Tecnologías bioinformáticas para el análisis de secuencias de ADN

Bioinformatics Technologies for the Analysis of DNA sequences.

Carlos Augusto Meneses Escobar[1], Lizeth Vanessa Rozo Murillo[2], Jhenifer Franco Soto[3]
Universidad Tecnológica de Pereira, Pereira, Colombia
cmeneses@utp.edu.co
mikeropi@gmail.com
jhefiner@gmail.com

***Resumen*— La información contenida en secuencias de ADN, por su contenido voluminoso requiere de técnicas inteligentes para el modelamiento de los datos y de métodos computacionales avanzados para el procesamiento de estos. Se busca optimizar el tiempo en el que se ejecutan cálculos e inferencias, y mejorar la confiabilidad de los análisis que se realizan a partir de los resultados obtenidos, los cuales pueden servir de base para el desarrollo de investigaciones científicas. El grupo de investigación GIA del programa Ingeniería de Sistemas y Computación de la Universidad Tecnológica de Pereira, se encuentra trabajando en la determinación de tecnologías informáticas que permitan hacer avances significativos en los desarrollos científicos en el campo de la biología. Este artículo explora que técnicas computacionales son pertinentes en el desarrollo de aplicaciones bioinformáticas..**

***Palabras clave*— ADN, ARN, bioinformática, biología, datos biológicos, proteínas, secuencias, tecnologías.**

***Abstract*— ADN sequences data require special processing systems due the corpus of its information toward data modeling and computational methods to process them. We want to optimize the time using to operations and inferences and improve reliability of the analysis from the obtained results, that you can use as a base to develop scientist's researches. GIA Group of Research in Systems Engineering Program from Universidad Tecnológica de Pereira is working in these new technologies looking for meaningful advances in the biology area. This article presents computational techniques using in bioinformatiques applications.**

***Key Word* — bioinformatics, biological data, biology, DNA, protein RNA, sequences, technologies.**

I. INTRODUCCIÓN

La comunidad científica que realiza investigación dentro del área biológica, en el afán de encontrar respuestas a estudios de la estructura molecular y las secuencias de ADN, dia a dia se enfrenta a mayores retos que implican el manejo de enormes volúmenes de datos que crecen de manera exponencial en tamaño y complejidad, debido a los avances tecnológicos que permiten hacer cálculos más precisos.

Afortunadamente, el desarrollo tecnológico tanto en el ámbito de la electrónica como el desarrollo de software y las telecomunicaciones han permitido un avance significativo en las técnicas para el procesamiento y análisis inteligente de los datos, beneficiando los estudios científicos que permiten conocer mejor las estructuras de los organismos vivos.

La complejidad que conlleva el manejo de grandes volúmenes de datos exige de procesos computacionales con alto nivel de desempeño en cuanto a espacio y tiempos de respuesta.

En este artículo se revisan algunas de las metodologías y técnicas computacionales que más se utilizan para el análisis de secuencias de ADN, que a su vez puedan ser utilizadas como base para desarrollar herramientas prácticas en trabajos futuros.

EL presente artículo, presenta en la sección 2, conceptos sobre bioinformática sus alcances y aplicaciones. En la sección 3, se hace referencia a las secuencias de ADN y su alineamiento. En la sección 4, se tratan las tecnologías computacionales orientadas al desarrollo y aplicación de soluciones bioinformáticas, entre las que se mencionan las bases de datos biológicas, las bodegas de datos, la minería de datos, las máquinas de aprendizaje y el uso de matlab. Finalmente, en la sección 5 se presentan conclusiones y trabajos futuros

[1] Ingeniero de Sistemas y Comp.

Fecha de Recepción: 26 de Agosto de 2011
Fecha de Aceptación: 12 de Octubre de 2011

[2] Estudiante de Ingeniería de Sistemas y Computación

[3] Estudiante de Ingeniería de Sistemas y Computación

II. BIOINFORMÁTICA

Las tareas más importantes de las que se ocupa la bioinformática consisten en entender las correlaciones, las estructuras y los patrones en los datos biológicos.

