventana con respecto a la variante clásica.
- Los métodos tienden a mantener respuestas similares para valores pequeños del factor paso, pero a medida que el paso aumenta disminuye la respuesta de los métodos, siendo estas diferencias significativas cuando el paso es grande.
- Los métodos en una misma población obtienen mejores respuestas en su variante borrosa que la clásica con respecto al factor ventana o paso. (Rodríguez et al. 2007b).

2.4.2 Algoritmos bioinspirados: optimización basada en enjambre de partículas

La Inteligencia Artificial ha jugado un papel importante como fuente inagotable de técnicas, métodos, modelos y algoritmos tanto para el análisis de datos biológicos como para el modelado y simulación de sistemas biológicos. Técnicas tales como algoritmos evolutivos, autómatas celulares, modelos ocultos de Markov, redes neuronales artificiales y redes bayesianas, resultan ser enfoques ideales para dominios

que se caracterizan por una explosión de datos y muy poca teoría, como es el caso de la Bioinformática.

En la actualidad los modelos bioinspirados se muestran eficientes en la solución de problemas prácticos, y en particular se pretende utilizar la técnica PSO en la búsqueda de parámetros adecuados en las técnicas Scan en general. Este método muestra similitudes con otras técnicas de la computación evolutiva, como los algoritmos genéticos (AG) (Davis 1991), pero no usa operadores de mutación y cruce, y tiene pocos parámetros a ajustar por lo que resulta más fácil de implementar (Beielstein et al. 2002; Mahamed et al. 2005).

Para la aplicación del PSO a la solución del problema de la detección de un parámetro adecuado en el método Scan se siguen los siguientes pasos:

Cada partícula se define por:

- $\mathbf{x}_k^i$ – Es el vector (venta-na móvil, paso, suavizado) en la iteración k , la longitud de las restricciones puede definir las el investigador, aunque las implícitas son las siguientes:
 - $1 \leq \text{Ventana móvil} \leq \text{Tamaño de la secuencia}$
 - $1 \leq \text{Paso} \leq \text{Ventana móvil}$
 - $0 \leq \text{Suavizado} \leq (\text{Ventana móvil}) / 2$
- $\mathbf{p}_k^i$ – Es el mejor vector (mejor ventana móvil, mejor paso, mejor suavizado) de la partícula i , hasta la iteración k .
- $\mathbf{p}_k^g$ – Es el mejor vector (la mejor ventana móvil, el mejor paso, el mejor suavizado) hasta la iteración k .
- $\mathbf{v}_k^i$ – Velocidad de la partícula i en la iteración k . Como se explicó anteriormente, la velocidad se define por:

$$\mathbf{v}_{k+1}^i = \mathbf{v}_k^i + c_1 r_1 (\mathbf{p}_k^i - \mathbf{x}_k^i) + c_2 r_2 (\mathbf{p}_k^g - \mathbf{x}_k^i).$$
- f_k^i – Valor de la función objetivo evaluada en $\mathbf{x}_k^i$.
- f_{best}^i – Mejor valor de la función objetivo evaluada en la partícula i .
- f_{best}^g – Mejor valor de la función objetivo evaluada en el grupo.

Se comprobó la estabilidad del PSO en varias corridas con las mismas secuencias y parámetros diferentes.

2.4.3 Métodos de Monte Carlo combinado con el PSO y los métodos Scan

En este epígrafe se explica el uso de la simulación de Monte Carlo combinada con los algoritmos presentados con anterioridad, para tener una certeza mayor en la respuesta final.

A partir de la secuencia binaria original se pueden “generar” tantas secuencias “similares” como se desee, por ejemplo diez. La generación se hace introduciendo “mutaciones” en la secuencia original, es decir cambiando los valores en algunas de sus posiciones, (Buckley y Jowers 2007).

El investigador controla la cantidad de secuencias mutantes a generar y el “grado de similaridad” con la secuencia original (por defecto 3%). La elección de las posiciones que cambiarán su valor, se realiza al azar, como lo muestra el algoritmo siguiente:

Paso 1: Calcular *cantidad de secuencias mutantes* a generar. (Este valor lo introduce el usuario, diez por defecto).

Paso 2: Repetir hasta *cantidad de secuencia mutantes*:

- a. Calcular *cantidad de posiciones a modificar* (Este valor lo introduce el usuario, 3% por defecto).
- b. Para $i = 1$ hasta *Cantidad de posiciones a modificar* hacer:
 - i. Generar *Posición a cambiar* (Generar un número aleatorio con distribución uniforme entre uno y el largo de la secuencia)
 - ii. $Secuencia[Posición\ a\ cambiar] = 1 - Secuencia[Posición\ a\ cambiar]$
(Donde se cambia valor de de 0 a 1 o viceversa)
- c. Se siguen los pasos de los algoritmos deseados

Paso 3: Terminar.

De esta forma se garantiza que las secuencias generadas sean similares a la original, pues se diferencian de ella en un porcentaje pequeño de sus valores.

Ante secuencias similares, el resultado de cualquiera de los métodos Scan, y del algoritmo PSO para optimizar los parámetros del Scan, no debe diferenciarse demasiado.

La aplicación del método de Monte Carlo fortalece los resultados que el PSO puede hallar, pero aumenta de manera notable el tiempo de ejecución de los algoritmos, sobre todo en caso de secuencias largas.

2.4.4 Resumen de recomendaciones para la selección de valores adecuados para los parámetros

Los resultados experimentales encontrados permiten resumir las recomendaciones para la selección de los valores adecuados de los parámetros en función de la longitud de la secuencia con distintas alternativas:

1^{ero}: Si el tamaño de la secuencia es menor o igual a 500 elementos (secuencias estudiadas minuciosamente) entonces utilizar Scan Generalizado en ambas variantes según caso con:

- *Ventana móvil* = valor entre 20 - 25% de la longitud de la secuencia
- *Paso* = 1

Si hay duda en los resultados utilizar Scan Borroso según caso con:

- *Suavizado* = 3 ó 4

2^{do}: Si el tamaño de la secuencia es mayor a 500 elementos entonces utilizar Scan Generalizado en ambas variantes según caso y aplicar:

- PSO sobre ambos parámetros (*ventana móvil y paso*)
- Si se quiere mayor certeza PSO + Técnicas de Monte Carlo

Si hay duda en los resultados utilizar Scan Borroso en ambas variantes según caso y aplicar:

- PSO sobre los tres parámetros (*ventana móvil, paso y suavizado*)
- Si se quiere mayor certeza PSO + Técnicas de Monte Carlo

2.5 Análisis del comportamiento de los algoritmos

Como los algoritmos que se proponen para reconocer conglomerados tienen estructuras parecidas, sólo se explican detalladamente el análisis teórico de la complejidad algorítmica del Scan Generalizado sobre una línea y en los demás se harán las notaciones necesarias.

Para realizar el análisis de la complejidad temporal se tiene en cuenta los siguientes parámetros:

$t \rightarrow$ longitud de la ventana móvil.

$T \rightarrow$ longitud de la secuencia analizada.

$p \rightarrow$ paso con que se mueve la ventana móvil.

$g \rightarrow$ cantidad de elementos que suaviza la ventana móvil.

Análisis de la Complejidad temporal del Scan Generalizado Lineal

El análisis de la complejidad temporal se hace sobre la base de la descripción por pasos del algoritmo descrito previamente en el epígrafe 2.1.1:

Paso 1: La complejidad temporal es $\Theta(T)$, pues se recorre exactamente la longitud de la secuencia.

Paso 2: Se realizan cuatro operaciones independientes y la suma de los t elementos de la ventana; por lo que su complejidad es $\Theta(t)$.

Paso 3: Al mover la ventana móvil con un paso fijo a lo largo de la línea de longitud y realizar tres operaciones independientes en cada momento su orden de complejidad es un $\Theta(t^*(T-t)/p)$.

Paso 4: La complejidad es $\Theta(1)$, se realiza dos operaciones independientes.

Paso 5: La Significación se calcula por un algoritmo descrito por Naus (1982), la complejidad depende sólo del máximo encontrado (w) en el paso 3 y su complejidad es del orden $O(w^2)$, por lo que se puede acotar superiormente por un $O(t^2)$.

La complejidad general del método modificado es el número de operaciones que se realizan en el algoritmo en los pasos 2 y 3, expresada en la función:

$$C(t, p) = t \left(1 + \frac{T-t}{p} \right) \quad \text{con } 1 \leq p \leq t \leq T$$

Cuyos valores extremos son:

$$C(1, 1) = T; \text{ mínimo.}$$

$$C(T, T) = T; \text{ este valor es despreciado porque está fuera de la fronteras del problema.}$$

$$C(T, 1) = T; \text{ mínimo.}$$

$$C\left(\frac{T+1}{2}, 1\right) = \frac{(t+1)^2}{4}; \text{ máximo.}$$

Los valores mínimos se corresponden con los valores extremos de los parámetros los cuales no obtienen una adecuada solución (observe Figura 2.2), mientras que el valor máximo es precisamente el de mayor complejidad algorítmica.

Esto significa que hay que buscar un compromiso entre ambos factores a la hora de determinar el tamaño de la ventana y del paso. Las pruebas realizadas demuestran que de forma general la mejor opción para la selección de los parámetros del método le corresponde a los valores alrededor del 20 y 25 % de T como la ventana móvil y el paso igual a uno, en dependencia de cómo se encuentra distribuida la secuencia binaria.

Análisis de la complejidad temporal del Scan Generalizado Circular

En este método es necesario añadir al final de la secuencia los elementos del inicio, por lo que solamente varía la cantidad de elementos a analizar de T a $T + t - 1$, quedando el número de operaciones expresado de la forma $t * (1 + (T-1)/p)$; lo que no afecta el orden de la complejidad temporal analizada.

Análisis de la complejidad temporal del Scan Borroso Lineal

Este método utiliza una ventana suavizada descrita previamente en el epígrafe 2.3.1, que provoca que el número de operaciones para cada ventana se incremente de t a $t+2g$. Del epígrafe 2.3.3.2 se obtiene que si el grado de suavizado es grande se

incluyen muchos falsos positivos, y el valor de g debe ser pequeño por lo que su complejidad se aproxima a la del Scan Generalizado Lineal.

Análisis de la complejidad temporal del Scan Borroso Circular

En este método es similar al anterior, añadiéndole al final de la secuencia los elementos del inicio, por lo que solamente se variará en el número de operaciones de Scan Borroso Lineal sustituyendo la variable T por $T + t - 1$, donde el número de operaciones queda expresado como $(t+2g)*(1+T-1)/p$. Como el valor de g es pequeño entonces su complejidad está en el mismo orden de la del Scan Generalizado Circular.

2.6 Consideraciones finales del capítulo

En este capítulo se describen y fundamentan matemáticamente las contribuciones propuestas. Se presentan los algoritmos de los métodos Scan Generalizado en ambas variantes, se enfatiza en sus desventajas y ventajas en sus diferentes variantes. Todo ello se encuentra justificado con estudios de simulación.

Con el objetivo de resumir un conjunto de recomendaciones que puedan ayudar a un investigador inexperto, o a un experto ante un nuevo problema, a seleccionar correctamente los valores de los parámetros de los algoritmos propuestos, se realiza un diseño experimental con dos factores. Para el cálculo de su significación se utiliza una variante no paramétrica de un análisis bifactorial no paramétrico implementada sobre el paquete *Mathematica* por no aparecer en los paquetes estadísticos tradicionales. Se realizan estudios en secuencias muy grandes, en las que resulta imposible realizar estudios de simulación intensivos. Los métodos PSO y de Monte Carlo ayudan también a este propósito y aparecen descritos en el capítulo.

CAPÍTULO III. APLICACIONES A PROBLEMAS BIOINFORMÁTICOS Y BIOMÉDICOS

En este capítulo se describen las implementaciones computacionales realizadas y se presentan tres aplicaciones bioinformáticas: dos sobre los orígenes de replicación del ADN, aplicando el Scan Generalizado sobre una línea en virus y el sobre un círculo en bacterias, la otra aplicación determina la existencia de conglomerados de gaps en el alineamiento de secuencias de ADN del virus de la Influenza A/H1N1, demostrando que los gaps pueden ser la quinta base de un nuevo modelo evolutivo (Grau y Sánchez 2009). Además se muestra otra aplicación real sobre diagnóstico de epidemias, lo que ilustra la factibilidad de usar los algoritmos desarrollados en otras áreas además de la Bioinformática.

3.1 Sobre la implementación de los algoritmos

Se cuenta en el mercado internacional con numerosos productos de software para cubrir las principales funciones y procedimientos de la vigilancia de enfermedades y diferentes tipos de estudios epidemiológicos, muchos de ellos apoyados en el uso de los Sistemas de Información Geográfica (SIG) como herramientas espaciales para fortalecer las capacidades de los mismos (Fernández 2006; Martínez-Piedra et al. 2004; Santovenia et al. 2009). Muchos de ellos tienen implementados diferentes métodos de detección de conglomerados espaciales, temporales y de ambos escenarios, una variante nacional es el EpiDet que contiene estos métodos incluyendo factores de riesgo (Casas 2003). Aunque se tiene acceso a algunos de ellos, estos analizan la secuencia de enfermos en el tiempo, excepto la variante r-Scan explicada en el epígrafe 1.2.2 cuyo análisis se basa en un estudio de casos y controles, por tal razón puede ser fácilmente modificado para otros estudios no relacionados con el tiempo, pero éste no cumple todas las perspectivas propuestas del capítulo II. Además, en múltiples casos estos sistemas tienen un alto precio, debido esencialmente a los beneficios que les reportan a las organizaciones que los utilizan y no puede contarse con los códigos fuentes para realizar algunas modificaciones que mejoren sus resultados. Estas son las causas por las que se hace necesario que las investigaciones desarrollen productos de software para los nuevos modelos que se proponen.

En los inicios de esta investigación, se propuso la generalización de los métodos y la

utilización de la Lógica Borrosa para lograr mejorar los resultados, explicado en el epígrafe 2.1 y 2.3, se necesitaba de forma inmediata la comprobación de su efectividad, alcance y comportamiento de los métodos, programándose sobre el paquete *Mathematica*, plataforma con un conjunto de funciones implementadas de fácil utilización. En esta primera etapa las aplicaciones se dedicaron, fundamentalmente, a problemas de Bioinformática sencillos y a analizar secuencias simuladas de diferentes longitudes, para comprobar cómo se comportaban los diferentes parámetros.

Con el objetivo de validar esta modificación con secuencias de longitudes mayores, mayor paso y mayor grado de suavizado, los métodos al estar programados en un intérprete eran lentos y tediosos, por lo que se reprogramaron en un software libre, Java, lenguaje de propósito general y con tendencia a ser usado por la comunidad científica, simple, orientado a objetos, robusto, de arquitectura neutra, seguro, multihilos, dinámico, etc., pudiéndose ejecutar en cualquier equipamiento que posea sistema operativo Windows o Linux y la máquina virtual de Java. En una primera versión los datos de entrada estaban en un fichero de texto que contiene una o varias secuencias binarias, en dependencia de si se analiza un problema concreto o una población de secuencias de un tamaño fijo para analizar el comportamiento de sus parámetros, los resultados pueden ser obtenidos directamente en la pantalla o en un fichero según requerimiento del usuario.

Con el método Scan Borroso se ha logrado ampliar el rango de significación para el parámetro longitud de la ventana móvil en las secuencias con verdaderos conglomerados, es aún difícil encontrar los parámetros adecuados. Con el objetivo de ayudar a los investigadores a encontrar estos valores, se incorpora el algoritmo de optimización de enjambre de partículas que tendrá como función objetivo el método Scan de forma general y la técnica de Monte Carlo para evitar errores de decisión debido a la posición de los datos.

Se elabora el sistema computacional “**Optimus**”, que incorpora todas las técnicas explicadas en el capítulo II, El sistema utiliza adecuadamente las facilidades de las componentes visuales del lenguaje, en aras de brindar un ambiente cómodo y sencillo. De forma general se encuentran las siguientes facilidades:

- ✓ Los datos de entrada son ficheros textos que poseen una secuencia binaria sin restricciones de longitud.

- ✓ Selección del método Scan a utilizar.
- ✓ Selección del algoritmo PSO (opcional).
- ✓ Selección del método Monte Carlo. (opcional).
- ✓ Otras facilidades generales, tales como guardar los resultados del proyecto, abrir un proyecto, etc.

3.2 Problemas sobre orígenes de replicación del ADN

Los orígenes de replicación²² son los lugares del cromosoma donde se inicia la replicación²³ de las cadenas de ADN. Debido a que la replicación del ADN es el paso central en la reproducción de muchos virus y bacterias, entender los mecanismos moleculares involucrados en este proceso es de gran importancia en las estrategias y vías para controlar el crecimiento y propagación de los mismos (Delecluse y Hammerschmidt 2000). Por ejemplo, para el virus de Epstein-Barr, las réplicas originales han mostrado la asociación con proteínas celulares que regulan la iniciación de la síntesis del ADN en las células humanas (Sugden 2002). Esto sugiere que estas réplicas originales también son importantes para estudiar posibles mecanismos de infección de células de diferentes organismos. El conocimiento de las localizaciones de las réplicas originales reforzará el desarrollo de agentes antivirales, bloqueando la replicación del ADN viral o interviniendo en el proceso de infección.

Debido a que los orígenes de la replicación del ADN son considerados lugares de gran importancia para regular la replicación del genoma en general, se han usado extensos procedimientos en los laboratorios para buscar dichos orígenes en varios organismos (Hamzeh et al. 1990; Newlon y Theis 2002; Zhu et al. 1998). Con la disponibilidad creciente de la secuenciación del ADN del genoma, ya se ha reconocido el valor de usar los métodos computacionales para predecir situaciones posibles de los orígenes de la replicación antes de hacerse los experimentos, aunque hasta ahora no existe ningún esquema para la predicción en el ADN en general. El éxito de la predicción

²² Determinada secuencia de nucleótidos a partir de la cual se desarrolla una horquilla de replicación que dará lugar a dos cadenas idénticas de ADN.

²³ Mecanismo que permite al ADN duplicarse, obteniéndose dos "clones" de la molécula. Esta duplicación se produce de acuerdo con un mecanismo semiconservador donde cada nueva doble hélice contiene una de las cadenas del ADN original.

computacional depende principalmente de la observación de los modelos caracterizados en la secuencias de nucleótidos alrededor de los orígenes de la replicación de los organismos que se están investigando. Por ejemplo, el algoritmo de Salzberg (1998) predijo los orígenes de la replicación para varias bacterias, basándose en el hallazgo de oligómeros de siete u ocho bases cuya orientación está, preferentemente, sesgada alrededor de las réplicas originales. Sin embargo y como lo especifica el autor, este algoritmo no está preparado para las moléculas de ADN donde existen múltiples orígenes de la replicación, como ocurre en muchos virus y organismos eucariotas. En estos casos, se necesitaría confiar en otros modelos de patrones de secuencia para localizar dichos orígenes (Service y Tauritz 2009; Wolpert y Macready 2005).

