ALINEACIÓN GLOBAL DE PARES DE SECUENCIAS DE BIOMOLECULAS

UNIVERSIDAD DE GUADALAJARA

CENTRO UNIVERSITARIO DE CIENCIAS EXACTAS E INGENIERIA (CUCEI).

Por: Karla Avila Cárdenas

Asesor: Dr. Israel Román G.

RESUMEN.

Uno de los procesamiento más comunes en bioinformática es determinar la relación de una secuencia nucleótidos o aminoácidos (en adelante, biomoléculas) con otra diferente. Esa similitud entre secuencias de biomoléculas, indica si los organismos han evolucionado de un ancestro común, es decir; son homólogas, si tienen funciones comunes o (en el caso de las proteínas) si tienen formas similares [1]. Existen dos tipos de alineamiento de secuencias: alineación global y local [2].

Este trabajo se enfocará, particularmente, en el alineamiento global. Se desarrollará una Interfaz de programación de aplicaciones (API, por sus siglas en inglés) basada en el algoritmo de Needleman – Wunsch [3] y el paradigma de programación orientado a objetos, utilizando C++. De modo que permita a los bioinformáticos, estudiantes y especialistas del área de biología molecular trabajar componentes de software para el análisis de secuencias global de biomoléculas sin importar si cuentan con conocimientos de programación, para lo cual se desarrollará una interfaz gráfica de usuario (GUI) que utilice el API antes mencionada.

ANTECEDENTES.

La bioinformática es, por sus características, una disciplina integradora que surge del encuentro o reunión de aquellos proyectos en los que un sistema biológico es analizado, moldeado o simulado, aplicando teorías, métodos y procedimientos informáticos [4]. Existen diferentes métodos aplicados para la alineación global de secuencias por ejemplo: método heurístico de palabra corta, matrices de puntuación (matrices de puntuación para aminoácidos, matrices PAM, matrices BLOSUM) y programación dinámica [5].

El algoritmo alineamiento global de secuencias fue propuesto en 1970 por Saul Needleman y Christian Wunch. Consiste en crear una matriz de puntuación de dos dimensiones para el alineamiento de secuencias, en cuyos ejes están las dos secuencias a comparar. Las puntuaciones se calculan fila por fila de forma que el escaneo de la segunda fila tiene en cuenta los resultados obtenidos en por la primera fila: el mejor resultado de ésta se coloca en la esquina inferior derecha (diagonal) de la matriz interna. Así pues, el mejor camino de alineación es el que tenga la puntuación máxima total. El camino también puede moverse horizontalmente o verticalmente en un punto determinado que corresponde a la introducción de un hueco, inserción o borrado de un nucleotido o aminoácido en una de las dos secuencias. Existen distintos tipos de sistemas de puntuación, los cuales dependen de diferentes modelos para la inserción, borrado e inserción de los nucleotidos o aminoácidos en las secuencias. El algoritmo funciona del mismo modo independientemente de la complejidad o longitud de las secuencias y garantiza la obtención del mejor alineamiento [3] [5].

La información de las secuencias de ADN y aminoácidos es, generalmente, guardados en archivos. Dos de los formatos mas usado son GenBank y FASTA. GenBank es una base de datos que almacena información de diferentes secuencias de biomoléculas. Contiene cerca de 15.000 entradas, entre secuencias de ADN y aminoácidos, que han sido reportados desde 1967 [6], [7]. La información en dicha base de datos puede ser accedida a través de la internet y descargada en alguno de los dos formatos antes mencionados. El formato GenBank tiene la ventaja de brindar información y metainformación de cada una de las secuencias almacenadas. Por otro lado, los archivos FASTA, mantiene las secuencias en un formato plano, con una cabecera y la secuencia determinada [8].

C++ es un lenguaje de programación de propósito general. Una de sus principales cualidades es la reutilización de código en forma de librerías de usuario. Con la excepción del ensamblador, genera los programas más compactos y rápidos. Otra de sus características es que es un lenguaje portable ya que la mayor parte de compiladores están respaldados por enormes librerías escritas en C/C++ [9] [10].

Por otro lado, Eclipse es un conjunto de herramientas de programación utilizado para la construcción de aplicaciones. El equipo de desarrollo de Eclipse, desarrolló Eclipse 3.0 con el que introdujo el RCP, para facilitar la creación de aplicaciones de cliente enriquecido [11], de modo que también nos apoyaremos de esta herramienta en este proyecto.

JUSTIFICACIÓN.