En los últimos años, la Bioinformática ha atraído la conjugación de varias disciplinas, entre las que están la informática, las matemáticas, la estadística, la química y las ciencias biológicas no tradicionales. Esto se debe a la disponibilidad de enormes cantidades de datos biológicos públicos y privados, y a la necesidad imperiosa de transformar datos en información biológica útil y en conocimiento.

La aplicación de estas disciplinas con técnicas computacionales inteligentes, sirven para la creación de proyectos que conlleven el descubrimiento y desarrollo de fármacos, análisis del genoma y control biológico, entre otros. Esto implica el uso de tecnologías informáticas y métodos estadísticos para manejar y analizar un gran volumen de datos biológicos sobre el ADN, el ARN y las secuencias de proteínas, estructuras de las proteínas, los perfiles de expresión genética y las interacciones de la proteína.

A. Alcance de la Bioinformática.

La Bioinformática se compone de dos subcampos complementarios entre sí:

- El desarrollo de herramientas informáticas y bases de datos, y
- La aplicación de estas en la generación de conocimientos biológicos para comprender mejor los sistemas vivos.

El desarrollo de herramientas incluye el software de grabación de secuencias, el análisis estructural y funcional de estas, así como la construcción y la conservación de bases de datos biológicas.

El análisis de los datos biológicos a menudo genera nuevos problemas y desafíos que a su vez estimulan el desarrollo de mejores herramientas computacionales.

B. ¿Cómo se puede aplicar la Bioinformática?

La Bioinformática no sólo se ha convertido en una ciencia esencial para la genómica básica y la investigación en biología molecular, también está teniendo un gran impacto en muchas áreas de la biotecnología y las ciencias biomédicas. Tiene aplicaciones, que están basadas por ejemplo, en los conocimientos de diseño de fármacos, análisis forense de ADN y Biotecnología agrícola.

Un enfoque basado en la bioinformática reduce significativamente el tiempo y el costo requerido para desarrollar medicamentos con mayor potencia, con menos efectos secundarios, y una menor toxicidad que el uso del tradicional ensayo y error.

En medicina forense, los resultados de los análisis filogenéticos moleculares han sido aceptados como pruebas en los tribunales penales. Alguna estadística bayesiana sofisticada y basada en la verosimilitud de los métodos de análisis de ADN se han aplicado en el análisis forense de la identidad.

Vale la pena mencionar que la genómica y la bioinformática están a punto de revolucionar los sistemas de salud mediante el desarrollo de la medicina personalizada. La secuencia genómica de alta velocidad junto con la tecnología informática sofisticada permitirá por ejemplo que un médico en una clínica, pueda secuenciar el ADN de un paciente de forma rápida, y detectar así posibles mutaciones dañinas del genoma, convirtiéndose en protagonista para participar en el diagnóstico precoz y el tratamiento eficaz de enfermedades.

Herramientas bioinformáticas se están utilizando también en agricultura. Las bases de datos del genoma de plantas y análisis de expresión génica de este perfil han desempeñado un papel importante en el desarrollo de nuevas variedades de cultivos que tienen una mayor productividad y más resistencia a las enfermedades.

En concreto, la bioinformática abarca el desarrollo de bases de datos o de conocimiento para almacenar y recuperar datos biológicos, algoritmos para analizar y determinar sus relaciones con los datos biológicos, y las herramientas estadísticas para identificar e interpretar conjuntos de datos.

III. ANÁLISIS DE SECUENCIAS DE ADN

El análisis de la secuencia de ADN, es el descubrimiento de similitudes funcionales y estructurales, y las diferencias entre múltiples secuencias biológicas. Esto puede hacerse comparando las nuevas (desconocidas) con las bien-estudiadas y anotadas (conocidas) secuencias.

Este análisis incluye la alineación de secuencias, la búsqueda en la base de datos de secuencias, el descubrimiento de patrones, la reconstrucción de las relaciones evolutivas, y la formación y la comparación del genoma.