3.2.1 Concentraciones de palíndromos en los orígenes de replicación del ADN en herpesvirus

En algunos estudios se ha reportado la existencia de altas concentraciones de palíndromos en la proximidad de los orígenes de la replicación de herpesvirus (Masse et al. 1992; Reisman et al. 1985; Weller et al. 1985). Este fenómeno se le atribuye, generalmente, al hecho de que la iniciación de la replicación del ADN requiere normalmente de un agrupamiento de enzimas para desmontar la estructura helicoidal del ADN y separar las dos cadenas complementarias. Masse et al. (1992) ha demostrado que a través de la existencia de clusters de palíndromos se predicen regiones que contienen los orígenes de replicación.

La Figura 3.1 (a) muestra que los palíndromos son palabras simétricas de ADN, en el sentido que ellos pueden leerse exactamente igual que al leer las secuencias complementarias en la dirección inversa. Es importante señalar (Figura 3.1 (b)) que la longitud en un palíndromo de ADN tiene, necesariamente, que ser un cordón de nucleótidos par ($2L$), para que cada porción L del cordón pueda tener su complemento.

En el artículo “Nonrandom Clusters of Palindromes in Herpesvirus Genomes” de Leung (2005), se estudia una colección de genomas de 16 herpesvirus, donde se identifican clusters de palíndromo no aleatorios utilizando el método r-Scan que calcula su significación estadísticas con la distribución binomial (Dembo y Karlin 1992; Glaz 1989) descrita brevemente en el epígrafe 1.2.2, donde cada una de las posiciones de los elementos de la secuencia de ADN son independientes e idénticamente distribuidos,

implicando que la ocurrencia de palíndromos puede aproximarse por un proceso de Poisson utilizando una cota superior obtenida por la distancia de Wasserstein (Barbour et al. 1992). Esta cota se usa como una guía para escoger la longitud óptima (L) del palíndromo en el análisis de la colección de cada genoma de los 16 herpesvirus.

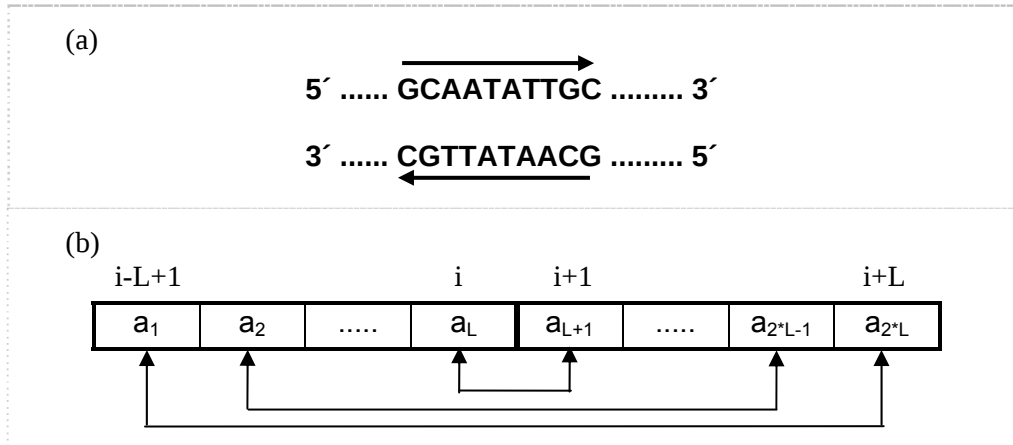


Figura 3.1: Palíndromo de ADN.

- (a) Se muestra una secuencia palíndromo de nucleótidos con sus dos cuerdas complementarias de ADN, que se lee en las direcciones de 5' a 3' como lo señalan las flechas. Los segmentos se leen exactamente igual en ambas cuerdas.
- (b) En cada cuerda, la primera base del palíndromo es complementaria a la última, la segunda a la segunda última, y así sucesivamente. Ésta es una representación esquemática de este tipo de apareamiento complementario entre las bases en un palíndromo $2L$ centrado en la base i .

Análisis de los datos

Las bases de datos comprenden todas las secuencias completas del genoma de la familia del herpesvirus, cargadas del GenBank del sitio NCBI²⁴. En la Tabla 3.1 se muestra el listado con cada nombre del virus y su abreviatura, identificación de la base de datos del GenBank, longitud de la secuencia del genoma en número de bases, las probabilidades pA , pC , pG , pT de las cuatro bases de nucleótidos del genoma y la longitud mínima (L) de los palíndromos obtenida por el límite superior de la distancia de Wasserstein, de forma tal, que cada secuencia genómica puede lograrse por un

²⁴ National Center for Biotechnology Information, EE.UU.

proceso de Poisson, capítulo 10 de (Barbour et al. 1992).

Tabla 3.1. Lista de los genomas de los Herpesvirus Analizados.

	Nombre	Abrev.	Registro	Longitud	Prob. bases	Valor L
1	Alcelaphine herpesvirus1	AHV1	NC_002531	130608	(.27, .24, .22, .26)	5
2	Ateline herpesvirus 3	AtHV3	NC_001987	108409	(.32, .19, .17, .31)	5
3	Bovine herpesvirus 1.1	BHV1	NC_001847	135301	(.14, .36, .37, .14)	6
4	Equine herpesvirus 1	EHV1	NC_001491	150223	(.22, .29, .28, .22)	5
5	Equine herpesvirus 4	EHV4	NC_001844	145597	(.25, .25, .25, .25)	5
6	Gallid herpesvirus 1	MDV2	NC_002530	110637	(.24, .25, .25, .25)	5
7	Gallid herpesvirus 2	MDV	NC_002229	138675	(.28, .22, .21, .29)	5
8	Human herpesvirus 1	HSV1	NC_001806	152261	(.16, .34, .34, .16)	6
9	Human herpesvirus 2	HSV2	NC_001798	154746	(.15, .35, .35, .15)	6
10	Human herpesvirus 3	VZV	NC_001348	124884	(.27, .23, .23, .27)	5
11	Human herpesvirus 4	EBV	NC_001345	172281	(.20, .30, .29, .20)	5
12	Human herpesvirus 5	HCMV	NC_001347	229354	(.22, .28, .29, .21)	5
13	Human herpesvirus 6	HHV6	NC_001664	159321	(.29, .22, .21, .29)	5
14	Human herpesvirus 7	HHV7	NC_001716	144861	(.32, .18, .17, .32)	5
15	Ictalurid herpesvirus	CCV1	NC_001493	134226	(.21, .28, .28, .22)	5
16	Saimiriine herpesvirus 2	HVS2	NC_001350	112930	(.33, .18, .16, .32)	5

De las anotaciones de las secuencias del GenBank y las referencias de los mapas genéticos y otros artículos biomédicos (Masse et al. 1992) se compilaron una lista de orígenes de replicación en 10 de los 16 herpesvirus. Éstos incluyen un herpesvirus en la vaca, dos en el caballo, y siete en los humanos. Estos virus se han estudiado más que los otros debido a su importancia agrícola y médica. Las localizaciones de estos orígenes muestran en la Tabla 3.2, indicándose los clusters significativos con el número de palíndromos que contienen y por último los resultados cercanos entre las regiones de rupturas y los clusters significativos encontrados. Las filas de la Tabla 3.2 indican cada uno de los genomas de los 16 herpesvirus, en la parte superior de cada fila están los resultados obtenidos por Leung (2005) y en la parte inferior se encuentran los resultados obtenidos por el Scan Generalizado sobre una línea.

Leung (2005) al usar el r-Scan en los genomas de los herpesvirus HSV1 y VZV no encuentra clusters significativos que contengan a los orígenes de replicación, pero plantea que en un análisis más detallado estos sitios se encuentran dentro de palíndromos de longitudes grandes. Al aplicar en método Scan Generalizado se encontraron clusters significativos en estos dos genomas que coinciden con los orígenes de replicación.

Tabla 3.2: Localización de los orígenes de replicación de los Herpesvirus.

#	Nombre	GenBank	Orig.Replicación	Clusters	#P	Coincidencia
1	AHV1-5	NC_002531	No conocida	113456 - 113759	5	
				112518 - 113759	8	
2	AtHV3-5	NC_001987	No conocida	95350 - 100098	17	
				95817 - 100330	17	
3	BHV1-6	NC_001847	111080 - 111300 126918 - 127138	77155 - 77168	3	1.61 μ del origen 1.75 μ del origen
				102895 - 106948	22	
3	BHV1-6	NC_001847	111080 - 111300 126918 - 127138	113462 - 113636	5	1.61 μ del origen 1.75 μ del origen
				124582 - 124756	5	
3	BHV1-6	NC_001847	111080 - 111300 126918 - 127138	131268 - 135221	21	1.61 μ del origen 1.75 μ del origen
				77156 - 77171	3	
3	BHV1-6	NC_001847	111080 - 111300 126918 - 127138	102897 - 106945	22	1.67 μ del origen 1.75 μ del origen
				113464 - 113635	5	
3	BHV1-6	NC_001847	111080 - 111300 126918 - 127138	124583 - 124754	5	1.67 μ del origen 1.75 μ del origen
				131273 - 135235	21	
4	EHV1-5	NC_001491	126187 - 126338	115125 - 119094	17	
				144064 - 148033	17	
4	EHV1-5	NC_001491	126187 - 126338	115127 - 119095	17	
				144065 - 148033	17	
5	EHV4-5	NC_001844	73900 - 73919 119462 - 119481 138568 - 138587	No Existen		
				No Existen		
6	MDV2-5	NC_002530	No conocida	93143 - 93243	4	
				109331 - 110590	8	
6	MDV2-5	NC_002530	No conocida	93143 - 93243	4	
				109331 - 110590	8	
7	MDV-5	NC_002229	No conocida	No Existen		
				106 - 475	7	
7	MDV-5	NC_002229	No conocida	141145 - 142428	9	
				176016 - 177299	9	
8	HSV1-6	NC_001806	62475 131999 146235	No Existen		
				62470 - 82905	30	Contiene origen de replicación
8	HSV1-6	NC_001806	62475 131999 146235	126339 - 126354	3	3.71 μ del origen
				151881 - 151896	3	3.72 μ del origen
9	HSV2-6	NC_001798	62930 132760 148981	No Existen		
				No Existen		
10	VZV-5	NC_001348	110087 - 110550 119547 - 119810	No Existen		
				110196 - 110738	26	0.12 μ del origen
10	VZV-5	NC_001348	110087 - 110550 119547 - 119810	119181 - 119701	26	Contiene origen de replicación
				6772 - 11675	19	Contiene origen de replicación
11	EBV-5	NC_001345	7315 - 9312 52589 - 53582	49460 - 54858	25	Contiene origen de replicación
				6772 - 11675	19	Contiene origen de replicación
11	EBV-5	NC_001345	7315 - 9312 52589 - 53582	49460 - 54858	25	Contiene origen de replicación
				89585 - 94183	19	Contiene origen de replicación
12	HCMV-5	NC_001347	92270 - 93715	195029 - 195268	8	
				91182 - 94541	17	Contiene origen de replicación
12	HCMV-5	NC_001347	92270 - 93715	195966 - 196205	6	
				No Existen		
13	HHV6-5	NC_001664	67617 - 67993	No Existen		
				No Existen		
14	HHV7-5	NC_001716	66685 - 67298	120758 - 124422	16	
				124986 - 128652	16	
15	CCV1-5	NC_001493	No conocida	No Existen		
				No Existen		
16	HVS2-5	NC_001350	No conocida	No Existen		
				109081 - 112860	16	

Nota: μ unidad de medida que representa 1% de la longitud del genoma. Esta distancia es calculada del punto medio de la región del cluster, al punto medio más cercano al origen de replicación.

En la Tabla 3.3 se resumen los resultados de ambos métodos en los diez herpesvirus que se conocen los orígenes de replicas, se observan porcentajes ligeramente superiores a favor del Scan Generalizado.

Tabla 3.3: Resultados de utilizar los métodos r-Scan y Scan Generalizados en los 10 herpesvirus donde se conocen los orígenes de la replicación.

Herpesvirus (10)	r-Scan		Scan Generalizado	
	Número	Por ciento	Número	Por ciento
- Con <u>clusters</u> significativo	5	50.00	7	70.00
- Coincidencias de <u>cluster</u> con orígenes de la replicación	3	30.00	5	50.00
		60.00*		71.43*
Cantidad de clusters				
- Significativos	12		17	
- Coincidencias de <u>cluster</u> con orígenes de la replicación	5	41.67**	10	58.82**

Nota: * Por ciento con respecto a la cantidad de Herpesvirus con clusters significativos

** Por ciento con respecto a la cantidad de clusters significativos

3.2.2 Patrones específicos alrededor de los orígenes de replicación en bacterias

Se han publicado numerosos estudios relacionados con el ADN de la *Escherichia coli*, en ellos se ha determinado que 245 pb es la secuencia más corta en la cual se puede encontrar el origen de replicación del ADN de esta bacteria, región muy conservada que se caracteriza entre otras por la existencia de conglomerados de sitios Dam, lo que implica un cluster de cuartetos de nucleótidos en el orden **GATC**, otorgándole una importancia especial desde el punto de vista bioquímico, (Cardellá y Hernández 1999; Glaz y Balakrishnan 1999; Hénaut et al. 1996; Karlin y Brendel 1992). Esta cuarteta es un palíndromo de $L=2$, y a diferencia de lo que ocurre en herpesvirus, es el único patrón característico de clusters de palíndromos.

El ADN de la *E. coli* es circular y tiene una longitud aproximada de 4.7 millones de pares de bases, por ese motivo se aplicarán las variantes circulares del método Scan

Generalizado y Borroso.

Análisis del ADN circular *E. coli* usando los métodos Scan

Por los estudios de laboratorio se puede determinar que el parámetro ancho de la ventana móvil puede ser igual a 245 elementos (Langrand 2005), no se tiene información acerca de los valores posibles de los demás parámetros, por lo que se decidió tomar el paso igual a la unidad y la parte borrosa de la ventana (en el caso de los métodos borrosos), como los valores 2 y 4. La Tabla 3.4 muestra los resultados obtenidos.

Tabla 3.4: Resultados obtenidos con el Scan sobre un círculo y parámetro “paso” igual uno

<i>Escherichia coli</i> IAI1, GenBank, NC_011741, 4.7Mb			
Ancho de la ventana móvil: 245bp			
Scan sobre un círculo	# 'GATC'	Resultado	Localización
- Generalizado	14	p = 0.00	4002141 - 4002422
- Borroso (g=2)	14	Significativo	4002141 - 4002422
- Borroso (g=4)	14	Significativo	4002141 - 4002422

Los valores de la significación demuestran la existencia de conglomerados de sitios Dam dentro del genoma de la *E. coli* localizados en las bases 4002141 – 4002422 que a su vez contiene el origen de replicación de la *E.coli* situados en las bases 4002160-4002400.

Análisis de la secuencia de *E. coli* usando los métodos Scan y el PSO

Si se supone que no se conoce un valor adecuado para los parámetros de los métodos Scan y que se desea de la misma forma, determinar la existencia de conglomerados de sitios Dam dentro del genoma de la *E. coli*.

Los resultados de la aplicación del método Scan con el PSO aparecen recogidos en la Tabla 3.5. Puede observarse que en ambos casos, los valores hallados para el tamaño de la ventana son inferiores a 245, pero en ambos se demuestra la existencia de conglomerados de sitios Dam que era el objetivo fundamental de la aplicación. Estos conglomerados formados en cada caso están alrededor de los pares de bases

4002141 – 4002422.

Tabla 3.5: Resultados obtenidos utilizando conjuntamente el Scan sobre un círculo, PSO y métodos de Monte Carlo.

Escherichia coli IAI1, GenBank, NC_011741, 4.7Mb											
10 partículas (posición 1-300) 10 iteraciones						+	5 mutaciones				
Scan sobre un círculo	PSO					PSO+ Monte Carlo					
	S*	Vent	Pas.	DAM	Resultado	S*	Vent	Pas	DAM	Resultado	
- Generalizado	0	258	67	13	p = 0.00	0	250	53	14	p = 0.00	
- Borroso	4	265	206	11	Significativo	6	262	31	15.2	Significativo	

Nota;

- S* grado de suavizado utilizado en el Scan Borroso.

3.3 Problemas sobre alineamiento de secuencias

Un alineamiento de secuencias en bioinformática es una forma de representar y comparar dos o más secuencias o cadenas de ADN, ARN, o estructuras primarias proteicas para resaltar sus zonas de similitud, que podrían indicar relaciones funcionales o evolutivas entre los genes o proteínas consultadas. Las secuencias alineadas se escriben con las letras (representando aminoácidos o nucleótidos) en filas de una matriz en las que, si es necesario, se insertan espacios para que las zonas con idéntica o similar estructura se alineen (Brudno et al. 2003; Schneider y Stephens 1990).

El alineamiento múltiple de secuencias es una extensión del alineamiento de pares que incorpora más de dos secuencias al mismo tiempo. Los métodos de alineamiento múltiple intentan alinear todas las secuencias de un conjunto dado. Se usa a menudo en la identificación de regiones conservadas en un grupo de secuencias que hipotéticamente están relacionadas evolutivamente. Los alineamientos múltiples son también utilizados para ayudar al establecimiento de relaciones evolutivas mediante la construcción de árboles filogenéticos. Los multi-alineamientos múltiples perfectos son computacionalmente difíciles de producir pues exigen la solución de problemas de optimización combinatoria NP-completos. Sin embargo, su utilidad en bioinformática ha dado lugar al desarrollo de una variedad de métodos o heurísticas suficientemente

adecuados para la alineación de varias secuencias, que aún cuando no producen una solución óptima, brindan un resultado bastante bueno, que además puede ser “retocado” manualmente por un especialista experimentado.

Secuencias muy cortas o muy similares pueden alinearse manualmente. Pero los problemas más interesantes necesitan alinear secuencias largas, muy variables y extremadamente numerosas que no pueden ser alineadas por humanos. Existen diferentes productos de software en Internet que realizan el alineamiento de secuencias, como el *Mega4* (Tamura K 2007) y el *Clusta/W* (Thompson et al. 1994).

Producto de la alineación de varias secuencias de ADN se necesitan ciertos desplazamientos de bases dentro de las secuencias y surgen “espacio vacíos” a los que se les denomina gaps. Cuando se trata del multialineamiento en un estudio evolutivo, los gaps pudieran representar mutaciones tipo “indel”, esto es mutaciones que consisten en la “inserción” o “delección” de bases en un momento dado.

