ALINEACIÓN GLOBAL DE DOS SECUENCIAS DE ADN POR MEDIO DE PROGRAMACIÓN DINÁMICA

UNIVERSIDAD DE GUADALAJARA

CENTRO UNIVERSITARIO DE CIENCIAS EXACTAS E INGENIERIA (CUCEI).

Por: Karla Avila Cárdenas

Asesor: Dr. Israel Román G.

RESUMEN.

El procesamiento más común en bioinformática es el obtener información de las cadenas de ADN que conforman a los organismos ya que la alineación de secuencias puede determinarse si una secuencia de nucleótidos o aminoácidos está relacionada con otra. Esa similitud entre secuencias, 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.

Existen dos tipos de alineación de secuencias de ADN: alineación global y local. Este trabajo se enfocará en la alineación global que es aquella donde las secuencias se alinean a lo largo de toda su longitud lo cual tiene una gran funcionabilidad para secuencias de longitudes. En la actualidad existen distintos métodos para su solución. Nosotros nos enfocaremos en el método de programación dinámica que se basa en el principio de optimalidad de Bellman donde el problema es dividido en subproblemas de tamaños menores que son fáciles de resolver.

ANTECEDENTES.

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++.

Eclipse fue utilizado para la construcción de aplicaciones fuera del dominio para el que Eclipse fue diseñado originalmente. Aprovechándose de esa tendencia, el equipo de desarrollo de Eclipse desarrollo Eclipse 3.0 con el que introdujo el RCP, para que facilita la creación de aplicaciones de cliente enriquecido.

Este proyecto aprovechara las herramientas antes mencionadas para la implementación el algoritmo de Needleman - Wunch para la alineación global de secuencias de ADN. Este algoritmo fue propuesto en 1970 por Saul Needleman y Christian Wunch consiste en crear una matriz de puntuación para el alineamiento de dos dimensiones 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 dilección en una de las dos secuencias. Existen distintos tipos de sistemas de puntuación para este caso en específico se utilizará un sistema de puntuación básico donde se le otorgará: 1 para iguales, 0 para distintos y -1 para penetraciones. El algoritmo funciona del mismo modo independientemente de la complejidad o longitud de las secuencias y garantiza la obtención del mejor alineamiento.

JUSTIFICACIÓN.

El proyecto brindará un entorno agradable para todos aquellos profesionistas o estudiantes que tengan la necesidad de realizar una alineación global de secuencias de ADN junto con todos los beneficios que el algoritmo otorga. De modo que sirva como apoyo para cualquier proyecto de bioinformática o áreas similares. Además de que sea fácil para cualquier usuario el manejo de cadenas de ADN, sin importan su longitud.

OBJETIVOS Y METAS.

* OBJETIVO GENERAL.

Brindar a la comunidad una solución eficiente y fácil de manejar para la alineación global de secuencias de ADN

* OBJETIVOS ESPECIFICOS.

Implementar la solución en lenguaje C++ a modo de librería e Interfaz de usuario por medio de Eclipse RCP.

* PRODUCTO.

Interfaz de usuario y librería de alineación global de secuencias de ADN.

METODOLOGIA.

La implementación de este proyecto será realizado en dos fases principales:

* La primer fase cubrirá el primer proyecto modular (Arquitectura y programación de sistemas), donde primeramente se estudiarán a profundidad los distintos algoritmos para la alineación global de secuencias para seleccionar el algoritmo con mayores beneficios para ser implementado. Luego, se diseñará el sistema a desarrollar en lenguaje c++ del algoritmo y la interfaz que conformarán nuestra API, junto con su documentación correspondiente. Seguidamente se realizará la codificción del algoritmo.

Cabe mencionar que nos apoyaremos de un sistema de control de versiones (github) de modo que nuestra API será de código abierto disponible en todo momento para toda la comunidad.

* La segunda fase cubrirá el segundo proyecto modular (Sistemas distribuidos) En el cual se planea optimización del algoritmo generado en la primera fase medio de programación paralela en clúster.

REFERENCIAS.

-”Material Suplementario 1: Algoritmos de Programación

Dinámica para alns. Globales y locales”, Pablo Vinuesa 2007.

-”Algoritmos de programación dinámica con R para resolver problemas de alineamiento de secuencias.”, Óscar Sánchez Becerro 2014.

-”Alineamiento: Análisis computacional de secuencias”, Andrés M. Pizón. Centro de bioinformática del instituto de biotecnología de la Universidad Nacional de Colombia.

-”Alineamiento de pares de secuencias”, Rodrigo Santamaría.

-”Curso de C++”, Salvador Pozo Coronado 2003.

-”Developing Eclipse Rich Client Applications”, Dr. Frank Gerhardt y Dr. Christian Wege 2005.