Los científicos han encontrado que dos secuencias similares poseen el mismo papel funcional. La comparación se puede hacer desde aspectos de comportamiento bioquímico o de acuerdo a la estructura de la proteína. Si hay dos secuencias de diferentes organismos son similares, se dice que son *secuencias homólogas*.[1]

A. Alineación de secuencias de ADN.

La comparación de la secuencia de ADN está centro del análisis bioinformático. Se trata de un importante primer paso hacia el análisis estructural y funcional de las secuencias recientemente determinadas.

El proceso más fundamental en este tipo de comparación es la *alineación de secuencias*. Este es el proceso por el cual, se comparan las secuencias mediante la búsqueda de patrones de caracteres comunes y el establecimiento de los residuos de correspondencia entre las secuencias relacionadas. El alineamiento de pares de secuencias es fundamental en la búsqueda de similitudes dentro de la base de datos y el alineamiento de secuencias múltiples.[2]

Un concepto importante en el análisis de la secuencia es una *homología de secuencia*. Cuando dos secuencias son descendientes de un origen evolutivo común, se dice que tienen una relación homóloga u homología.

Un término relacionado, pero diferente es la *similitud de secuencias* (estos términos suelen utilizarse indistintamente de manera errónea), y corresponde al porcentaje de residuos alineados, similares en propiedades físico-químicas tales como el tamaño, el costo, y la hidrofobicidad.

IV. TECNOLOGIAS COMPUTACIONALES APLICADAS A LA BIOINFORMATICA.

La biología al igual que todas las ciencias que son base de la investigación científica, proveen (dependiendo de los objetivos planteados) grandes volúmenes de información que requieren de técnicas computacionales avanzadas para permitir hacer procesamiento en tiempo real.

Muchas de estas técnicas se enmarcan dentro de temas de investigación y desarrollo informático que tienen que ver con el almacenamiento y procesamiento de datos, entre las cuales podemos mencionar las bases de datos (BD) relacionales y semánticas, las bodegas de datos, minería de datos y algunas técnicas de inteligencia artificial, entre otras.

A. Bases de Datos Biológicas.

Las actuales bases de datos biológicas usan generalmente tres tipos de estructuras de base de datos: ficheros planos (a pesar de las obvias desventajas de su uso), relacionales y orientados a objetos. La razón es la falta de dimensionamiento de los modelos reales con el volumen de datos requeridos.

Con base en su contenido, las bases de datos biológicas se pueden dividir en tres categorías :
- *Bases de datos primarias,* las cuales contienen datos biológicos originales. Son archivos de secuencia en bruto o datos estructurales (por ejemplo *GenBankm* y *Protein Data Bank).*
- *Bases de datos secundarias* que contienen información procesada computacionalmente (o manualmente curada), con base en datos primarios. Las bases de datos de secuencias de proteínas traducidas contienen la anotación funcional perteneciente a esta categoría (ejemplo *Swiss-Prot* y *PIR).*
- *Bases de datos especializadas,* aquellas que atienden a un interés de investigación en particular (por ejemplo *Flybase).* La base de datos de secuencias del VIH, y *Ribosomal Database Project* son ejemplos de bases de datos que se especializan en un determinado organismo o un determinado tipo de datos.

Muchos de los problemas detectados en las investigaciones científicas, radican en la necesidad de conectar las bases de datos secundarias y especializadas a las bases de datos primarias. Es conveniente que las entradas en una base de datos sean de referencia cruzada y vinculadas o ¨linkeadas¨ a las entradas relacionadas en otras bases de datos que contengan información adicional.

La barrera principal al enlazar diversas bases de datos biológicas es la incompatibilidad del formato que actualmente utilizan los tres tipos de estructuras de base de datos mencionadas, limitando la comunicación entre ellas.

Una solución para estandarizar la comunicación entre bases de datos en sistemas distribuidos, es el uso de un lenguaje de especificación llamado *Common Object Request Broker Architecture (CORBA)*, que permite a los programas de bases de datos en diferentes lugares comunicarse en una red a través de un "corredor de interfaz" sin tener que entender cada estructura de manera independiente.