Tradicionalmente los estudios evolutivos requerían prescindir de las zonas del multialineamiento donde aparecían gaps. Así se hace por ejemplo cuando se desea utilizar el modelo clásico evolutivo de Tamura y Nei (1993). Pero evidentemente esta simplificación del problema está descartando información que podría ser importante. Ello motivó al Grupo de Bioinformática de la Universidad Central de Las Villas, a desarrollar un nuevo modelo evolutivo basado en cinco bases: las cuatro del ADN y el “gap” y demostrar la factibilidad de su aplicación en la construcción de árboles filogenéticos más verosímiles (Sánchez y Grau 2009) y en el desarrollo de estudios evolutivos, que incluyen por ejemplo, el pronóstico de las mutaciones del virus de la influenza (Grau y Sánchez 2009).

La distribución de los gaps que surgen producto de la alineación, dentro de la secuencia no es el mismo que el de las cuatro bases nucleotídicas. Al parecer ellos tienden a aparecer en zonas concentradas. Es interesante teóricamente comprobar la existencia de conglomerados de gaps dentro de secuencias alineadas. Pero además puede ser importante prácticamente. Por ejemplo una vez predichas las mutaciones de un virus como el de la influenza, las zonas del mutante donde se concentran tales gaps tienen que ser descartadas como “blancos” o “dianas” de sistemas de diagnóstico, virus o antivirales.

Para comprobar estadísticamente la existencia de conglomerados de gaps en el

alineamiento de mutaciones se alinearon los genomas completos de 167 mutantes del virus de la influenza A H1N1, obtenido de Internet. Las bases nucleotídicas se sustituyeron por 0, mientras que los “gaps” se sustituyeron por 1. Con esas secuencias binarias se ejecutaron los métodos Scan Generalizado y Borroso sobre una línea. Para ambos algoritmos se obtuvieron los resultados mostrados en la Tabla 3.6. En ella se muestran las secuencias que tienen conglomerados mayores de 20 gaps. Cada celda se divide en dos: número (#) y por ciento (%). El # contiene la cantidad de secuencias con los conglomerados correspondientes. El porcentaje a su vez se divide en dos valores, el superior es el porcentaje por columnas y el inferior por filas. Por ejemplo, la celda que resulta de la intercepción entre la fila 450-499 y la columna correspondiente a 9 conglomerados tiene 26 secuencias con 9 conglomerados de más de 20 gaps. Este número representa el 53.06% de secuencias con 9 conglomerados y el 96.30% de las secuencias que poseen de 450 a 499 del total de gaps.

Tabla 3.6: Resultados del virus de la influenza A H1N1 en 167 genomas con longitud de 14158 pares de bases

Total de <u>Gaps</u>	Cantidad de conglomerados mayores o iguales a 20 <u>gaps</u> .										Total	
	6		8		9		10		14		15	
	#	%	#	%	#	%	#	%	#	%	#	%
400-449			3	$\frac{100.00}{100.00}$							3	$\frac{1.80}{100.00}$
450-499					26	$\frac{53.06}{96.30}$	1	$\frac{0.92}{3.70}$				$\frac{16.17}{100.00}$
500-549	1	$\frac{100.00}{3.57}$			22	$\frac{44.90}{78.57}$	5	$\frac{4.59}{17.86}$				$\frac{16.77}{100.00}$
550-599					1	$\frac{2.04}{1.75}$	56	$\frac{51.38}{98.25}$				$\frac{34.13}{100.00}$
600-649							35	$\frac{32.11}{100.00}$				$\frac{20.96}{100.00}$
700-749							1	$\frac{0.92}{100.00}$				$\frac{0.60}{100.00}$
750-799							11	$\frac{10.09}{100.00}$				$\frac{6.59}{100.00}$
800-849									2	$\frac{100.00}{66.67}$	1	$\frac{33.33}{3.33}$
850-899											1	$\frac{33.33}{100.00}$
900-949											1	$\frac{33.33}{100.00}$
Total	1	$\frac{100.00}{0.60}$	3	$\frac{100.00}{1.80}$	49	$\frac{100.00}{29.34}$	109	$\frac{100.00}{65.27}$	2	$\frac{100.00}{1.20}$	3	$\frac{100.00}{1.80}$
											167	$\frac{100.00}{100.00}$

Como puede apreciarse, los resultados fueron altamente significativos en todos los casos. La cantidad de “gaps” oscila de 435 a 914 y alrededor del 68% de la secuencias tienen 10 o más conglomerados cada uno de ellos con 20 o más gaps consecutivos.

El trabajo fue replicado con subsecuencias más cortas pero en mayor número. Específicamente se trabajó con secuencias de dos de los segmentos del virus que representan los principales sitios antigénicos, los correspondientes a las proteínas Hemaglutinina (HA) y Neuraminidasa (NA). Ellas son especialmente importantes pues constituyen el blanco hacia el cual se dirigen los antivirales o vacunas y sus eventuales mutaciones pueden reducir o inhibir la unión de anticuerpos neutralizantes.

En ambos casos se obtuvieron resultados similares a los del genoma completo, lo cual demuestra que los conglomerados pueden aparecer efectivamente en las mutaciones de estos sitios de antigénicos.

Así se comprueba que efectivamente existen conglomerados de “gaps” en las secuencias alineadas, lo que desde el punto de vista bioinformático, era lo que se quería demostrar. La información sobre la localización de los gaps en mutaciones futuras del H1N1 se añade a la localizaciones más conservadas de los sitios del genoma de la HA y es usada hoy en día por el Centro Nacional de Salud Agropecuaria (CENSA) de La Habana en el análisis de la efectividad del sistema de diagnóstico y su perfeccionamiento.

3.4 Problemas sobre detección de conglomerados de enfermos

Este epígrafe se dedica a solucionar un problema no bioinformático para mostrar las posibilidades de aplicación de los métodos propuestos en otras áreas del saber.

Se realiza un estudio sobre la mortalidad y morbilidad en el municipio de Cifuentes, Villa Clara. Se seleccionaron las enfermedades Cerebrovasculares, Corazón, Tumores malignos, Suicidios y Accidentes, que constituyen las cinco primeras causas de muerte en el territorio, se estudiaron además, Hepatitis A, Meningoencefalitis Viral e Intentos Suicidas por ser las enfermedades que incrementaron notablemente su incidencia en los últimos diez años (Díaz 2010).

Los Suicidios e Intentos Suicidas no son enfermedades como tal, se definen como “trastornos de la conducta” y están incluidas en las Enfermedades de Declaración

Obligatoria (EDO). En cualquier caso, resultó muy interesante para los médicos especialistas que participaron en esta investigación su inclusión en el estudio. En lo adelante se utilizará el término “enfermedades” de una forma general, para referirse también a ellos, sin que eso afecte la claridad del objetivo de este epígrafe.

Los datos utilizados fueron obtenidos de las bases de datos de mortalidad y morbilidad de la dirección Provincial de Salud en Villa Clara, correspondiente al municipio de Cifuentes. En el caso de la morbilidad se realizó un trabajo mucho más intenso pues estos datos no están informatizados, sólo se encuentran archivadas sus tarjetas de EDO.

En Higiene y Epidemiología existen sus propias técnicas para detectar epidemias, se utilizan métodos de detección de conglomerados cuando tienen dudas en algunos casos, es obvio que estos métodos pueden ser utilizados de forma general, aunque se llegan a las mismas conclusiones, por tal razón esta información fue procesada utilizando dos software de detección de conglomerados implementados con objetivos diferentes, ellos son:

- El “**Optimus**”, recibiendo como datos de entrada una cadena no binaria formada por la cantidad de pacientes con una enfermedad determinada en cada día del período analizado.
- El “**EpiDet**” (Casas 2003), recibiendo como datos de entrada las fecha de los pacientes de una enfermedad en el período analizado.

Con ambos softwares se obtienen los mismos resultados, pero con el Optimus se puede utilizar el Scan Borroso sobre una línea para identificar la posición en tiempo en que se encuentran los enfermos que favorecen a la formación de focos de enfermedades. Es esta la razón por la cual sólo se hará referencia a los resultados finales sin referirnos al software utilizado.

3.4.1. Metodología para la aplicación de los métodos Scan en la detección de conglomerados de enfermos

Como parte de este trabajo, se decidió formalizar un conjunto de pasos que sirven de guía a los epidemiólogos para la correcta aplicación de los métodos Scan en la detección de conglomerados de enfermos. A continuación se describen y se comentan

cada uno de ellos:

Paso 1:	Recopilación de datos (selección de las enfermedades, afecciones, trastornos de la conducta etc. a evaluar).
Paso 2:	Determinación de los valores de los parámetros del método Scan (Se recomienda que sean varios valores).
Paso 3:	Aplicar el método Scan Clásico. Si los resultados coinciden para todos los valores de los parámetros seleccionados, concluir.
Paso 4:	Si “hay dudas” (no coincidencia de los resultados para todos los valores de los parámetros seleccionados), entonces aplicar el método Scan Borroso. En base a los resultados que arroje este último método, concluir.

Para realizar el paso 1 debe consultarse las bases de datos de mortalidad y morbilidad existentes en los departamentos de estadísticas de salud en la forma ya explicada con anterioridad.

La diferencia fundamental entre los problemas anteriormente estudiados y este, es que existe un conjunto de médicos epidemiólogos expertos en el tema que pueden determinar los valores de los parámetros, ancho de la ventana móvil y el paso del desplazamiento. Estos valores no tienen que ser los mismos para todas las enfermedades estudiadas, pero dependen mucho de la forma en la que se recopila la información: semanal, quincenal, mensual, etc. Debido a la selección subjetiva de estos parámetros, pueden variar en dependencia de los criterios de los epidemiólogos (no siempre se ponen de acuerdo), se recomienda probar con varias configuraciones. Es importante mencionar que, para evitar sesgos en los resultados, los valores de los parámetros deben elegirse sin haber revisado previamente los datos.

Parámetros			
Ventana Móvil	Pasos		
60	30	15	7
30			
15			

Figura 3.2: Valores de los parámetros de Scan aplicado en cada una de las enfermedades.

Los especialistas en Higiene y Epidemiología de la Unidad de Higiene de Cifuentes proponen los siguientes valores para los parámetros, resumidos en la Figura 3.2.

El paso 3 se refiere concretamente a la aplicación de los métodos clásicos. Se recomienda aplicar las siguientes reglas:

- Si para todos los valores de los parámetros previamente seleccionados, los resultados son significativos, se concluye que existen conglomerados de enfermos. Terminar.
- Si para todos los valores de los parámetros previamente seleccionados, los resultados son no significativos, se concluye que no existen conglomerados de enfermos. Terminar.

Al paso 4 se llega si existen dudas, es decir si los resultados no coincidieron para todas las configuraciones de parámetros seleccionadas. En estos casos se debe aplicar el método Scan Borroso. Recuérdese que este método tiene un parámetro adicional: la longitud de la parte borrosa de la ventana móvil.

Al aplicar el método Scan Borroso los resultados pueden seguir discrepando unos con otros. En este paso es crucial realizar el análisis con los especialistas. Sólo una opinión conjunta de los resultados estadísticos unido a los criterios de epidemiólogos será definitiva (Díaz 2010).

A continuación se describen los resultados obtenidos de la aplicación del método Scan. En dependencia de las conclusiones que se extrajeron, se formaron los tres grupos siguientes:

- **Resultados no significativos:** resume la información de aquellas enfermedades en las que no se demostró la presencia de conglomerados.
- **Resultados significativos para todos los valores de los parámetros:** resume la información de aquellas enfermedades en las que se demostró la presencia de conglomerados para todos los valores de los parámetros considerados.
- **Resultados significativos para algunos valores de los parámetros:** resume la información de aquellas enfermedades en las que el método Scan arrojó dudas. Para algunos valores de los parámetros los resultados fueron

significativos mientras que para otros no. Por lo que se decidió aplicar además el método Scan Borroso para llegar a conclusiones más certeras.

3.4.2. Análisis y discusión de las enfermedades estudiadas en Cifuentes

Para las enfermedades Cerebrovasculares, Accidentes, Suicidios, Meningoencefalitis Viral y Hepatitis A, se obtienen idénticos resultados para cualquier juego de parámetros, motivo por el cual no se discutirán los resultados. Sólo se discutirán las enfermedades restantes porque en ellas es necesario utilizar el Scan Borroso para ayudar a las autoridades de salud a tomar una decisión. El método se aplicó utilizando como máximo un suavizado de siete días (una semana).

Enfermedades del Corazón

Las Enfermedades del Corazón son la primera causa de muerte en Cuba. Producidas por un desbalance entre la oferta y la demanda de oxígeno al miocardio, debido a lesiones orgánicas (aterosclerosis) o funcionales (espasmo) y que provocan varios cuadros, desde fenómenos asintomáticos (isquemia silente, disfunción diastólica) hasta cuadros de necrosis miocárdica extensa (Penichet et al. 2007).

La forma clínica más grave de estas enfermedades es el Infarto Agudo del Miocardio (IMA), esta entidad se caracteriza por un fuerte dolor precordial, que puede irradiarse a la axila, ambos brazos o el izquierdo, o al cuello, acompañado de sudoración profundas, vómitos y mareos. El dolor habitualmente dura más de 10 minutos y requiere, con frecuencia, el uso de opiáceos para su alivio (Toledo 2007).

Se observó que en el análisis general de todos los casos que se muestra en la Tabla 3.7, la técnica del Scan Clásico expresa que existen la presencia de conglomerados para todos los valores de los parámetros considerados, excepto para dos juegos de parámetro, corroborándose la presencia de conglomerados en los mismos utilizando el Scan Borroso con un suavizado no superior a dos días, es decir la disposición de los pacientes en el tiempo favorece a la formación de los conglomerados.

Tabla 3.7 Resultados obtenidos con los métodos Scan para las Enfermedades del Corazón.

Vent. M.	Paso	Scan sobre una línea											
		General				Factores (Sexo)							
		Clásico		Borroso		Masculino				Femenino			
		Est.	p.	S*	Res.	Clásico		S*	Res.	Clásico		S*	Res.
						Est.	p.			Est.	p.		
60	30	32	0.000	—	—	18	0.059	1	Sig	14	0.289	7	No S.
	15	32	0.000	—	—	18	0.060	1	Sig	14	0.294	7	No S.
	7	32	0.000	—	—	18	0.058	1	Sig	14	0.298	7	No S.
30	30	16	0.246	1	Sig	11	0.302	6	Sig	10	0.234	7	No S.
	15	20	0.004	—	—	13	0.029	—	—	10	0.231	3	Sig
	7	24	0.000	—	—	14	0.006	—	—	10	0.218	3	Sig
15	15	12	0.084	2	Sig	9	0.068	1	Sig	7	0.385	7	No S.
	7	14	0.005	—	—	9	0.063	4	Sig	8	0.087	7	No S.

Nota;

- S* grado de suavizado utilizado en el Scan Borroso.

La figura 3.3 muestra una representación gráfica de los datos procesados. Pueden apreciarse picos con una incidencia más elevada de la enfermedad alrededor de los años 1997 - 1998 y 2004 - 2005.

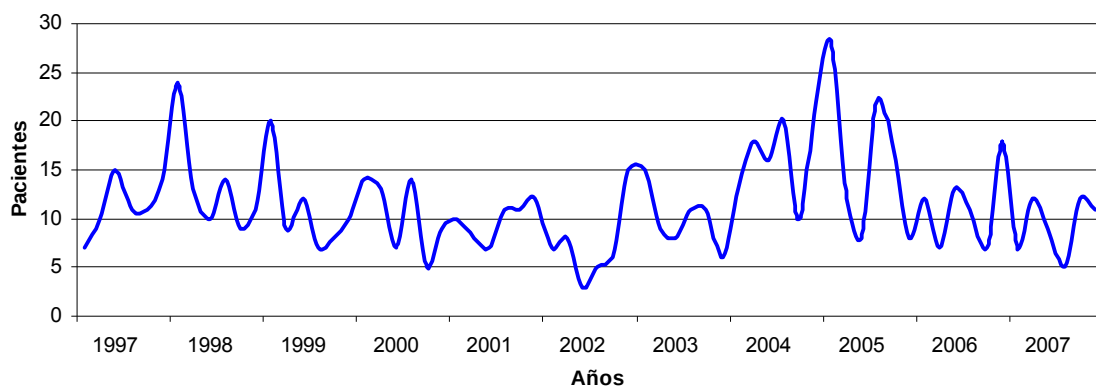


Figura 3.3. Distribución de la mortalidad por Enfermedades del Corazón en Cifuentes en el período 1997 – 2007.

En los años 1997 y 1998, se incrementó la mortalidad por Enfermedades del Corazón, según los especialistas en Higiene y Epidemiología del municipio pues coincide con la

etapa del período especial, donde se modificaron los estilos de vida de la población por la difícil situación económica que existió en el país durante esa fecha, se incrementó el consumo de grasa de origen animal, disminuyó la realización de ejercicios físicos, y aumentó el estrés, todo esto condujo a un aumento de la incidencia de hipertensión arterial, que constituyen los principales factores de riesgo de esta enfermedad.

El mayor número de fallecidos por enfermedades del corazón se produjo alrededor de los años 2004 y 2005, debemos tener en cuenta que la edad es uno de los principales factores de riesgo de estas patologías y la provincia de Villa Clara y en particular el municipio de Cifuentes presenta una de las poblaciones más envejecidas del país, el grupo de edad de 65 años y más representa el 21% de la población total de estos años. Además se incrementaron los hábitos tóxicos como el consumo de café, tabaco y alcohol fundamentalmente en la población masculina, existe un mal seguimiento en consulta de la hipertensión arterial y hay una tendencia al abandono del tratamiento por parte de los pacientes, todo esto pudo contribuir al incremento de la mortalidad por esta causa.

En la tabla 3.7 se hace también el análisis separado para ambos sexos. Se sigue la misma metodología: en los casos en los que el Scan Clásico no brinda resultados satisfactorios, se aplica el método Scan Borroso, concluyendo que existe un foco de mortalidad masculina para todos los juegos de parámetros, no ocurriendo lo mismo para el sexo femenino para todos los juegos de parámetros.