Este proyecto se lleva a cabo con la finalidad de integrar un entorno de trabajo en dos niveles diferentes: 1) nivel *intermedio* orientado a desarrolladores de aplicaciones del área de bioinformática; y 2) nivel *alto* una interfaz gráfica de usuario, para especialistas y estudiantes del campo de la biología molecular, para que tengan una herramienta (i.e. algoritmo de alineación global) que faciliten su trabajo. Además, dicha interfaz gráfica tendrá la facilidad de ampliarse a través de componentes de software conocidos como plugins y permitirá la lectura de archivos FASTA y GenBank.

OBJETIVOS Y METAS:

* OBJETIVO GENERAL.

Brindar a la comunidad, tanto de desarrolladores como de estudiantes y profesionistas de áreas relacionadas con la biología molecular, un entorno de trabajo en dos niveles diferentes con el que puedan realizar alineamientos globales de secuencias de biomoléculas y lectura de archivos FASTA y GenBank con la facilidad de ampliarse a través de componentes de software conocidos como plugins.

* OBJETIVOS ESPECIFICOS.
* Desarrollar una API en lenguaje C++ implementando el algoritmo de Needleman – Wunsch por medio de programación dinámica.
* Desarrollar una Interfaz Gráfica de Usuario (GUI) con Eclipse RCP.
* Desarrollar una API para la lectura de archivos FASTA y GenBank.
* META:

Interfaz de usuario y librería de alineación global de secuencias de biomoléculas junto con lectura de archivos FASTA y GenBank.

METODOLOGIA.

Este proyecto se llevará a cabo en base a las siguientes etapas:

1. Investigación sobre las características de biomoléculas y su funcionabilidad en los organismos.
2. Investigación sobre la bioinformática y el papel de la alineación de secuencias de biomoléculas para las áreas biológicas.
3. Investigación sobre los principales métodos de alineación global de biomoléculas.
4. Investigación y asimilación del algoritmo de Needleman – Wunsch.
5. Investigación sobre la composición de archivos FASTA y GenBank.
6. Investigación sobre API´s y GUI (Diseño), así como su implementación.
7. Investigación sobre las herramientas a utilizar: Elipse RCP y C++.
8. Levantamiento de requerimientos.
9. Diseño de la API para la lectura de archivos FASTA y GenBank.
10. Implementación de la API anterior en lenguaje C++.
11. Diseño de la API de alineación global de pares de secuencias de biomoléculas.
12. Implementación de la API en lenguaje C++.
13. Diseño de la GUI.
14. Implementación de la GUI en Eclipse RCP.

PROTOTIPO:

Al final de este proyecto se obtendrá: un entorno de trabajo en dos niveles diferentes: 1) nivel intermedio orientado a desarrolladores de aplicaciones del área de bioinformática; y 2) nivel alto una interfaz gráfica de usuario, para especialistas y estudiantes del campo de la biología molecular, para que tengan una herramienta (i.e. algoritmo de alineación global) que faciliten su trabajo. Dicha interfaz gráfica tendrá la facilidad de ampliarse a través de componentes de software conocidos como plugins y permitirá la lectura de archivos FASTA y GenBank.

REFERENCIAS.

[1] Perezleo Solórzano, L., Arencibia Jorge, R., Conill González, C., Achón Veloz, G., & Araújo Ruiz, J. A. (2003). Impacto de la bioinformática en las ciencias biomédicas. *Acimed*, *11*(4), 0-0.

[2] Zimic, M. Alineamiento Simple (global y local) de secuencias.

[3] Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, *48*(3), 443-453.

[4] Lahoz-Beltrá, R. (2010). *Bioinformática: Simulación, vida artificial e inteligencia artificial*. Ediciones Díaz de Santos.

[5] Sánchez Becerro, Ó. (2014). Algoritmos de programación dinámica con R para resolver problemas de alineamiento de secuencias.

[6] Bilofsky, H. S., & Christian, B. (1988). The GenBank® genetic sequence data bank. *Nucleic acids research*, *16*(5), 1861-1863.

[7] Fourment, M., & Gillings, M. R. Una comparación de lenguajes de programación usados en bioinformática.

[8] Zaliz, R. C. R. Extracción de conocimiento en repositorios biológicos estructurados.

[9] Stroustrup, B. (1986). *The C++ programming language*. Pearson Education India.

[10] Borrás, J. L. L. (2001). *Curso práctico de programación en C y C++* (No. 8). Universitat Jaume I.

[11] Rubel, D. (2006). The heart of eclipse. *Queue*, *4*(8), 36-44.