Un protocolo similar llamado *eXtensible Markup Language (XML)* también ayuda en el enlace de las bases de datos. En este formato, cada registro biológico se divide en pequeños componentes básicos que se marcan con etiquetas de agrupamiento jerárquico.

Las secuencias de genes se pueden contaminar con secuencias de vectores de clonación y por redundancia de datos primarios causada por la adquisición repetida de secuencias idénticas o coincidentes. Para reducir esta redundancia, el *National Center for Biotechnology Information* (*NCBI)* ha creado una base de datos no redundante llamada *RefSeq*, en el que las secuencias idénticas del mismo organismo y los fragmentos de secuencia asociadas se fusionan en una sola entrada.

Otra manera de abordar el problema de la redundancia es crear las bases de datos de secuencia-*cluster* tales como *UniGene* que unen secuencias de etiquetas expresadas (*EST)* que son derivadas del mismo gene.

A menudo, la secuencia del gen se puede encontrar bajo diferentes nombres como resultado de múltiples entradas. Para aliviar este problema, es necesaria la re-anotación de genes y

proteínas utilizando un vocabulario común para describirlos. *Gene Ontology* proporciona un sistema coherente e inequívoco de nomenclatura para todos los genes y las proteínas.

B. Bodegas de Datos.

Un *Data Warehouse (DW)* es un conjunto de datos integrados orientados a una materia, que varían con el tiempo y que no son transitorios, los cuales soportan el proceso de toma de decisiones de la administración [3].

A partir de la revisión de los proyectos de Bioinformática se encuentra que los requerimientos de este campo exigen el almacenamiento de grandes volúmenes de datos, con múltiples dimensiones, de periodos de tiempo extensos y con formatos heterogéneos al igual que sus fuentes.

Wang *et al*[4] describe su propuesta de modelamiento multidimensional para datos biomédicos, basados en una bodega de datos llamado esquema *BioStar,* que puede capturar la rica semántica de datos biomédicos y proporcionar una mayor extensibilidad y flexibilidad para la rápida evolución de las metodologías de investigación biológica. Esto se garantiza con el almacenamiento de las diferentes medidas en n-tablas separadas, las cuales son usadas para manejar las relaciones de muchos-a-muchos entre la entidad central y las dimensiones pudiendo estar diseñados para soportar características específicas de una medida.

Ligand Depot es una fuente de datos integrados para encontrar información acerca de moléculas pequeñas, proteínas y ácidos nucleicos. Se centra en proporcionar información química y estructural para pequeñas moléculas. A su vez acepta consultas basadas en palabras clave, también proporciona una interfaz gráfica para la realización de búsquedas en subestructura química, y permite el acceso a una amplia variedad de recursos Web. Se plantea como trabajo futuro la implementación de capacidades mejoradas de búsqueda y la incorporación de una más sofisticada interfaz gráfica de usuario [5].

C. Minería de Datos en Bioinformática.

La minería de datos se orienta hacia el estudio de técnicas para extraer información valiosa de una gran cantidad de datos biológicos. Para ello, son necesarias herramientas de software eficientes que permitan recuperar datos, comparar secuencias biológicas, descubrir patrones y visualizar el descubrimiento del conocimiento.[6]

Entre las técnicas de minería de datos en Bioinformática más comunes se pueden destacar:
- *KDD,* que es el proceso completo de extracción de conocimientos, no triviales, previamente desconocidos y potencialmente útiles a partir de un conjunto de datos;
- *minería textual o KDT,* que se orienta a la extracción de conocimiento a partir de datos (no-estructurados en lenguaje natural) almacenados en las bases de datos textuales, se identifica con el descubrimiento de conocimiento en los textos, y
- *Estadistica en la minería de datos,* que se puede dividir en dos grupos: *Aprendizaje supervisado* en el que se tiene conocimiento por adelantado de los grupos de secuencias y en el que el objetivo es deducir la forma de clasificar las futuras observaciones, y el *Aprendizaje no supervisado,* el cual consiste en la detección previa de los grupos hasta ahora desconocidos de casos "similares" en los datos.