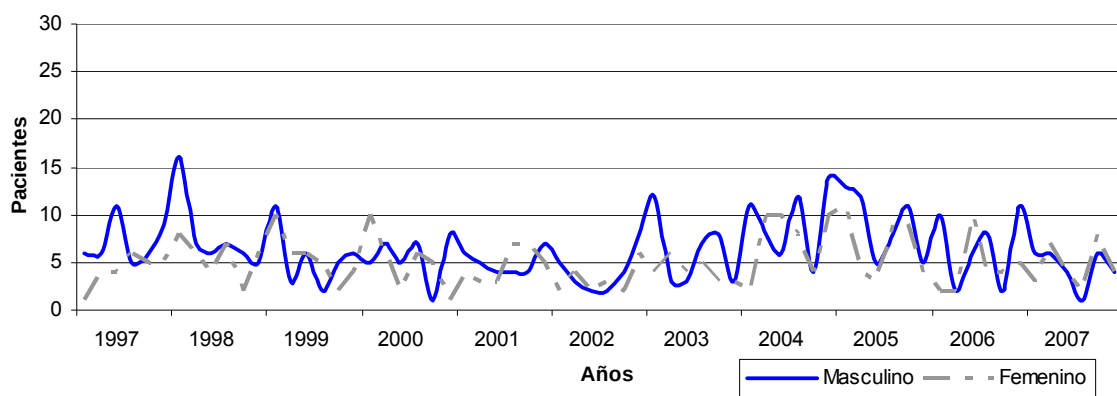


Figura 3.4 Distribución de la mortalidad por Enfermedades del Corazón, según sexo, en Cifuentes en el período 1997 – 2007.

De manera descriptiva puede apreciarse que esta enfermedad atacó más al sexo masculino que al femenino en el municipio de Cifuentes. Existiendo focos de

mortalidad masculina alrededor de los años 1997 al 1998 y del 2004 al 2005, lo que se puede apreciar en el gráfico de series de tiempo de las incidencias por sexo que aparece en la figura 3.4.

Tumores Malignos

El Cáncer es una proliferación celular desordenada debido a la pérdida de los controles normales, que da lugar a crecimiento desordenado, ausencia de diferenciación, invasión tisular local y, con frecuencia, metástasis. El Cáncer puede aparecer en cualquier tejido u órgano, a cualquier edad. Con frecuencia hay una respuesta inmunitaria frente a los tumores. Las neoplasias malignas pueden producir dolor, adelgazamiento, neuropatías, náuseas, anorexia, convulsiones, hipercalcemia, hiperuricemia y obstrucción. La muerte se produce típicamente como consecuencia de la insuficiencia súbita o progresiva de uno o más sistemas de órganos (Beers et al. 2007).

Se estima que el cáncer provoca la pérdida de más de cuatro millones de personas anualmente, lo que lo convierte en la tercera causa de muerte a escala mundial. Constituye un problema de salud especialmente relevante en los países desarrollados, en los cuales se ha logrado controlar otras causas de muerte, pero esto se ha convertido en un verdadero flagelo. En Europa, uno de cada cuatro ciudadanos muere por esta causa; en España se ha convertido en la segunda causa de muerte, y se conoce que un elevado porcentaje de tales pacientes (50-90%) padece dolor. En Cuba, constituye la segunda causa de muerte desde 1958. Se estima que al iniciarse el siglo XXI haya superado a la cardiopatía (Lovellette et al. 2007).

La Tabla 3.8 muestra los resultados de la aplicación de los métodos Scan para la detección de conglomerados a los casos de mortalidad por Tumores Malignos en el período comprendido entre los años 1997 y 2007. Al igual que en las Enfermedades del Corazón en la población general del municipio existen conglomerados para todos los juegos de parámetros al utilizar el Scan Clásico y en sus excepciones el método del Scan Borroso lo corrobora con un suavizado de 7 o menos días, en la Figura 3.5, se observa evidentemente un foco de mortalidad por cáncer alrededor de los años 2002, que son precisamente los picos que están detectando los métodos aplicados.

Tabla 3.8. Resultados obtenidos con los métodos Scan para la mortalidad por Tumores Malignos.

Vent. M.	Paso	Scan sobre una línea											
		General				Factores (Sexo)							
		Clásico		Borroso		Masculino				Femenino			
		Est.	p.	S*	Res.	Clásico		S*	Res.	Clásico		S*	Res.
						Est.	p.			Est.	p.		
60	30	27	0.001	—	—	16	0.093	4	Sig	13	0.062	1	Sig
	15	28	0.000	—	—	18	0.011	—	—	13	0.063	5	Sig
	7	28	0.000	—	—	18	0.012	—	—	13	0.062	1	Sig
30	30	15	0.092	4	Sig	9	0.748	7	No S.	8	0.365	7	No S.
	15	15	0.092	2	Sig	12	0.032	—	—	8	0.368	7	No S.
	7	16	0.032	—	—	12	0.034	—	—	8	0.367	7	No S.
15	15	9	0.594	7	Sig	7	0.515	7	No S.	5	0.919	7	No S.
	7	9	0.632	6	Sig	7	0.553	7	Sig	5	0.921	7	No S.

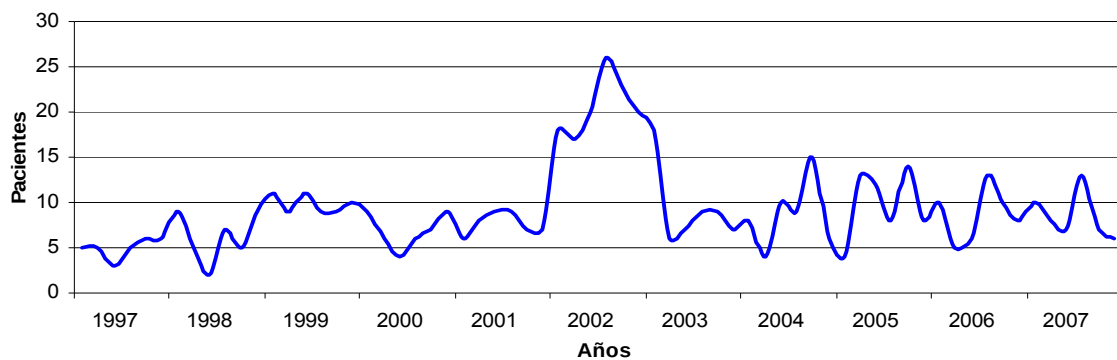


Figura 3.5. Distribución de la mortalidad por Tumores Malignos en Cifuentes en el período 1997 – 2007.

Según el criterio de especialistas del municipio, la mortalidad por Tumores Malignos aumentó durante los años, 2002, 2004 y 2005. Esta es una enfermedad de etiología desconocida, cuya aparición se asocia a factores de riesgo. Según estudios realizados (Beers et al. 2007) los factores ambientales constituyen un riesgo a largo plazo en la aparición de estas enfermedades, coincidiendo con el uso de productos químicos para la maduración de las frutas y el abuso de insecticidas en la agricultura durante el

período especial, además se incrementó el consumo de café y alcohol, sobre todo de bebidas de fabricación casera con alto grado de sustancias tóxicas, y en estos años se reporta que el 23.6% de la población del municipio es fumadora.

Al analizar la mortalidad del cáncer por sexo se observa en la Tabla 3.8. y en la Figura 3.6 que hay una tendencia a existir un foco de mortalidad en los masculinos alrededor de los años 2002 y principios del 2003, no existiendo evidencias marcadas de conglomerados en el sexo femenino.

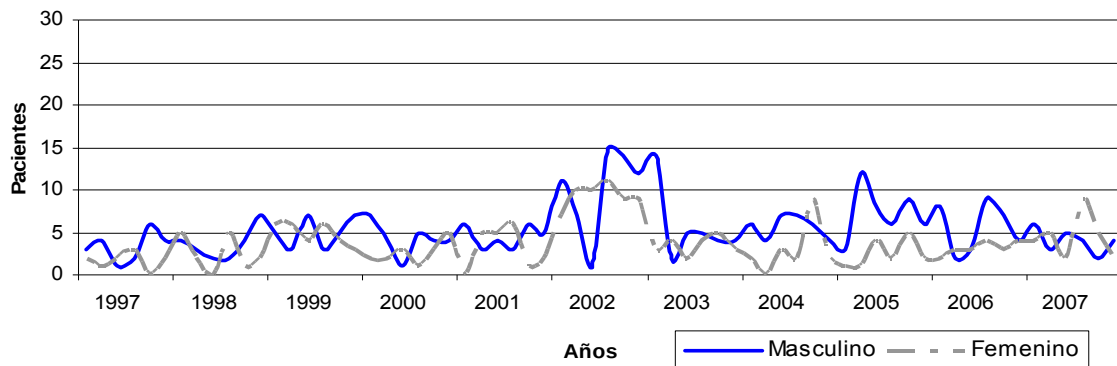


Figura 3.6. Distribución de la mortalidad por Tumores Malignos, según sexo, en Cifuentes en el período 1997 – 2007.

Intentos Suicidas

El Intento Suicida o parasuicidio es definido por la OMS, "como un acto con una consecuencia no fatal en la cual el individuo realiza deliberadamente una conducta no habitual con amenaza de muerte, que sin la intervención de otros le causará autodaño, o ingiere una sustancia superior a las dosis terapéuticas generalmente reconocidas y cuyo objetivo es producir cambios que él o ella desean a través de las consecuencias físicas y psíquicas reales o esperadas cercanas a la muerte" (Guibert y Torres 2001).

Lo intentan más los adolescentes, en especial el sexo femenino, mediante el uso de los métodos más suaves como la ingestión de tabletas, aunque esto está relacionado con los medios disponibles a su alcance en el momento de la crisis (Rodríguez 2006).

Los factores psicosociales de riesgo individuales que, de acuerdo con las investigaciones científicas más actuales sobre los intentos suicidas son: presencia generalizada de sentimientos de desesperanza y culpa, presencia de depresión mayor, personas que han sobrevivido al intento suicida, personas que han llamado la atención

por presagiar o amenazar con el suicidio (proyecto suicida), antecedentes familiares de suicidio o de intento suicida, personas sin apoyo social y familiar y presencia de impulsividad o de ansiedad y hostilidad (Guibert 2003).

Tabla 3.9. Resultados obtenidos con los métodos Scan para la morbilidad por Intentos Suicidas.

Vent. M.	Paso	Scan sobre una línea											
		General				Factores (Sexo)							
		Clásico		Borroso		Masculino				Femenino			
		Est.	p.	S*	Res.	Clásico		Borroso		Clásico		Borroso	
						Est.	p.	S*	Res.	Est.	p.	S*	Res.
60	30	14	0.346	7	No S.	5	0.854	7	No S.	11	0.600	7	No S.
	15	14	0.343	7	No S.	5	0.854	7	No S.	11	0.593	7	No S.
	7	14	0.361	7	No S.	5	0.856	7	No S.	12	0.295	7	No S.
30	30	10	0.263	7	No S.	4	0.806	7	No S.	8	0.502	7	No S.
	15	10	0.259	7	No S.	4	0.800	7	No S.	8	0.495	7	No S.
	7	10	0.273	7	No S.	4	0.830	7	No S.	8	0.515	7	No S.
15	15	8	0.098	5	Sig	2	1.000	7	No S.	6	0.497	7	No S.
	7	8	0.097	1	Sig	3	0.965	7	No S.	6	0.488	4	Sig

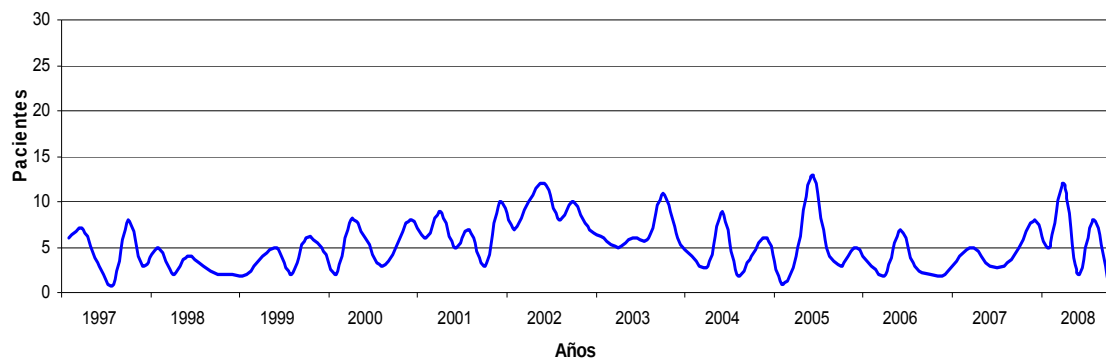


Figura 3.7. Distribución de la morbilidad por Intentos Suicidas en Cifuentes en el período 1997 – 2008.

La mayoría de los resultados que aparecen en la Tabla 3.9 muestran la no existencia de diferencias significativas, o lo que es lo mismo la no presencia de conglomerados de Intentos Suicidas en el municipio de Cifuentes en el período de 1997 a 2008 con el Scan Clásico, cuando se corrobora con el Scan Borroso

se obtiene resultados significativos con la ventana móvil de tamaño 15 y el paso del desplazamiento de 15 y 7 días con un suavizado de 5 y 1 respectivamente.

En la Figura 3.7 se muestra la serie de tiempo de los enfermos con unidad de medida dos meses (60 días), no observándose evidencias de picos en las mismas. Sin embargo, al volver a graficar los datos de la incidencia de intentos suicidas, mostrando las cantidades de casos reportados cada 15 días. La Figura 3.8 muestra que alrededor de los años 2002 y 2008 existen efectivamente dos picos notables. Esos son los que detecta el método borroso.

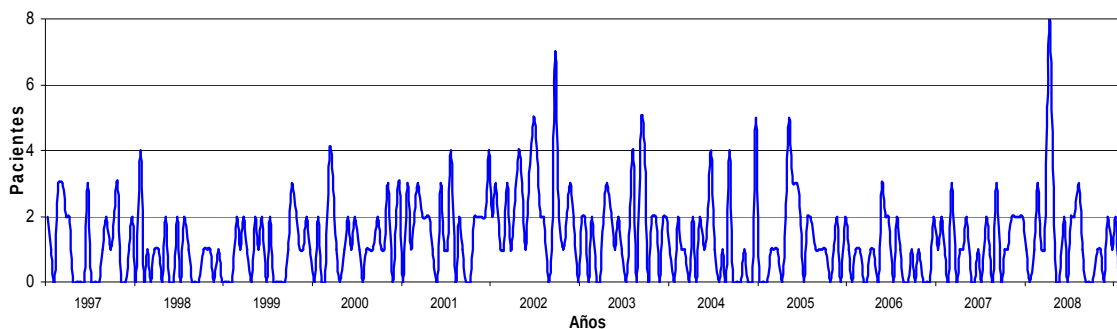


Figura 3.8. Distribución de la morbilidad quincenal por Intentos Suicidas en Cifuentes en el período 1997 – 2008.

De manera general los picos no son tan elevados. Ello unido a los resultados no significativos del método Scan para las otras configuraciones de los parámetros, y a los criterios de los epidemiólogos, permitieron concluir que no existían clusters de enfermos en el período analizado.

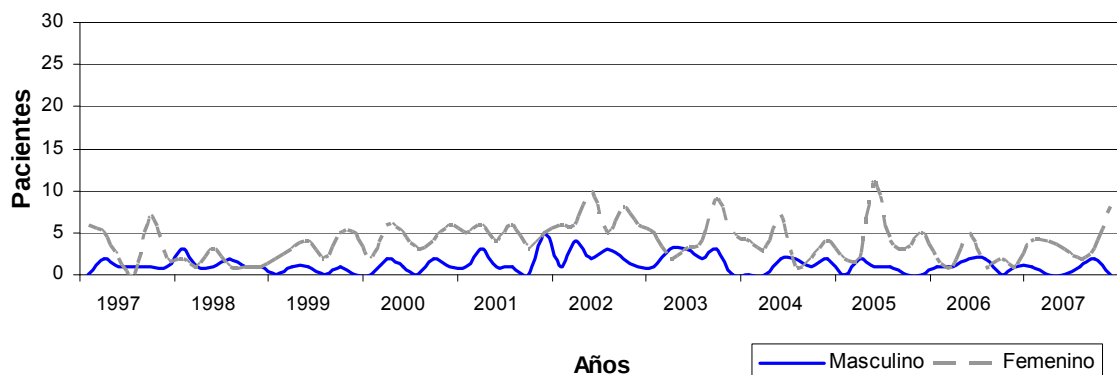


Figura 3.9. Distribución de la morbilidad por Intentos Suicidas, según sexo, en Cifuentes en el período 1997 – 2008.

Al analizarse el comportamiento por sexo de los Intentos Suicidas se aprecia en la Tabla 3.9, ambos métodos no detectan conglomerados para ninguna combinación de parámetros en el sexo, excepto para el sexo femenino que para ventana móvil igual a 15, paso 7 y un suavizado de 4 días el Scan Borroso detecta que hay conglomerado, al observarse la Figura 3.9 hay un pico no prolongado alrededor de los años 2002 y 2005.

Los Intentos Suicidas han tenido un comportamiento habitual en los años analizados, un ligero incremento de estos casos estuvo relacionado con los síndromes depresivos en el anciano que vive solo, siendo más evidente en el sexo femenino pues está demostrado por estudios realizados que en las mujeres son más frecuentes los intentos suicidas y en los hombres el suicidio.

3.4.3. Consideraciones sobre la detección de conglomerados de enfermos

En estos problemas en particular la secuencia utilizada es no binaria y cuando se usa el Scan Borroso el resultado depende del ancho de la ventana, paso y suavizado. El resultado puede ser afectado de forma positiva cuando en los extremos de la ventana móvil hay una cantidad considerable de enfermos, ya que su peso es mucho mayor y puede afectar considerablemente al estadígrafo obteniendo como resultado que los datos contribuyen a la formación de conglomerados.

Además los tamaños de las ventanas móviles que determinan los Epidemiólogos son relativamente muy pequeños, menores al 2% del tamaño total de la secuencia, lo que implica en general según la teoría, mejores resultados del Scan Borroso.

3.5 Consideraciones finales del capítulo

En este capítulo se describen brevemente los productos de softwares utilizados en esta investigación y se muestran varias aplicaciones de las contribuciones propuestas en dominios bioinformáticos y biomédicos.

Se presentó un estudio sobre los orígenes de replicación en secuencias de herpesvirus y bacterias, otra sobre la existencia de conglomerados de gaps en secuencias de H1N1 y finalmente se presentó una metodología para el uso de los métodos propuestos en investigaciones de Epidemiología.

CONCLUSIONES Y RECOMENDACIONES

Al finalizar este trabajo se arriba a las siguientes conclusiones:

1. Se crearon e implementaron los métodos Scan Borrosos para la detección de conglomerados en secuencias, a partir de la combinación de sus variantes clásicas con elementos de la lógica borrosa. Estas técnicas tienen eficiencia similar o superior a las ya reportadas en la literatura.
2. Se implementaron los métodos propuestos computacionalmente en plataformas de software libre, utilizando Java como lenguaje de programación. Además se desarrollaron otras implementaciones en el lenguaje basado en listas que soporta el paquete *Mathematica*.
3. Se realizó un estudio de simulación en secuencias relativamente pequeñas para analizar la influencia de los valores de los parámetros en la capacidad de respuesta de los métodos. Se concluyó que no deben utilizarse valores demasiado pequeños (cerca de uno) y valores demasiado grandes (valores cercanos al tamaño de la secuencia analizada).
4. Se aplicó el análisis bifactorial no paramétrico para analizar de forma general el comportamiento de los parámetros de los métodos en secuencias grandes.
5. Se utilizó un algoritmo bioinspirado con el objetivo de optimizar los métodos Scan, aplicados fundamentalmente en secuencias largas para encontrar un juego de parámetros que favorecen, si existe, a la formación de conglomerados.
6. Se ejemplificó el uso de los métodos desarrollados en problemas de análisis de secuencias genómicas en bioinformática, así como en problemas médicos de detección de epidemias. En todos los casos se obtuvieron buenos resultados.