Las herramientas de software que facilitan la investigación en bioinformática pueden clasificarse en cuatro clases:
a. *Herramientas de recuperación de datos*. Por ejemplo, *Entrez*[7], que es un sistema integrado de datos de recuperación desarrollado por la *NCBI* que proporciona un acceso integrado a una amplia gama de dominios de datos, incluyendo secuencias de la literatura, nucleótidos y proteínas, genomas completos, estructuras 3D y más.

b. *Comparación de la secuencia y las herramientas de alineación.* Un ejemplo es *BLAST* (su principal característica es la velocidad), que realiza búsquedas en la totalidad de una base de datos no redundante en poco tiempo.

GenBank y *EMBL*, son dos de las herramientas principales de gestión de bases de datos biológicas para alineamiento local por pares de secuancias.

FASTA se puede utilizar para hacer una comparación rápida de proteínas o de nucleótidos. Alcanza un alto nivel de sensibilidad para la búsqueda de similitud mediante la realización de búsquedas optimizadas para alineamientos locales utilizando una matriz de sustitución.

Para alineación de secuencias múltiples, la herramienta disponible es *ClustalW*, la cual se puede utilizar para alinear las secuencias de ADN o de proteínas con el fin de dilucidar sus relaciones, así como su origen evolutivo.

c. *Herramientas de descubrimiento de patrones*, que se utilizan para buscar patrones o características de los datos. *Análisis de Cluster* es una herramienta que se utiliza para encontrar grupos en un determinado conjunto de datos de tal manera que los objetos en el mismo grupo sean similares entre sí y diferentes a los de otros grupos.

Otra herramienta útil integrada para el descubrimiento de patrones de expresión es *GeneQuiz*, como un sistema integrado de gran escala para el análisis de secuencias de ADN y proteínas, usando una variedad de métodos de búsqueda y análisis.

d. *Herramientas de visualización.* Permiten una visualización interactiva y gráfica de los datos genómicos.

Los más grandes paquetes de análisis, tales como *Expression Profiler y GeneQuiz*, tienen una herramienta de visualización integrada en ellos. Además, muchos paquetes de software de visualización también se encuentran disponibles gratuitamente en Internet.

Algunos ejemplos son los siguientes: *Protein Explorer* que proporciona una visualización en 3D de la estructura de proteínas en un sistema interactivo *y TreeView* que proporciona una representación gráfica de los resultados de la agrupación y la imagen de navegación basada en los árboles jerárquicos.

D. Máquinas de Aprendizaje.

Una máquina de aprendizaje es un proceso adaptativo que permite a las computadoras aprender de la experiencia, aprender con el ejemplo, y aprender por analogía. *La red neuronal* es una de las máquinas de varios enfoques de aprendizaje que se han aplicado con éxito a la solución de una amplia variedad de problemas bioinformáticos. Por ejemplo, un sistema de red neuronal basado en el conocimiento neuronal fue aplicado al análisis de la secuencia de ADN[8].
Existen dos buscadores de genes más populares que dieron lugar a las Redes Neuronales Artificiales. *GRAIL*[9] es el primer programa buscador de genes, que fue diseñado para identificar genes, exones, y varias características en las secuencias de ADN; éste utiliza una red neural que combina una serie de algoritmos de codificación de predicción para reconocer el potencial de codificación en ventanas de longitud fija sin buscar características adicionales.

Otro sistema de buscador de genes es *Gene Parser*, que fue diseñado para identificar y determinar la fina estructura de los genes de la proteína en las secuencias de ADN genómico.

Un sistema neural artificial para clasificación de genes llamado *GenCANS* fue desarrollado para analizar y gestionar un gran volumen de datos de secuenciación molecular del Proyecto del Genoma Humano.

El algoritmo genético ha sido aplicado con éxito para resolver muchos problemas prácticos en muchas disciplinas, en particular, en la bioinformática, estos se han utilizado para resolver los problemas de alineación de secuencias múltiples.

Un enfoque bien conocido es *SAGA*, el cual crea aleatoriamente una población inicial de alineaciones y evoluciona a través de generaciones, mejorando gradualmente el *fitness* de la población.

E. Soft Computing en Bioinformática.

Una de las técnicas más usadas de *Soft computing* son los S*istemas expertos,* los cuales se construyen mediante la recopilación de conocimientos de expertos humanos específicos. A menudo es difícil para los expertos a decir cuáles son las reglas que utilizan. Los problemas se resuelven extrayendo la descripción de la situación oculta, en términos de los factores y las normas que coinciden con el comportamiento del experto humano.

Con los avances en la biotecnología, se generan enormes volúmenes de datos biológicos. Además, es posible que existan importantes relaciones ocultas y correlaciones en los datos. Algunos métodos de *Soft computing* están diseñados para manejar grandes conjuntos de datos, y también pueden ser utilizados para extraer este tipo de relaciones [10].

En bioinformática, los *sistemas difusos* juegan un papel importante para la construcción de sistemas basados en el conocimiento. Hay muchas áreas de aplicación de la ciencia biomédica y la bioinformática, donde las técnicas de lógica difusa pueden ser aplicadas con éxito. Algunas de las aplicaciones importantes de la lógica difusa son las siguientes: aumentar la flexibilidad de los motivos de proteínas, estudiar las diferencias entre polinucleótidos [11], analizar los datos experimentales de expresión [12] utilizando la teoría difusa de resonancia adaptativa, alinear las secuencias basadas en una difusa refundición de un algoritmo de programación dinámica[13], la secuenciación del ADN genético utilizando sistemas difusos[14], analizar los datos de expresión génica[15], analizar las relaciones entre los genes y descifrar una red genética[16], y clasificar las secuencias de aminoácidos en diferentes super familias[17].

F. *MATLAB* aplicado a la Bioinformatica.

MATLAB es el nombre abreviado de "*MATrix LABoratory*". Es un entorno de computación y desarrollo de aplicaciones totalmente integrado orientado para llevar a cabo proyectos en donde se encuentren implicados elevados cálculos matemáticos y la visualización gráfica de los mismos.

Dispone también en la actualidad de un amplio abanico de programas de apoyo especializado, denominados *Toolboxes. Bioinformatics* que los biólogos moleculares y a otros investigadores científicos un entorno abierto y extensible, en el cual pueden explorar ideas, hacer prototipos de nuevos algoritmos, y construir aplicaciones en investigación de drogas, ingeniería genética, y otros proyectos genómicos y proteómicos. *Toolbox* provee acceso a formatos de datos genómicos y proteomicos, técnicas de análisis y visualizaciones especializadas para secuencias genómicas y proteomicas y análisis de *microarrays*.

V. CONCLUSIONES Y RECOMENDACIONES

El almacenamiento de datos aparece en la bioinformática para apoyar el descubrimiento de los conocimientos biológicos y

también para facilitar la investigación y el intercambio de información.

Se considera que las aplicaciones web biológicas colaborativas han revolucionado la investigación biológica.

Es indispensable el acompañamiento del profesional en informática para que el científico haga un uso más correcto y pueda aprovechar al máximo todas las características técnicas de las BD.

Muchas de las investigaciones recientes probablemente sean en las ciencias biológicas y de la salud, por lo que se busca que se busca un enfoque investigativo desde la informática hacia esta área.

Como trabajo futuro se requiere definir las plataformas tecnológicas y la implementación de los procesos asociados a la solución propuesta. No obstante en esta fase los investigadores ven representados sus intereses y requerimientos lo que es condición fundamental para el éxito del sistema planteado.

REFERENCIAS

[1] PHOEBE CHEN, Yi-Ping. *Bioinformatics Technologies*. Alemania: *Springer-Verlag* Berlin Heidelberg, 2005. 396p. ISBN 3-540-20873-9.