Los resultados obtenidos de ninguna forma agotan el desarrollo de esta temática. Al igual que los resultados de cualquier desarrollo teórico, constituyen las bases para nuevas líneas de investigación. A continuación se enumeran algunos temas que pudieran ser fuentes de trabajos futuros a manera de recomendaciones:

1. Realizar un análisis de los algoritmos propuestos para determinar si es posible obtener versiones paralelizadas. Ello aumentaría notablemente las posibilidades de aplicación en dominios Bioinformáticos.
2. Analizar la posible aplicación de otros algoritmos bioinspirados, como el de colonia de hormigas o bandadas de insectos en sustitución del algoritmo PSO utilizado.
3. Investigar la posibilidad de utilizar funciones de pertenencia aplicadas a la categoría de interés, para intentar solucionar problemas como los relacionados con la detección de cajas TATA.

REFERENCIAS BIBLIOGRÁFICAS

- Aldrich, T. y Wanzer, D. (1993). "'Cluster', The agency for Toxic Substances and Disease Registry Division of Health Studies."
- Anderson, C. (2008). "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete " Wired **16**(7). www.wired.com/science/discoveries/magazine/16-07/pb_theory.
- Bailey, N. T. J. (1975). "The mathematical theory of infectious diseases and it's applications." Charles Griffin & Company Limited, Second Edition.
- Baldi, P. y Brunak, S. (2001). Bioinformatics.. the Machine Learning Approach. Cambridge, England, The MIT Press.
- Baldi, P. y Pollastri, G. (2003). "The principled design of large-scale recursive neural network architectures--dag-rnns and the protein structure prediction problem." The Journal of Machine Learning Research **4**: 575-602.
- Barbour, A. D., Holst, L. y Janson, S. (1992). Poisson Approximation, Clarendon Press, Oxford.
- Beers, H., Porter, R. y Jones, T. (2007). "Hematología y oncología." El manual Merck. E. española **1119**.
- Beielstein, T., Parsopoulos, K. E. y Vrahatis, M. N. (2002). Tuning PSO parameters through sensitivity analysis. , Technical Report of the Collaborative Research Center, University of Dortmund: <http://sfhci.cs.uni-dortmund.de/home/English/Publi>.
- Bell, G., Hey, T. y Szalay, A. (2009). "Computer science. Beyond the data deluge." Science **323**(5919): 1297-1298.
- Benson, D. A., Karsch-Mizrachi, I., Ostell, O. y Wheeler, D. L. (2005). "GenBank." Nucleic Acids Research **33**.
- Bird, A. (1987). "CpG islands as gene markers in the vertebrate nucleus." Trends in Genetics **3**: 342–347.
- Bonet, I., Grau, R., Rodríguez, A. y García, M. M. (2007). Predicción de splice sites usando redes neuronales recurrentes. XII Convención y Expo Internacional de Informática, INFORMÁTICA 2007, La Habana.,
- Bonet, I., Rodríguez, A., Grau, R. y García, M. M. (2008). Combining classifiers for Bioinformatics. Second International Workshop on Bioinformatics, Cuba- Flanders, 2008, Villa Clara,
- Boutros, P. (2006). "Why biologist can't count?: An overview of the gene-finding problem." Hypoth: 26-29.
- Brender, J., Talmon, J., Egmont-Petersen, M. y McNair, P. (1994). Measuring quality of

- medical knowledge. Medical Informatics in Europe, Lisbon,
- Brubaker, D. y Cedric, S. (1992). "Fuzzy-logic system solves control problem." EDN **18**: 121-127.
- Brudno, M., Malde, S., Poliakov, A., Do, C. B., Couronne, O., Dubchak, I. y Batzoglou, S. (2003). "Glocal alignment: finding rearrangements during alignment." Bioinformatics **19**(1): 54-62.
- Buckley, J. y Jowers, L. (2007). Monte Carlo Methods in Fuzzy Optimization. 978-3-540-76289-8, Heidelberg.
- Calviño, M. H. (2003). "Aclarando la Lógica borrosa (Fuzzy Logic)." Revista Cubana de Física **20**(2): 5.
- Cardellá, L. y Hernández, R. (1999). Bioquímica Médica. Tomo II. La Habana Ciencias Médicas
- Casas, G. (2003). Técnicas de detección de conglomerados incluyendo factores adicionales. Departamento de Computación. Santa Clara Universidad Central "Marta Abreu". **Tesis presentada en opción al grado científico de Doctor en Ciencias Técnicas: 113.**
- Casas, G., Grau, R. y 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.
- Consortium, I. H. G. (2004). "Finishing the euchromatic sequence of the human genome. International Human Genome Sequencing Consortium." Nature **431**(7011): 931-45.
- Cox, R. T. (1946). "Probability, Frequency and Reasonable Expectation." American Journal of Physics **14**(1): 1-13.
- Cromie, G., Millar, C., Schmidt, K. y Leach, D. (2000). "Palindromes as substrates for multiple pathways of recombination in Escherichia coli." Genetics **154**(2): 513-522.
- Chávez, M., Silveira, P., Casas, G. y Grau, R. (2007a). Aprendizaje estructural de redes bayesianas utilizando PSO. COMPUMAT, Holguín, Cuba 5., Holguín,
- Chávez, M., Casas, G., Moreira, J., González, E., Bello, R. y Grau, R. (2008a). "Uso de redes bayesianas obtenidas mediante Optimización de Enjambre de Partículas para el diagnóstico de la Hipertensión Arterial. ." Revista Investigación Operacional **30**(1). 52-59.
- Chávez, M. C., Casas, G. y Grau, R. (2007b). "Uso de las redes bayesianas combinado con técnicas estadísticas para el diagnostico de la Hipertensión arterial." Revista Automática Comunicaciones y Electrónica **XXXVIII**(2): 45- 48.
- Chávez, M. C., Casas, G., Moreira, J., Silveira, P., Moya, I., Bello, R. y Grau, R.

- (2008b). "Predicción de mutaciones en secuencias de la proteína transcriptasa inversa del VIH usando nuevos métodos para Aprendizaje Estructural de Redes Bayesianas." Avances en Sistemas e Informática. 4(2): 77-85.
- Cheng, J. y Baldi, P. (2005). "Three-stage prediction of protein beta-sheets by neural networks, alignments and graph algorithms." Bioinformatics 21: 75-84.
- Cheng, J., Arlo, R. y Baldi, P. (2006). "Baldi P: Prediction of protein stability changes for single-site mutations using support vector machines." Proteins 62(1125--1132).
- Daalen, V. C. (1992). Evaluating Medical Knowledge Based Systems. Annual International Conference of the IEEE Engineering in Medicine and Biology Society. 3: 888-889.
- Davis, J. y Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. In ICML '06: Proceedings of the 23rd international conference on Machine learning, Pittsburgh, Pennsylvania,
- Davis, L. (1991). "Handbook of Genetics Algorithms." Van Nostrand Reinhold Company, New York II: 100 páginas
- Delecluse, H. J. y Hammerschmidt, W. J. (2000). "The genetic approach to the Epstein-Barr virus: From basic virology to gene therapy." Mol. Pathol 53(5): 270–279.
- Delvin, T. (2006). Bioquímica. Libro de Texto con aplicaciones clínicas. Barcelona, España, Editorial Reverté, S.A.
- Dembo, A. y Karlin, S. (1992). "Poisson approximations for r-scan processes." Ann. Appl. Probab. 2(2): 329–357.
- Díaz, F., Rodríguez, L., Casas, G. y Grau, R. (2009). Análisis de los parámetros del Scan Lineal utilizando diseño de experimento. Memorias del Primer Taller Internacional FIMAT XXI ISBN. Holguín.
- Díaz, J. L. (2010). Uso de los métodos Scan en la detección de conglomerados de enfermos en Cifuentes. Centro de Estudios Informáticos. Santa Clara. Villa Clara, Universidad Central "Marta Abreu" de Las Villas. **Tesis presentada en opción al grado académico de Máster en Computación Aplicada.**
- Donald, M., Spiegelhalter, C. y Taylor, J. (1994). Machine Learning, Neural and Statistical Classification Reviews.
- Dopazo, J. y Valencia, A. (2002). "Bioinformática y Genómica." Genómica y mejora vegetal 147-198
- Durbin, R., Eddy, S., Krogh, A. y Mitchison, G. (2003). Biological sequence analysis. Australia, The press syndicate of the University of Cambridge.
- EBI. (1999). "The European Bioinformatics Institute." from <http://www.ebi.ac.uk>.
- Ezura, Y., Sekiya, I., Koga, H., Muneta, T. y Noda, M. (2009). "Methylation status of

- CpG islands in the promoter regions of signature genes during chondrogenesis of human synovium-derived mesenchymal stem cells." InterScience **60**(5): 1416 – 1426.
- Fawcett, T. (2004). "ROC Graphs: Notes and Practical Considerations for Researchers." Machine Learning: <http://citeseer.ist.psu.edu/fawcett04roc.html>.
- Feller, W. (1971). An Introduction to Probability Theory and Its Applications. Reino Unido (INGLATERRA), JOHN WILEY & SONS,.
- Fernández, H. M. (2006). "SIG-ESAC: Sistema de Información Geográfica para la gestión de la estadística de salud de Cuba." Rev Cubana Hig Epidemiol **44**(3).
- Galperin, M. (2007). "The Molecular Biology Database Collection 2007 update. ." Nucleic Acids Research, **35**: D3 - D4.
- Giegerich, R. (2000). "A systematic approach to dynamic programming in bioinformatics." Bioinformatics **16**(8): 665-677.
- Glaz, J. (1989). "Approximations and bounds for the distribution of the scan statistics." Statist. Assoc. **84**(406): 560–566.
- Glaz, J. (1993). "Approximations for the tail probabilities and moments of the Scan statistics." Statistics in medicine **12**: 1845-1852.
- Glaz, J. y Balakrishnan, N. (1999). Scan Statistics and Applications. Boston, Hardcover.
- Glaz, J., Naus, J., Roos, M. y Wallenstein, S. (1994). "Poisson approximations for the distribution and moments of ordered m-spacings." Appl. Prob. **31**(A): 271-281.
- Grau, R. y Sánchez, R. (2009). Nuevos modelos algebraicos y markovianos del proceso evolutivo. Aplicaciones a la predicción de mutaciones de la influenza. Memorias del XI Congreso de Matemática y Computación, Compumat 2009. La Habana.
- Guibert, W. (2003). "Factores psicosociales de riesgo de la conducta suicida." Revista Cubana Medicina General Integral **5**(19).
- Guibert, W. y Torres, N. (2001). "Intento suicida y funcionamiento familiar." Rev Cubana Med Gen Integr **17**(5): 452-60.
- Halevy, A., Norvig, P. y Pereira, F. (2009). "The Unreasonable Effectiveness of Data." IEEE Intell. Syst. **24**(2): 8-12.
- Hamzeh, F. M., Lietman, P. S., Gibson, W. y Hayward, G. S. (1990). "Identification of the lytic origin of DNA replication in human cytomegalovirus by a novel approach utilizing ganciclovir-induced chain termination." J. Virol **64**: 6184–6195.
- Hénaut, A., Rouxel, T., Gleizes, A., Moszer, I. y Danchin, A. (1996). "Uneven Distribution of GATC Motifs in the Escherichia coli Chromosome, its Plasmids and its Phages." Molecular Biology **257**: 574–585.

- Hinkelmann, K. y kempthorne, O. (2005). Advanced Experimental Design. New Jersey, John Wiley & Sons.
- Hinkelmann, K. y kempthorne, O. (2008). Introduction to Experimental Design New Jersey, John Wiley & Sons.
- Iliende, R., Curotto, L. B., Valiente, G. A., Toro, J., Santa María, L. y González, R. M. (2007). "Diagnóstico citogenético-molecular del síndrome Xq frágil." Rev. chil. tecnol. méd **27**(1): 1339-1346.
- Irizarry, R., Ladd-Acosta, C., Wen, B., Wu, Z., Montano, C., Onyango, P., Cui, H., Gabo, K., Rongione, M., Webster, M., Ji, H., Potash, J., Sabunciyan, S. y Feinberg, A. (2008). "Genome-wide methylation analysis of human colon cancer reveals similar hypoand hypermethylation at conserved tissue-specific CpG island shores." Nature Genetics **Available online**.
- Jacquez, G. y Waller, L. (1996). "Disease cluster statistics for imprecise space-time locations." Statistics in Medicine **15**: 873-85.
- Jacquez, G., Waller, L., Grimson, R. y Watenberg, D. (1996a). "The analysis of Disease Clusters, Part I: Stat of the Art." Infection Control and Hospital Epid. **17** (5): 319-27.
- Jacquez, G., Waller, L., Grimson, R. y Watenberg, D. (1996b). "The analysis of Disease Clusters, Part II: Introduction to techniques." Infection Control and Hospital Epid.. **17** (6): 385-97.
- Jain, A. K., Murty, M. N. y Flynn, P. J. (1999). "Data Clustering: A Review." ACM Computing Surveys **31**(3): 264-323.
- Janssens, D., Wets, G., Brijs, T. y Vanhoof, K. (2005). "The development of an adapted Markov chain modelling heuristic and simulation framework in the context of transportation research." Expert Syst. Appl. **28**(1): 105-117.
- Jaronski, W., Vanhoof, K. y Bloemer, J. (2005). "Inductive Development of Customer e-Loyalty Theory with Bayesian Networks." CORES 187-194.
- Karlin, S. y Brendel, V. (1992). "Chance and Statistical Significance in Protein and DNA Sequence Analysis." Science **39-49**, **257**, No. **5066**. : 39-49.
- Kennedy, J. (1997). The particle swarm: social adaptation of knowledge. IEEE International Conference on Evolutionary Computation,
- Kennedy, J. y Eberhart, R. (1995a). A new optimizer using particle swarm theory. Sixth International Symposium on Micro Machine and Human Science, Nagoya;
- Kennedy, J. y Eberhart, R. (1995b). Particle swarm optimization. IEEE International Conference on Neural Networks,, Perth, ,
- Kennedy, J., Spears, W. y 43, P. o. t.-. (1998). Matching algorithms to problems: an experimental test of the particle swarm and some genetic algorithms on the

- multimodal problem generator. IEEE International Conference on Evolutionary Computation, 39- 43,
- Kennedy, J., Eberhart, R. y Shi, Y. (2001). Swarm Intelligence. Morgan Kaufmann Series in Artificial Intelligence, Hardcover,
- Knox, E. (1964). "The detection of space-time interactions." Applied Statistics **13**: 25-9.
- Kron, K., Pethe, V., Briollais, L., Sadikovic, B., Ozcelik, H., Sunderji, A., Venkateswaran, V., Pinthus, J., Fleshner, N., Kwast, T. y Bapat, B. (2009). "Discovery of novel hypermethylated genes in prostate cancer using genomic CpG island microarrays." PLoS ONE **4**(3).
- Kulldorff, M. (1997). "A spatial scan statistic. Communications in Statistics." Theory and Methods **26**: 1481–1496.
- Kulldorff, M. (1998). "Evaluating cluster alarms: A space-time scan statistic and brain cancer in Los Alamos." American Journal of Public Health **88**: 1377-80.
- Kulldorff, M. (1999). "Geographic information systems (GIS) and community health: Some statistical issues," Journal of Public Health Management and Practice **5** 100-106.
- Kulldorff, M. (2001). "Prospective time-periodic geographical disease surveillance using a scan statistic." Journal of the Royal Statistical Society **164**: 61-72.
- Kulldorff, M., Mostashari, F., Duczmal, L., Yih, K., Kleinman, K. y Platt, R. (2007). "Multivariate scan statistics for disease surveillance." Statistics in Medicine **26**(8): 1824-1833.
- Lambert, C., Campenhout, J., DeBolle, X. y Depiereux, E. (2003). "Review of common sequence alignment methods: clues to enhance reliability." Current Genomics **4**: 131-146.
- Langrand, C. (2005). Scan Statistics: definición y ejemplos. Seminario ANY 2005, Universidad Politécnica de Catalunya. España., Université Sciences et Technologies de Lille (Lille-1),
- Larrañaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., Lozano, J. A., Armañanzas, R., Santafé, G., Pérez, A. y Robles, V. (2005). "Machine learning in bioinformatics." Briefings in Bioinformatics **7**(1): 86-112.
- Leach, D. (2005). "Long DNA palindromes, cruciform structures, genetic instability and secondary structure repair " BioEssays **18**(12): 893-900.
- Leung, M., Pui Choi, K., Xia, A. y Chen, L. (2005). "Nonrandom Clusters of Palindromes in Herpesvirus Genomes." Journal of Computational Biology **12**(3): 331–354.
- Lovelle, J., Cordero, N., Álvarez, A., Gutiérrez, J., Méndez, M. y Rodríguez, I. (2007). "Comportamiento de la mortalidad por tumores malignos." Revista Medcentro

- 11(2).**
- Lu, L., Jia, H., Drög, P. y Li, J. (2007). "The human genome-wide distribution of DNA palindromes " SpringerLink **7**(3): 221-227.
- Lukasiewicz, J. (1910). "O zasadzie wylaczonego srodka." Przegląd Filozoficzny **13**: 372-373.
- Mahamed, G. H. O., Engelbrecht, A. P. y Salman , A. (2005). Dynamic Clustering using PSO with Application in Unsupervised Image Classification. . In proceedings of the World Academy of Science, Engineering and Technology,
- Marrero-Ponce, Y., Meneses-Marcel, A., Castillo-Garit, J. A., Machado-Tugores, Y., Escario, J. A., B:A., G., Montero, D., Nogal-Ruiz, J. J., Arán, V. J., Martínez-Fernández, A. R., Torrens, F., Rotondo, R., Ibarra-Velarde, F. y Alvarado Ysaías, J. (2006). "Predicting antitrichomonal activity: a computational screening using atom-based bilinear indices and experimental proofs." Bioorganic & medicinal chemistry **14**(19): 6502-24.
- Martin, A. W. (1981). "A Generalised Scan Statistic Test for the Detection of Clusters." International Journal of Epidemiology. **10**.(3): 289-293.
- Martín del Brío, B. y Sánchez, A. (2005). Redes Neuronales y Sistemas Difusos. México, Alfaomega.
- Martínez-Piedra, R., Loyola-Elizondo, E., Vidaurre-Arenas, M. y Nájera-Aguilar, P. (2004). "Paquetes de Programas de Mapeo y Análisis Espacial en Epidemiología y Salud Pública." Boletín Epidemiológico OPS **25**(4): 1-9.
- Masse, M. J., Karlin, S., Schachtel, G. A. y Mocarski, E. S. (1992). "Human cytomegalovirus origin of DNA replication (oriLyt) resides within a highly complex repetitive region." Proc. Natl. Acad. Sci. USA. **89**(5246–5250.).
- Montgomery, D. C. (2008). Diseño y Análisis de Experimentos. México, Limusa.
- Mott, M. L. y Berger, J. M. (2007). "DNA replication initiation: mechanisms and regulation in bacteria." Nat. Rev. Microbiol. **5**(5): 343–54.
- Nagarwilla, N. (1996). "A Scan statistic with a variable window." Stat. in Med. **15**: 845-50.
- Naus, J. I. (1965). "The distrution of the size of the maximum cluster of points on a line." Journal of the American Statistical Association **60**: 532-538.
- Naus, J. I. (1982). "Approximations for distributions of Scan statistics." Journal of the American Statistical Association **77**(No. 377): 177-183.
- Neiman, P., Elsaesser, K., Loring, G. y Kimmel, R. (2008). "Myc Oncogene-Induced Genomic Instability: DNA Palindromes in Bursal Lymphomagenesis." PLoS Genet **4**(7).