[2] XION, Jin. *Essential Bioinformatics*. Estados Unidos de América: *Cambridge University Press*, 2006. 331p. ISBN 978-0-511-16815-4.

[3] HARJINDER S, Gill y PRAKASH C, Rao. *Data Warehousing*. La Integracion de Informacion para la Mejor Toma de Decisiones. México: Prentice Hall, 1996. 382p. ISBN 968-880-792-3.

[4] *BioStar models of clinical and genomic data for biomedical data warehouse design* [en línea]. WANG, Liangjiang; RAMANATHAN, Murali y ZHANG, Aidong. *State University of New York at Buffalo:* New York, Estados Unidos de América, 2005 - [citado el 30 de marzo de 2011]. Disponible desde Internet en: <http://www.cse.buffalo.edu/DBGROUP/bioinformatics/papers/ijbra05.pdf>

[5] FENG, Zukang, *et al. Ligand Depot: a data warehouse for ligands bound to macromolecules.* En: *Bioinformatics Applications Note* [en línea]. 1 de abril de 2004. vol. 20. no. 13. Disponible desde Internet en: <http://bioinformatics.oxfordjournals.org/content/20/13/2153.full.pdf+html?sid=5fbc13fd-7bee-4364-829b-ef27e2d53032>

[6] PHOEBE CHEN, Yi-Ping. *Bioinformatics Technologies*. Alemania: *Springer-Verlag* Berlin Heidelberg, 2005. 396p. ISBN 3-540-20873-9.

[7] GEER, Renata C. y SAYERS, Eric W. *Entrez: Making use of its power*. En: *Briefings in Bioinformatics.* vol. 4, no. 2. Junio, 2003. p. 179.

[8] FU, Limin. *Knowledge Discovery Based on Neural Networks.* En: *Communications of the ACM (CACM).* vol. 42, Issue: 11, Noviembre 1999. p. 47-50.

[9] UBERBACHER, Edward y Mural, Richard. *Locating Protein Coding Regions in Human DNA Sequences Using a Multiple Sensor-Neural Network Approach.* En: *Proceedings of the National Academy of Sciences of United States of America*. vol. 88, Diciembre de 1991. p. 11261-11265.

[10] JENA, Rabindra Ku., *et al. Soft computing Methodologies in Bioinformatics.* En: *European Journal of Scientific Research.* vol 26, no.2. 2009. p. 193

[11] TORRES, Angela y NIETO, Juan. *The Fuzzy polynucleotide space: basic properties.* En: *Bioinformatics.* vol. 19, Issue: 5. 2003. p. 92

[12] TOMIDA, Shutta, *et al. Analysis of expression profile using fuzzy adaptive resonance theory*. En: *Bioinformatics.* vol. 18, Issue: 8. 2002. p.1073-1083

[13] SCHLOSSHAUER, Maximilian y OHLSSON, Mattias. *A novel approach to local reliability of sequence alignments*. En: *Bioinformatics.* vol 18, no.6. 2002. p. 847-854.

[14] CORDÓN, Oscar, *et al. Ten years of genetic fuzzy systems.* En: *Fuzzy Sets and Systems*. vol. 141, Issue: 1. 2004. p. 5-31.

[15] WOOLF, Peter y WANG, Yixing. *A fuzzy logic approach to analyzing gene expression data.* En: *Physiological Genomics*. vol.3, Issue: 1. 2000. p. 9-15.

[16] RESSOM, H.; REYNOLDS R. y VARGHESE R. *Increasing the efficiency of fuzzy logic based gene expression data analysis.* En: *Physiological Genomics*. vol. 13, Issue: 2. 2003. p. 107–117.

[17] BANDYOPADHYAY, Sanghamitra. *An efficient technique for super family classification of amino acid sequences: feature extraction, fuzzy clustering and prototype selection.* En: *Journal Fuzzy Sets and Systems.* vol. 152, Issue: 1. 2005. p. 5–16.