- Newlon, C. S. y Theis, J. F. (2002). "DNA replication joins the revolution: Whole-genome views of DNA replication in budding yeast." BioEssays **24**: 300–304.
- Notredame, C. (2002). "Recent progresses in multiple sequence alignment: a survey." Pharmacogenomics **3**(1): 1-14.
- Peeters, M., Könönen, V., Verbeeck, K. y Nowé, A. (2008). "A Learning Automata Approach to Multi-agent Policy Gradient Learning." KES **2**: 379-390.
- Penichet, M., Pérez, R. y Triolet, A. (2007). Cardiopatía isquémica. Medicina Interna. Diagnóstico y tratamiento.
- Pérez, M., Morales, A., Molina, R. y García, J. (2006). "2D Autocorrelation Modelling of the Inhibitory Activity of Cytokinin-Derived Cyclin-Dependent Kinase Inhibitors." Bulletin of Mathematical Biology **68**(4): 735-751.
- Pertusa, J. F. (2003). Técnicas de análisis de imagen: aplicaciones en biología. España, Valencia.
- Ponger, L. y Mouchiroud, D. (2002). "CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences." Bioinformatics **18**(4): 631-633.
- Prinzie, A. D. y Vanden, P. (2007). "Predicting home-appliance acquisition sequences: Márkov/Márkov for Discrimination and survival analysis for modeling sequential information in NPTB models." Decision Support Systems **44**(1): 28–45.
- Prioleau, M. N. (2009). "CpG Islands: Starting Blocks for Replication and Transcription." PLoS Genet **5**(4).
- Pupo, M., Rodríguez, L. y Phan, D. (2006). An amino acid property-based semantic analysis of a stochastic sequence of amino acids using dynamic complex systems concepts. First International Workshop on Bioinformatics Cuba-Flanders 2006, UCLV. Santa Clara. Cuba,
- Reisman, D., Yates, J. y Sugden, B. (1985). "A putative origin of replication of plasmids derived from Epstein-Barr virus is composed of two cis-acting components." Mol. Cell. Biol. **5**: 1822-1832.
- Rivera-Borroto, O. M., Marrero-Ponce, Y., Meneses-Marcel, A., J.A., E., Gómez, A., Arán, V. J., Martins, M. A., Montero, D., Nogal, J. J., Torrens, F., Ibarra-Velarde, F., Vera, Y., Huesca-Guillén, A., Rivera, N. y Vogel, C. (2008). "Discovery of Novel Trichomonacids Using LDA-Driven QSAR Models and Bond-Based Bilinear Indices as Molecular Descriptors." QSAR & Combinatorial Science **28**(1): 9 - 26.
- Rodríguez, A. y Bonet, I. (2007). Sistema Multiagente para Combinar Técnicas de Aprendizaje Automatizado sobre Plataforma Libre. XII Convención y Expo Internacional de Informática, INFORMÁTICA 2007, La Habana,

- Rodríguez, A., Lorenzo-Ginori, J. y Grau, R. (2006). Detection of Coding Regions in Large DNA Sequences Using the Short Time Fourier Transform with Reduced Computational Load. CIARP,
- Rodríguez, A., Lorenzo-Ginori, J. y Grau, R. (2007a). "Coding Region Prediction in Genomic Sequences Using a Combination of Digital Signal Processing Approaches." CIARP: 635-642.
- Rodríguez, L., Casas, G. y Grau, R. (2007b). Validación del método Scan Generalizado con verdaderos falsos conglomerados. X Congreso Nacional de Matemática y Computación, Holguín,
- Rodríguez, L., Casas, G. y Grau, R. (2008a). Linear Fuzzy Scan Method to Detect Clusters. A Bioinformatic Application. XIV Latin Ibero-American Congress on Operations Research (CLAIO 2008), Cartagena de Indias. Colombia,
- Rodríguez, L., Casas, G., Grau, R. y Pupo, M. (2008b). "Generalización de dos métodos de detección de conglomerados. Aplicaciones en Bioinformática." Revista de Matemática: Teoría y Aplicaciones. **15** (1): 27 - 40.
- Rodríguez, L., Casas, G., Grau, R. y Martínez, Y. (2008c). "Fuzzy Scan Method to Detect Clusters." International Journal of Biomedical Sciences, © www.waset.org Spring 2008 **3**: 111 -115.
- Rodríguez, L., Casas, G., Grau, R. y Gómez, O. (2009). "Approximations for the distribution of Fuzzy Scan Statistics." Investigación Operacional **30**(2): 131-139.
- Rodríguez, M. (2006). Conducta suicida. Salud Mental Infanto - Juvenil. La Habana: 182., Ciencias Médicas.
- Romero, M. (2007). "Bioinformática: del wet al dry, y al web lab." RevistaeSalud.com **3**(11).
- Ruiz-Shulcloper, J. y Abidi, M. A. (2002). "Logical Combinatorial Pattern Recognition." ScientificConns CiteSeerX - Scientific Literature Digital Library and Search Engine (United States).
- Sahu, S., Bendel, R. B. y P., S. C. (1993). "Effect of relative risk and cluster configuration on the power of the one-dimensional Scan statistics." Statistics in Medicine **12**: 1853-1865.
- Salzberg, S. L., Salzberg, A. J., Kerlavage, A. R. y Tomb, J.-F. (1998). "Skewed oligomers and origins of replication." Genetics **217**: 57-67.
- Sánchez, R. y Grau, R. (2009). "An algebraic hypothesis about the primeval genetic code architecture." Mathematical Biosciences **221**(1): 60-76.
- Santovenia, J., Tarragó, C. y Cañedo, R. (2009). "Sistemas de información geográfica para la gestión de la información." ACIMED **20**(5).

- Schneider, T. D. y Stephens, R. M. (1990). "Sequence logos: a new way to display consensus sequences." Nucleic Acids Res **18**: 6097-6100.
- Service, T. C. y Tauritz, D. R. (2009). Free lunches in pareto coevolution. Genetic And Evolutionary Computation Conference archive. Proceedings of the 11th Annual conference on Genetic and evolutionary computation table of contents, Montreal, Québec, Canada, 1721-1728,
- Shad, D. A. y Madden, L. V. (2004). "Nonparametric Analysis of Ordinal Data in Designed Factorial Experiments." The American Phytopathological Society **94**(1): 33-43.
- Shamsir, M. S. y Mohamed Hussein, Z. A. (2010). "Across and beyond the divide: the role of inter-departmental teaching in bioinformatics." Teaching and Learning in Higher Education **2**(1): 30-40.
- Shi, Y. y Eberhart, R. (1998). Parameter Selection in Particle Swarm Optimization. In Proceedings of the Seventh Annual Conference on Evolutionary Programming: ,
- Shortliffe, E. H. y Buchanan, B. G. (1975). "A model of inexact reasoning in medicine." Mathematical Biosciences **23**: 351-379.
- Sokal, R. R. y Rohlf, F. J. (1995). The principles and practice of statistics in biological research. New York, W. H. Freeman and Company.
- Sugden, B. (2002). "In the beginning: A viral origin exploits the cell." Trends Biochem. Sci. **27**(1): 1-3.
- Tamura, K. y Nei, M. (1993). "Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees." Mol. Biol. Evol. **10**(3): 512-526.
- Tamura K, D. J. (2007). "MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0." Mol Biol Evol **24**: 1596-1599.
- Tanaka, H., Bergstrom, D., Yao, M. y Tapscott, S. (2005). "Widespread and nonrandom distribution of DNA palindromes in cancer cells provides a structural platform for subsequent gene amplification." Nat Genet. **37**(3): 320-7.
- Thompson, J. D., Higgins, D. G. y Gibson, T. J. (1994). "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice." Nucleic Acids Res. **22**: 4673-4680.
- Toledo, G. (2007). Fundamentos de salud pública. La Habana, Ciencias Médicas.
- Van-Rijsbergen, C. J. (1979). Information Retrieval. London, Butterworths.

- Vanhulsel, M., Janssens, D., Wets, G. y Vanhoof, K. (2009). "Simulation of sequential data: An enhanced reinforcement learning approach. ." Expert Syst. Appl. **36**(4): 8032-8039.
- Vasconcelos, A., Maia, M. y de Almeida, D. (2000). "Short interrupted palindromes on the extragenic DNA of Escherichia coli K-12, Haemophilus influenzae and Neisseria meningitidis." Bioinformatics **16**(11): 968-77.
- Wang, X., Yang, J., Teng, X., Xia, W. y Jensen, R. (2007). "Feature selection based on rough sets and particle swarm optimization." Pattern Recognition Letters **28**: 459-471.
- Wang, Z., Chen, Y. y Li, Y. (2004). "A brief review of computational gene prediction methods." Geno. Prot. Bioinfo **2**(4): 216-221.
- Weller, S. K., Spadaro, A., Schaffer, J. E., Murray, A. W., Maxam, A. M. y Schaffer, P. A. (1985). "Cloning, sequencing, and functional analysis of oriL, a herpes simplex virus type 1 origin of DNA synthesis." Mol. Cell. Biol. **5**: 930-942.
- Wolpert, D. (1996). "The Lack of A Priori Distinctions between Learning Algorithms." Neural Computation **8**(7): 1341-1390
- Wolpert, D. H. y Macready, W. G. (1997). "No Free Lunch Theorems for Optimization." IEEE Transactions on Evolutionary Computation **1**(1): 67-82.
- Wolpert, D. H. y Macready, W. G. (2005). "Coevolutionary free lunches." IEEE Transactions on Evolutionary Computation **9**(6): 721-735.
- Yager, R. R. (2008). Entropy and Specificity in a Mathematical Theory of Evidence, Springer Berlin / Heidelberg.
- YU, U., LEE, S. H., KIM, Y. J. y KIM, S. (2004). "Bioinformatics in the Post-genome Era." Journal of Biochemistry and Molecular Biology **37**: 75-82.
- Zadeh, L. A. (1973). "Outline of a new approach to the analysis of complex systems and decision processes." IEEE Trans. Sys. Man. Gybern. **1**(28-48).
- Zadeh, L. A. (1975). "Fuzzy Logic and Approximate Reasoning." Synthese **30**: 407-428.
- Zadeh, L. A. (1986). "A Simple View of the Dempster-Shafer Theory of Evidence and its Implication for the Rule of Combination." AI Magazine **7**(2): 85-90.
- Zadeh, L. A. (2002). "Toward a perception-based theory of probabilistic reasoning with imprecise probabilities." Journal of Statistical Planning and Inference **105**: 233–264.
- Zadeh, L. A. (2004). "Precisiated Natural Language (PNL)." AI Magazine **25**(3): 74-91.
- Zhu, Y., Huang, L. y Anders, D. G. (1998). "Human cytomegalovirus oriLyt sequence requirements." J. Virol **72**: 4989–4996.

Producción científica del autor sobre el tema de la tesis

Publicaciones revistas y memorias de eventos (en orden cronológico)

1. Casas, G.M., **Rodríguez, L.**, Grau, R., Cardoso, G., Chávez, M.C. (2005) Metodología general para la Validación de técnicas conglomerados. Boletín de la Sociedad Cubana de Matemática y Computación. ISSN 17286042. Vol. 3 No.1, 2005.
2. **Rodríguez, L.**, Casas, G.M., Grau, R., Cardoso, G., Ortega, S. Pupo, M. (2006) Scan Statistics. Bioinformatics Applications., Proceedings of First International Workshop on Bioinformatics Cuba-Flanders'2006, Santa Clara, Feb. 7-10, ISBN: 959-250-239-0.
3. Pupo, M., **Rodríguez, L.**, Phan, D. (2006) An amino acid property-based semantic analysis of a stochastic sequence of amino acids using dynamic complex systems concepts. Proceedings of First International Workshop on Bioinformatics Cuba-Flanders'2006, Santa Clara, Feb.7-10, ISBN: 959-250-239-0
4. **Rodríguez, L.**, Casas, G.M., Grau, R., Cardoso, G. (2006) Aplicación de los métodos Scan en Bioinformática. Memorias de UCIENCIA 2006. II Conferencia Científica de la Universidad de Las Ciencias Informáticas. III Taller de Bioinformática de la UCI., La Habana, Julio 4-6. ISBN: 959-16-0463-7.
5. **Rodríguez, L.**, Casas, G.M., Grau, R. (2007) Validación del método Scan con verdaderos y falsos conglomerados. Memorias de COMPUMAT 2007. X Congreso Nacional de Matemática y Computación. Holguín Noviembre 21-23. ISBN: 1728-6042.
6. **Rodríguez, L.**, Casas, G.M., Grau, R., Martínez, Y. (2008) Fuzzy Scan Method to detect Clusters Proceedings of Second Workshop on Bioinformatics Cuba – Flanders, February, 2008. Publicado en la revista International Journal of Biomedical Sciences, © www.waset.org **Spring** Vol.3: 111 -115. 2008.
7. **Rodríguez, L.**, Casas, G.M., Grau, R., Pupo M. (2008) Generalización del método Scan. El método Scan Lineal Borroso. SIMMAC XVI. Simposio Internacional de Métodos Matemáticos Aplicados a las Ciencias. Costa Rica. Feb. 19-21. Trabajo aceptado para el evento.

8. **Rodríguez, L.**, Casas, G.M., Grau, R. (2008) Approximations for the distribution of Fuzzy Scan Statistics. ICOR 2008. 8th International Conference on Operations Research. Havana. February 25-29. Publicado en Revista Investigación Operacional Vol. 30, No.2, 131-139, 2009
9. **Rodríguez, L.**, Casas, G.M., Grau, R., Pupo M. (2008) Generalización de dos métodos de detección de conglomerados. Aplicaciones en Bioinformática. Revista de Matemática: Teoría y Aplicaciones. Vol. 15 No. 1; 27-40
10. **Rodríguez, L.**, Casas, G.M., Grau, R., Pupo M. (2008) Cluster Detection Using Fuzzy Logic. A Bioinformatic Application With Fuzzy Scan Method. BIOCOMP'08 International Conference on Bioinformatic and Computational Biology. USA July 14-17. Paper aceptado para el evento con número de inscripción BIC9158
11. **Rodríguez, L.**, Casas, G.M., Grau, R., Pupo M. (2008) Linear Fuzzy Scan Method to Detect Clusters. A Bioinformatic Application. Memorias de XIV Congreso Latino-Iberoamericano en Investigación de Operaciones (CLAIO 2008). Cartagena de Indias, Colombia. Sep. 9-12. ISBN: 978 958 825283-4
12. Díaz, J.E., Casas, G., Alvarez M., **Rodríguez, L.**, (2009) Detección de conglomerados de enfermos dados por tumores malignos. Municipio de Cifuentes. XVII Fórum de Ciencia y Técnica del Sectorial de Salud de Cifuentes. 4 de Abril.
13. Díaz, F., **Rodríguez, L.**, Casas, G.M., Grau, R. (2009) Análisis de los parámetros del Scan Lineal utilizando diseño de experimento. Memorias del Primer Taller Internacional FIMAT XXI. Holguín. Mayo 26-30. ISBN: 978-959-18-0498-3
14. Valdés, E., **Rodríguez, L.** y Casas, G. (2009). Herramienta computacional para la detección de conglomerados en secuencias de ADN usando los métodos Scan. Informática en Salud 2009, código SLD062. La Habana. Feb. 9-13. Online en Internet http://informatica2009.sld.cu/pageTemp_ListarTrab?b_start:int=100&desde=pública
[dos](#).
15. **Rodríguez, L.**, Casas, G.M., Grau, R. (2009) Cluster Detection in DNA Sequences using the Fuzzy Circular Method. Memorias RECPAT 2009. Congreso Nacional de Reconocimiento de Patrones. Santiago de Cuba. Dic. 8-10. ISBN: 978-959-207-381-4

16. **Rodríguez, L.**, Casas, G.M., Grau, R. (2010) Optimización basada en enjambres de partículas para detectar los parámetros óptimos del método Scan Borroso. ICOR 2010. 9th International Conference on Operations Research. Havana. Feb. 22-26.
17. **Rodríguez, L.**, Casas, G.M., Silveira, P., Grau, R., Díaz, F. (Noviembre 2010) Optimización de parámetros en los Métodos Scan Generalizados. Revista de la Facultad Ingeniería de la Universidad de Antioquia. Vol. 65

Se tiene además el siguiente registro de software:

Rojas, Y., Rodríguez, L., Casas, G.M. Registro de Software número 2382-2009 del Centro Nacional de Derecho de Autor a favor de: Optimus, Software para calcular valores óptimos de los parámetros del método Scan, mediante la unión de algoritmo bioinspirados (PSO) y el método de simulación de Mote Carlo. Octubre del 2009.

Anexos

Anexo 1: ANOVA bifactorial no-paramétrico

Implementación del ANOVA bifactorial no-paramétrico en el paquete *Mathematica*.

```
RankValues[values_]:= Module[{s,m,r,a,means,ranks,rules},
  s=Split[Sort[values]];
  m=Map[Length,s];
  a=Accumulate[m];
  r=Range[1,Length[values]];
  means=Map[Mean,Drop[MapThread[Function[{i,k},Take[Drop[r,k],i]],
    {Append[m,0],Prepend[a,0]}],-1]];
  ranks=MapThread[Function[{i,j},Table[i,{j}]],{means,m}]/N;
  rules=MapThread[Function[{i,j},i[[1]]->j[[1]]],{s,ranks}];
  ReplaceAll[values,rules]
];

test[nrep_,lf1_,lf2_,namef1_,namef2_,sqsumf1_,sqsumf2_,sqsumf1f2_]:=
Module[{cmtot,grlf1,grlf2,Hf1,Hf2,Hf1f2,sigf1,sigf2,sigf1f2,finalt},
  cmtot=nrep*lf1*lf2*(nrep*lf1*lf2+1)/12;
  {Hf1,Hf2,Hf1f2}=N[{sqsumf1,sqsumf2,sqsumf1f2}/cmtot,4];
  {grlf1,grlf2}={lf1,lf2}-1;grlf1f2=grlf1*grlf2;
  sigf1=N[1-CDF[ChiSquareDistribution[grlf1],Hf1],3];
  sigf2=N[1-CDF[ChiSquareDistribution[grlf2],Hf2],3];
  sigf1f2=N[1-CDF[ChiSquareDistribution[grlf1f2],Hf1f2],3];
  finalt=PaddedForm[TableForm[Transpose[{{Hf1,Hf2,Hf1f2},{sigf1,sigf2,sigf1f2}}],
    TableHeadings->{{namef1,namef2,namef1<>"*"<>namef2}, {" H", "Sign"}},{10,3}];
  Return[finalt]
];

BifactorialNonParamANOVA[data_,nrep_,lf1_,lf2_,namef1_,namef2_]:=
Module[{datanew,res},
  datanew=data;
  datanew=Transpose[datanew];
  datanew[[3]]=RankValues[datanew[[3]]];
```

```

datanew=Transpose[datanew];
res=ANOVA[datanew,{namef1,namef2,All},{namef1,namef2}];
test[nrep,lf1,lf2,namef1,namef2,res[[1]][[2]][[1]][[1]][[2]], res[[1]][[2]][[1]][[2]][[2]],
res[[1]][[2]][[1]][[3]][[2]]]
];

```

La función RankValues tiene el parámetro:

values: lista de valores de la variable dependiente que serán ranqueados.

La función test tiene los siguientes parámetros:

nrep: Representa el número de réplicas (constante en cada combinación de valores de los factores)

lf1: Niveles del factor 1

lf2: Niveles del factor 2

namef1: Nombre del factor 1

namef2: Nombre del factor 2

sqsumf1: Suma de cuadrados del factor 1

sqsumf2: Suma de cuadrados del factor 2

sqsumf1f2: Suma de cuadrados de la interacción

La función BifactorialNonParamANOVA tiene los siguientes parámetros:

nrep, lf1, lf2, namef1, namef2: Como en la función test

Una vez cargadas las funciones, será invocada la función BifactorialNonParamANOVA con los parámetros correspondientes a cada análisis. Un ejemplo, sería:

```

BifactorialNonParamANOVA[{{1,1,100.},{1,2,100.},{2,1,100.},{2,2,100.},{3,1,86.},{3,2,84.85},
{1,1,100.},{1,2,99.3},{2,1,100.},{2,2,100.},{3,1,81.65},{3,2,78.95},
{1,1,99.15},{1,2,87.1},{2,1,99.9},{2,2,96.25},{3,1,74.1},{3,2,68.2}},
3,3,2,"Ventana","Paso"]

```

La respuesta del Mathematica será una tabla como la siguiente:

	H	Sign
Ventana	11.556	0.000
Paso	0.329	0.566
Ventana * Paso	0.052	0.969

Anexo 2. Scan Lineal Generalizado

```

ScanValidation[sec_, AnchoW_, Paso_] :=
  CompoundExpression[ (* Buscar parámetro de la distribución de Poisson*)
    W = Partition[sec, AnchoW, Paso]; (* Particiona la secuencia en ventanas*)
    Win = N[Map[Function[lis, Plus@@lis], W], 8]; (* # de unos de cada ventana*)
    media = N[Mean[Win]]; (* Promedio de unos por ventana "Landa de Poisson" *)
    maximo = Max[Win]; (* Ventana de mayor número de unos *)
    L = N[Length[sec]/AnchoW, 9]; (* Fracción de ventanas mínimas a formar *)
    signifs = N[pFinal[media, maximo, L], 10]; (* Busca la significación del estadígrafo. *)
    Return[signifs]
  ];

```

Para calcular la significación se utiliza el procedimiento pFinal donde están programadas las formulas aproximadas de (Naus 1982):

```

Fnn[media_, i_] := Module[{}],
  If[max<0, p:=0, p:=CDF[PoissonDistribution[media], i]]; Return[N[p, 10]]
]
Psi[media_, i_] := Module[{}],
  p := PDF[PoissonDistribution[media], i]; Return[N[p, 10]]
]

```

```

pFinal[max_, media_, L_] := 1 - Q[max, media, L]
Q[max_, media_, L_] := Q2[max, media] ( Q3[max, media] / Q2[max, media] )L-2
Q2[max_, media_] := Fnn[media, max-1]2 - (max - 1) Psi[media, max] Psi[media, max-2] - (max - 1 - media)
  Psi[media, max] Fnn[media-3, max]
Q3[max_, media_] := Fnn[media, max-1]3 - A1[media, max] + A2[media, max] + A3[media, max] - A4[media, max]
A1[media_, max_] := 2Psi[media, max] Fnn[media, max-1] ((max-1) Fnn[media, max-2] - media Fnn[media, max-3])
A2[media_, max_] := 0.5 Psi[media, max]2 ((max - 1) (max - 2) Fnn[media, max - 3] - 2(max - 2) media
  Fnn[media, max - 4] + media2 Fnn[media, max - 5])
A3[media_, max_] :=  $\sum_{r=1}^{max-1} Psi[media, 2 max - r] Fnn[media, r-1]^2$ 
A4[media_, max_] :=  $\sum_{r=2}^{max-1} Psi[media, 2 max - r] Psi[media, r] ((r - 1) Fnn[media, r] - max Fnn[media, r - 3])$ 

```

Anexo 3. Scan Circular Generalizado

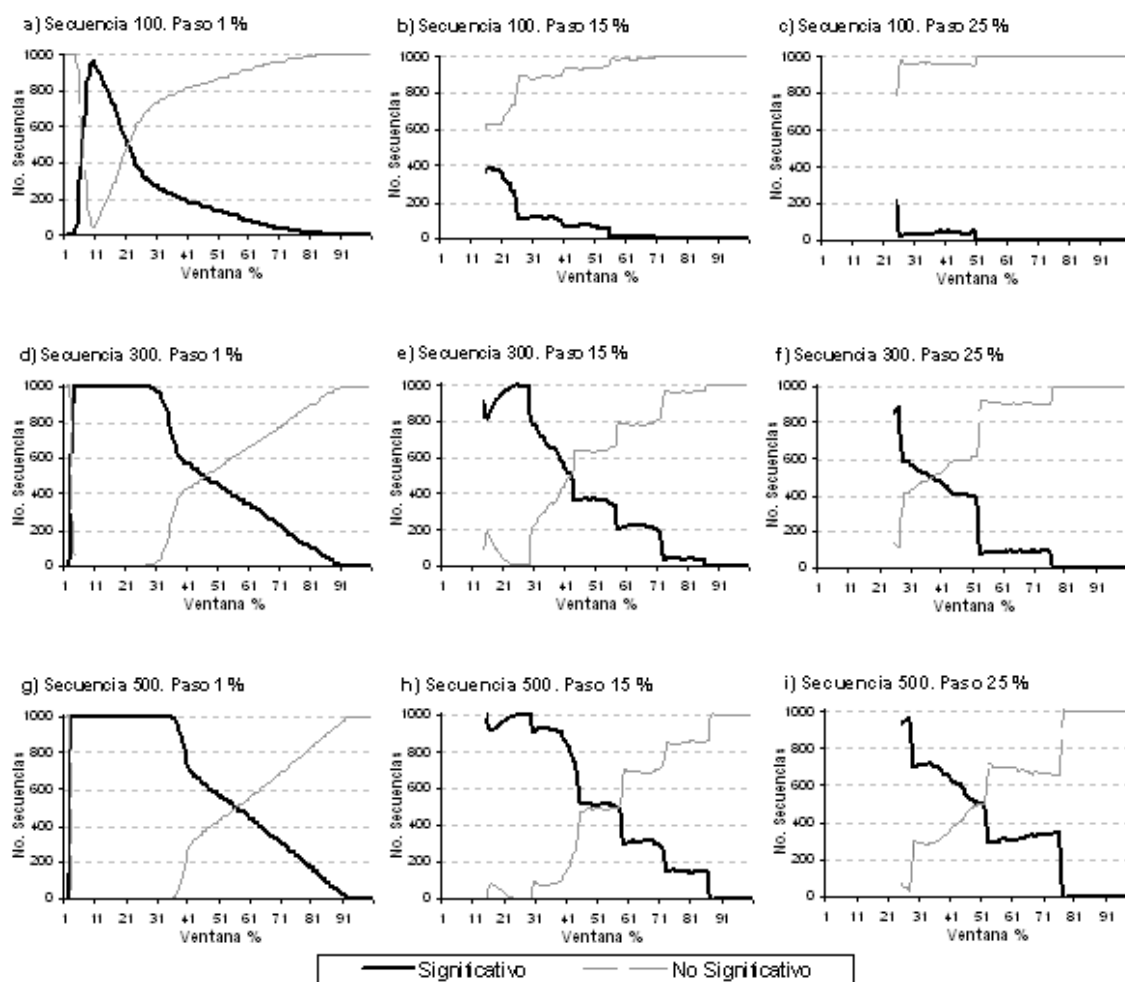
ScanValidation[se1_, AnchoW_, Paso_] :=

```
CompoundExpression[ (* Buscar parámetro de la distribución de Poisson*)
  sec=Join[se1,Take[se1,t-1]]; (* Convertir la lista en secuencia circular*)
  W = Partition[sec,AnchoW,Paso]; (* Dividir la secuencia en ventanas*)
  Win = N[Map[Function[lis,Plus@@lis], W], 8]; (* # de unos de cada ventana*)
  media = N[Mean[Win]]; (* Promedio de unos por ventana "Landa de Poisson" *)
  maximo = Max[Win];(*Print[Win];*) (*Ventana de mayor número de unos *)
  L = N[Length[sec]/AnchoW, 9]; (* Fracción de ventanas mínimas a formar *)
  signifs = N[pFinal[media,maximo,L],10]; (* Busca la significación del estadígrafo.*)
  Return[signifs]
];
```

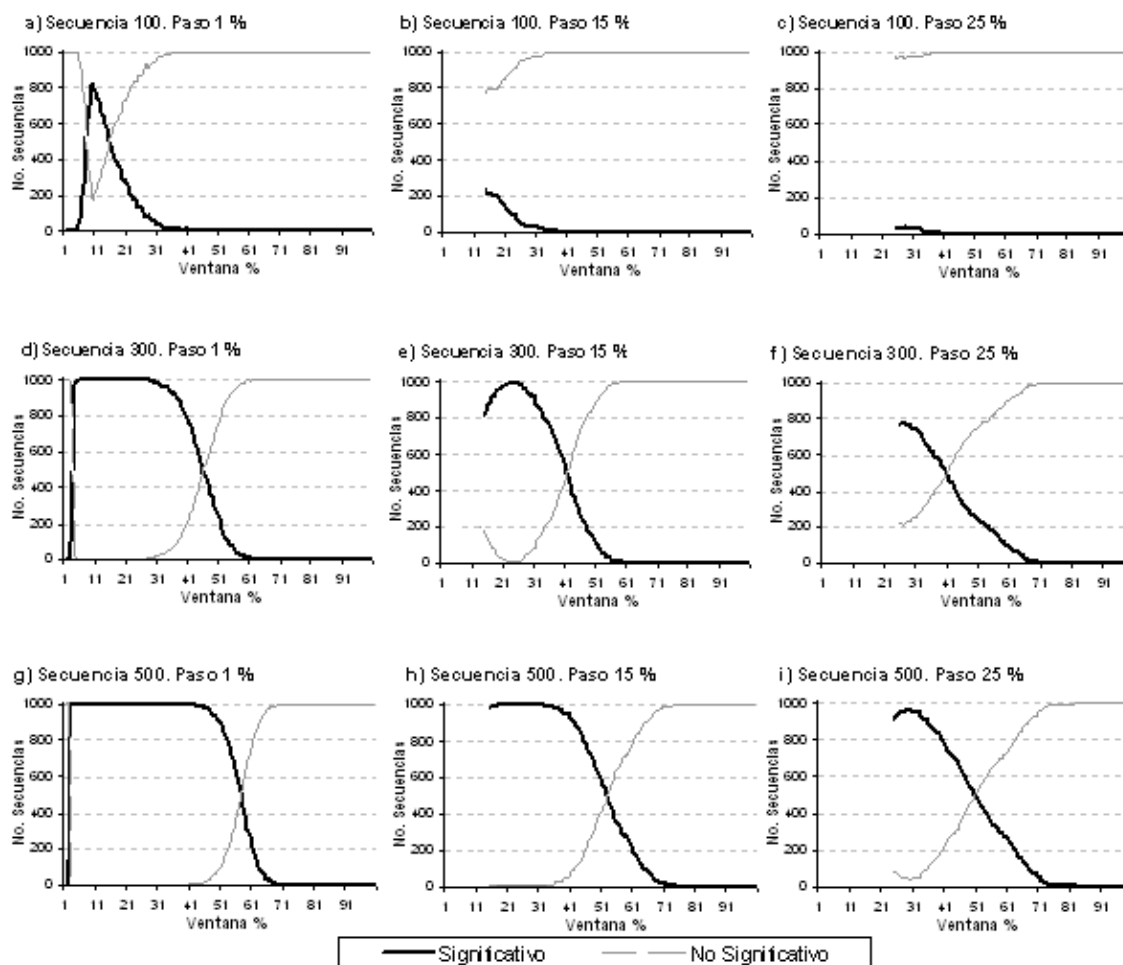
La significación se calcula utilizando las formulas aproximadas de (Naus 1982):

```
Fnn[media_, i_] := Module[{} ,
  If[max<0,p:=0, p:=CDF[PoissonDistribution[media], i]]; Return[N[p,10]]
]
Psi[media_, i_] := Module[{} ,
  p :=PDF[PoissonDistribution[media], i]; Return[N[p,10]]
]
pFinal[max_, media_, L_] := 1 - Q[max, media, L]
Q[max_, media_, L_] := Q4[max, media] Q3[max, media]L-2 Q2[max, media]L-1
Q2[max_, media_] := Fnn[media,max-1]2 - (max - 1) Psi[media,max] Psi[media, max-2] - (max - 1 - media)
  Psi[media, max] Fnn[media-3, max]
Q3[max_, media_] := Fnn[media, max-1]3-A1[media, max]+A2[media, max]+A3[media, max]-A4[media, max]
A1[media_, max_] := 2Psi[media,max]Fnn[media,max-1] ((max-1)Fnn[media,max-2]-media Fnn[media,max-3])
A2[media_, max_] := 0.5 Psi[media, max]2 ((max - 1) (max - 2) Fnn[media, max - 3] - 2(max - 2) media
  Fnn[media, max - 4] + media2 Fnn[media, max - 5])
A3[media_, max_] :=  $\sum_{r=1}^{max-1} Psi[media, 2 max - r] Fnn[media, r - 1]^2$ 
A4[media_, max_] :=  $\sum_{r=2}^{max-1} Psi[media, 2 max - r] Psi[media, r] ((r - 1) Fnn[media, r] - max Fnn[media, r - 3])$ 
Q4[max_, media_] := Q3[max, media]2 / Q2[max, media]
```

Anexo 4. Scan Lineal Modificado con verdaderos conglomerados creados con el 10% del tamaño total de la secuencia



Anexo 5. Scan Circular Modificado con verdaderos conglomerados creados con el 10% del tamaño total de la secuencia



Anexo 6. Scan Lineal Borroso

```

ScanValidation[sec_, AnchoW_, Paso_, Incr_] :=
CompoundExpression[ (* Buscar parámetro de la distribución de Poisson*)
W = Partition[sec, AnchoW, Paso]; (* Particiona la secuencia en ventanas*)
Win = N[Map[Function[lis, Plus@@lis], W], 8]; (* # de unos de cada ventana*)
media = N[Mean[Win]]; (* Promedio de unos por ventana "Landa de Poisson" *)
If[Incr > 0, W = Fuzzy[sec, AnchoW, Paso, Incr]]; (* Procedimiento que suaviza ventanas *)
Win = Map[Function[lis, Plus@@lis], Win]; (* suma los valores de cada ventana*)
maximo = Max[Win]; (*Print[Win];*) (*Ventana de mayor valor*)
L = N[Length[sec]/AnchoW, 9]; (* Fracción de ventanas mínimas a formar *)
signifs = If[Incr == 0, {N[Pfinal[media, maximo, L], 10]}; (*Significación Scan Clásico.*)
signifs = If[Incr <> 0 {DesFuzzificacion[N[Pfinal[media, Round[maximo], L], 10]],
DesFuzzificacion[N[Pfinal[media, maximo, L], 10]],
DesFuzzificacion[N[NausSignif[media, maximo, L], 10]]}];
(*Significación Scan Borroso, por las tres vías permitidas.*)

Return[signifs]
]

```

Para calcular la significación del Scan Lineal Borroso, se utilizan dos procedimientos, el primero para la aproximación borrosa 1 y 2, el segundo para la aproximación borrosa 3.

Primer procedimiento

```

Fnn[media_, i_] := Module[{},
If[max < 0, p := 0, p := CDF[PoissonDistribution[media], Floor[i]] +
PDF[PoissonDistribution[media], Ceiling[i]] * FractionalPart[i];
Return[N[p, 10]]
]
Psi[media_, i_] := Module[{},
p := Psi1[media, Floor[i]] - (Psi1[media, Floor[i]] - Psi1[media, Ceiling[i]]) * FractionalPart[i];
Return[N[p, 10]]
]
Psi1[media_, i_] := Module[{},
p := PDF[PoissonDistribution[media], i]; Return[N[p, 10]]
]

```

```

pFinal[max_, media_, L_] := 1 - Q[max, media, L]
Q[max_, media_, L_] := Q2[max, media] (Q3[max, media] / Q2[max, media])L - 2

```

$$Q2[\text{max_}, \text{media_}] := \text{Fnn}[\text{media}, \text{max}-1]^2 - (\text{max} - 1) \text{Psi}[\text{media}, \text{max}] \text{Psi}[\text{media}, \text{max}-2] - (\text{max} - 1 - \text{media}) \text{Psi}[\text{media}, \text{max}] \text{Fnn}[\text{media}-3, \text{max}]$$

$$Q3[\text{max_}, \text{media_}] := \text{Fnn}[\text{media}, \text{max}-1]^3 - A1[\text{media}, \text{max}] + A2[\text{media}, \text{max}] + A3[\text{media}, \text{max}] - A4[\text{media}, \text{max}]$$

$$A1[\text{media_}, \text{max_}] := 2 \text{Psi}[\text{media}, \text{max}] \text{Fnn}[\text{media}, \text{max}-1] ((\text{max}-1) \text{Fnn}[\text{media}, \text{max}-2] - \text{media} \text{Fnn}[\text{media}, \text{max}-3])$$

$$A2[\text{media_}, \text{max_}] := 0.5 \text{Psi}[\text{media}, \text{max}]^2 ((\text{max} - 1) (\text{max} - 2) \text{Fnn}[\text{media}, \text{max} - 3] - 2(\text{max} - 2) \text{media} \text{Fnn}[\text{media}, \text{max} - 4] + \text{media}^2 \text{Fnn}[\text{media}, \text{max} - 5])$$

$$A3[\text{media_}, \text{max_}] := \sum_{r=1}^{\text{max}-1} \text{Psi}[\text{media}, 2 \text{max} - r] \text{Fnn}[\text{media}, r - 1]^2$$

$$A4[\text{media_}, \text{max_}] := \sum_{r=2}^{\text{max}-1} \text{Psi}[\text{media}, 2 \text{max} - r] \text{Psi}[\text{media}, r] ((r - 1) \text{Fnn}[\text{media}, r] - \text{max} \text{Fnn}[\text{media}, r - 3])$$

Segundo procedimiento

```

FPsi1[max_, flpdf_] := Module[{},
  If[max<0, p=0, FPsi=Interpolation[flpdf]; p=FPsi[max]];
  Return[N[p,10]] (*Calcula probabilidad puntual usando función de interpolación *)
]
FFnn1[max_, flcdf_] := Module[{},
  If[n<0, p=0, FFnn=Interpolation[flcdf]; p=FFnn[max]];
  Return[N[p,10]] (*Calcula probabilidad acumulada usando función de interpolación *)
]
NausSignif[media_, maximo_, L_] :=
Module[{}, (*lp Función de interpol. de probabilidades lc Función de interpol. de probabilidades acumulada*)
  lp = Table[{k, PDF[PoissonDistribution[media], k]}, {k, -1, 2 max+1}];
  lc = Table[{k, CDF[PoissonDistribution[media], k]}, {k, -1, 2 max+1}];
  FA1 = 2 FPsi1[max, lp] FFnn1[max-1, lc] ((max-1) FFnn1[max-2, lc] - media FFnn1[max-3, lc]);
  FA2 := 0.5 (FPsi1[max, lp])^2 ((max-1) (max-2) FFnn1[max-3, lc] - 2(max-2) media FFnn1[max-4, lc] +
    media^2 FFnn1[max-5, lc]);
  FA3 := \sum_{r=1+FractionalPart[max]}^{\text{max}-1} \text{Fpsi1}[2 \text{max}, \text{lp}] \text{FFnn1}[r-1, \text{lc}]^2 ;
  FA4 := \sum_{r=2+FractionalPart[\text{max}]}^{\text{max}-1} \text{FPsi1}[2 \text{max}-r, \text{lp}] \text{FPsi1}[r, \text{lp}] ((r-1) \text{FFnn1}[r-2, \text{lc}] - \text{media} \text{FFnn1}[r-3, \text{lc}])
  FQ2 := FFnn1[max-1, lc]^2 - (max-1) FPsi1[max, lp] FPsi1[max-2, lp] - (max-1-media) FPsi1[max, lp]
    FFnn1[max-3, lc]
  FQ3 := FFnn1[n-1, lc]^3 - FA1 + FA2 + FA3 - FA4;
  FQ := FQ2 ( FQ3 / FQ2 )^{L-2};
  Pfin := 1-FQ;
  Return[N[Pfin,10]];
]

```

Para suavizar las ventanas con los procedimientos:

- **Fuzzy:** Procedimiento general borroso, que dirige los siguientes procedimientos:
 - **IncrementTamWindows:** Suaviza todas las ventanas de la secuencia, añadiendo los elementos adecuados por la izquierda de cada ventana y posteriormente hace el procedimiento por la derecha.
 - **Fuzzificacion:** Pesa los elementos suavizados de cada ventana en dependencia de su valor y posición dentro de la ventana.

Procedimiento General

```
Fuzzy[sec_, AnchoW_, Paso_, Incr_] :=
CompoundExpression[
W = IncrementTamWindows[sec, AnchoW, Paso, Incr];          (* Suaviza todas las ventanas *)
Inc1 = 1/(Incr+1);    (*Fracción general que aporta al peso cada elemento suavizado de una ventana*)
TW = Length[W];
W2 = W;                                                       (*W, W2 lista con las ventanas suavizadas*)
Map[Function[x,W2 = Fuzzificacion[W2, TW, Inc1*x, x]],Range[Incr]];  (*Valoriza parte izquierda*)
L2 = AnchoW+2*Incr+1;                                         (*L2 cantidad de elementos de una ventana suavizada*)
Map[Function[x, W2 = Fuzzificacion[W2, TW, Inc1*x, L2-x]],Range[Incr]]; (*Valoriza parte derecha*)
Return[W2]
]
```

Procedimiento que permite suavizar cada ventana de la secuencia

```
IncrementTamWindows[sec_, AnchoW_, Paso_, Incr_] :=
CompoundExpression[
W = Partition[sec, AnchoW, Paso];
TW = Length[W];
sec1=PadLeft[sec,Length[sec]+Incr];          (*Inserta ceros a la izquierda de la secuencia*)
W1 = Map[Function[z,Join[Take[sec1,{(z-1)*Paso+1,(z-1)*Paso+Incr}],W[[z]]],Range[TW]];
                                         (*Suaviza parte izquierda de las ventanas *)
sec1 = PadRight[sec,Length[sec]+Incr];       (*Inserta ceros a la derecha de la secuencia*)
W1 = Map[Function[z,Join[W1[[z]],Take[sec1,{(z-1)*Paso+1+AnchoW,(z-1)*Paso+ AnchoW
                                         +Incr}]]],Range[TW]];
                                         (*Suaviza parte derecha de las ventanas *)
Return[W1];
]
```

Procedimiento que pesa los elementos suavizados de cada ventana

Fuzzificacion[W_, L_, Val_, Pos_] :=

```
CompoundExpression[
  K=Map[Function[z,If[(W[[z, Pos]]!=0),ReplacePart[W[[z]],Val*W[[z,Pos]],Pos],W[[z]]],Range[L]];
  (* Dada la posición de un elemento móvil lo pesa según su valor en todas las ventanas *)
  Return[K]
];
```

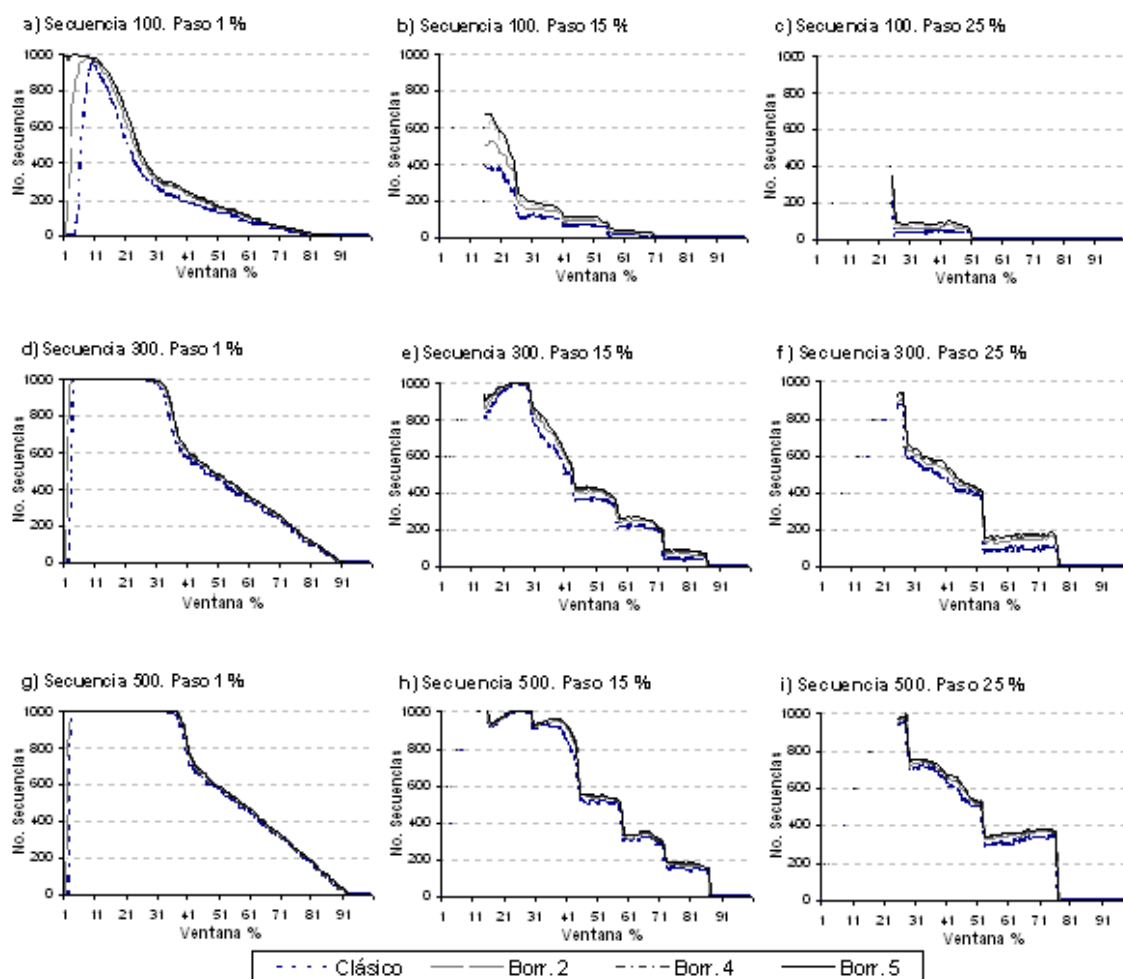
El valor borroso se desborrosifica utilizando la variante que toma como resultado final el conjunto borroso de mayor valor.

DesFuzzificacion[x1_] :=

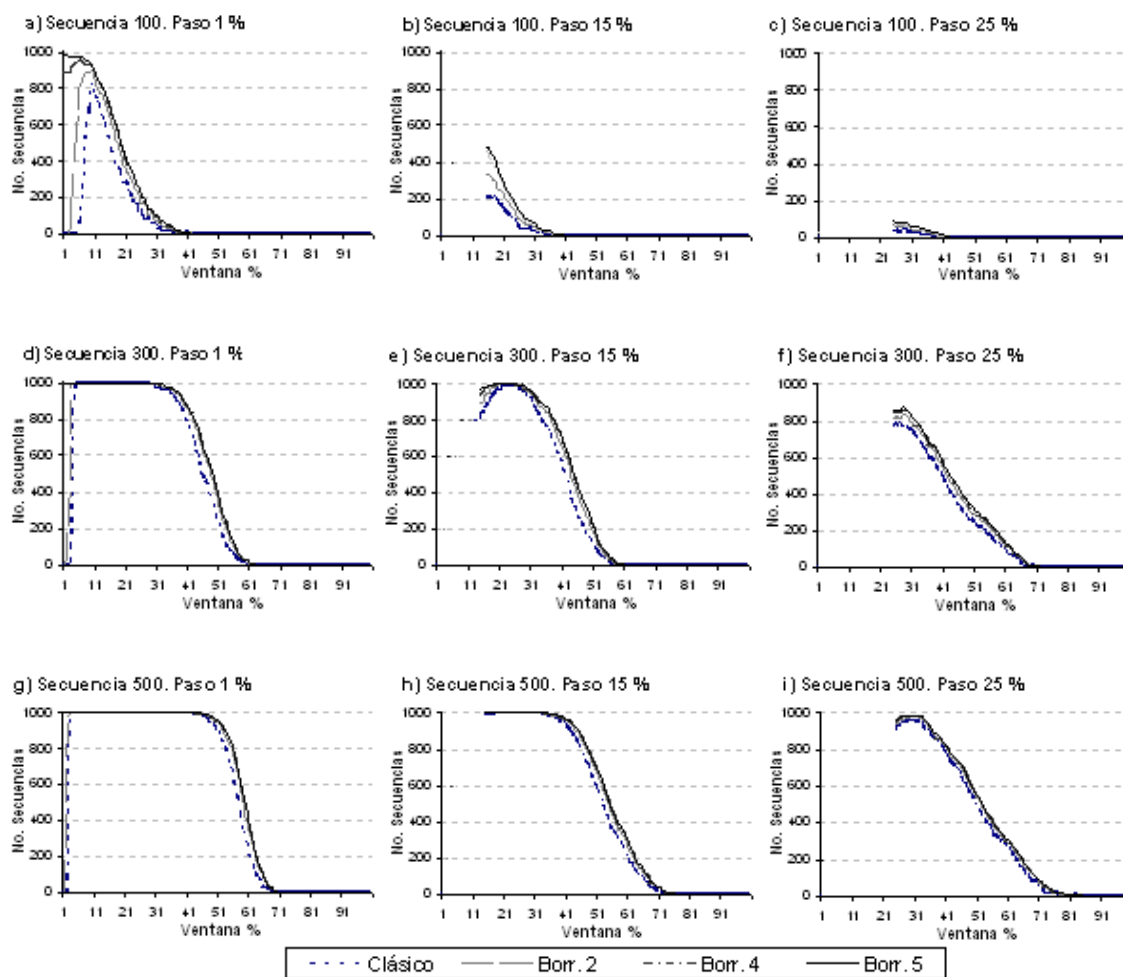
```
CompoundExpression[
  Which[
    x1 ≥ 0.075, gs = 0,
    x1 ≤ 0.05, gs = 1,
    x1 < 0.0625, gs = 1 - 2 * ((x1-0.05) / 0.025) 2,
    x1 < 0.075, gs = 2 * ((x1-0.075) / 0.025) 2
  ];
  (*Calcular grado de pertenencia de x al conjunto borroso significativo *)
  Which[
    x1 ≤ 0.05, ns = 0,
    x1 ≤ 0.075, ns = 1,
    x1 < 0.0625, ns = 2 * ((x1-0.05) / 0.025) 2,
    x1 < 0.075, ns = 1-2 * ((x1-0.075) / 0.025) 2
  ];
  (*Calcular grado de pertenencia de x al conjunto borroso no significativo*)
  DF1 = If[gs ≥ ns,"Signif.,"No Signif."]; (*Calcula definitivamente el conjunto al cual pertenece*)
  Return[DF1];
]
```

El Scan Circular Borroso posee estas mismas opciones los que hay que convertir la secuencia en una lista circular y para suavizar las ventanas iniciales y finales se le añade los elemento que le siguen a continuación en la lista.

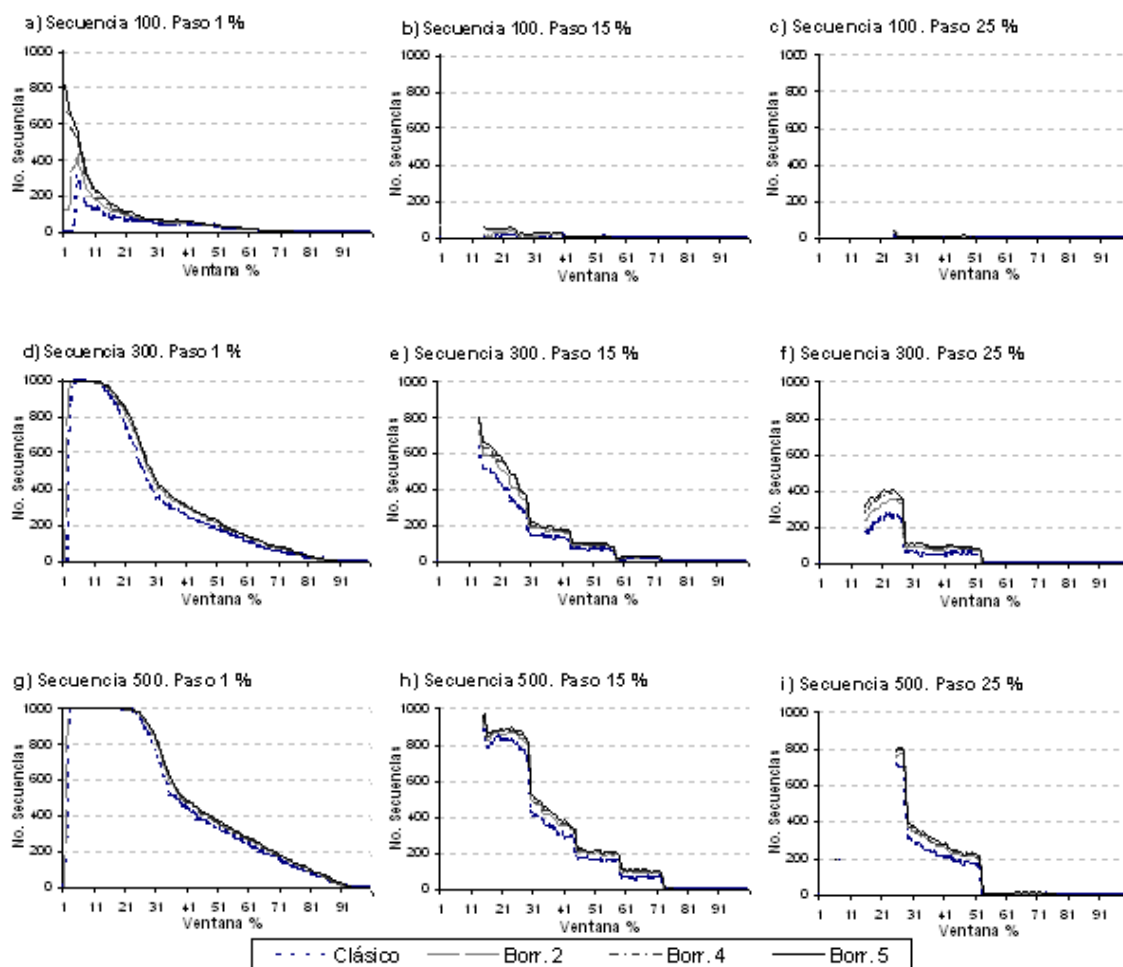
Anexo 7. Scan Lineal Borroso con verdaderos conglomerados creados con el 10 % del tamaño total de la secuencia



Anexo 8. Scan Circular Borroso con verdaderos conglomerados creados con el 10% del tamaño total de la secuencia



Anexo 9. Scan Lineal con verdaderos conglomerados creados con el 5% del tamaño total de la secuencia



Anexo 10. Scan Circular con verdaderos conglomerados creados con el 5% del tamaño total de la secuencia

