Interfaz de programación de aplicaciones (API) e Interfaz gráfica (GUI) para algoritmos de String­Matching.

Resumen:

Una de la tareas más frecuentes en la biología molecular es el análisis de secuencias (ácido desoxirribonucleico y/o aminoácidos). El fin de dicho análisis es encontrar características, intrínsecas en las secuencias, que permitan determinar si dicha secuencia tiene alguna función biológica o no. Dicho análisis puede ser abordado por un área de la ciencias computacionales conocido como Bioinformática.

El presente proyecto modular tiene como objetivo desarrollar una API, utilizando el paradigma orientada a objetos, que permita a los bioinformáticos construir componentes de software para el análisis de secuencias, utilizando algoritmos de string-matching. Además, con el fin de que esta herramienta sea utilizada por profesionales del área de biología molecular (sin conocimientos de programación), se desarrollará una interfaz gráfica de usuario (GUI) que utilice el API antes mencionada.

Antecedentes:

La bioinformática es la ciencia que se encarga de aplicar métodos computacionales para hacer descubrimientos biológicos. Se trata de un campo interdisciplinario que engloba la biología, ciencias de la computación, matemáticas y estadística.

[TesisMaestria.pdf?]

Existen diferentes algoritmos computacionales aplicados para el análisis de secuencias por ejemplo String-matching, entre los que se encuentran alineamiento de secuencias, memorias asociativas, redes neuronales, entre otros, los cuales sirven a diversos propósitos como clasificación de proteínas, identificación de genes, etc.

[Bioinformatics: sequence and genome analysis, TesisMaestria].

String-matching busca encontrar las ocurrencias de un patrón determinado dentro de una secuencia de caracteres, dicha búsqueda puede tener como objetivo encontrar ocurrencias de una cadena específica, lo que se conoce como *exact matching,* o un conjunto de cadenas que cumplan con un criterio determinado, posiblemente mediante expresiones regulares, lo que se conoce como *inexact matching.*

[Biological Sequences and the Exact String Matching

Problem]

**Desde un punto de vista computacional, las secuencias de ADN pueden ser vistas de un modo más abstracto como cadenas de caracteres este enfoque nos permite traducir problemas de bioinformática a problemas de *string-matching.***

Existen diferentes algoritmos para String-matching como Z-algorithm, set matching using keyword trees, suffix trees, suffix arrays que ayudan a resolver problemas como longest common substring, k-mismatches, entre otros.

[el fragmento de string – matching que me pasó]

Nosotros nos enfocaremos en algoritmo1, algoritmo2.

Actualmente existen diferentes API’s para el análisis de secuencias, algunos ejemplos son BioPerl, BioPhyton, BioJava, Bioc++, SeqAn entre otros. Algunos de éstos no fueron modelados utilizando el paradigma orientado a objetos (BioPerl, BioPhyton, BioJava, Bioc++) [Verificar si esto es cierto: parece que no :(]. En los casos antes mencionados, el principal objetivo fue el de mantener un mejor desempeño y eficacia de los algoritmos por sobre la facilidad en la construcción de nuevos componentes de software; la razón de esto último es que el paradigma orientado a objetos implicaría una disminución en el eficiencia de los algoritmos.

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 tenga herramientas (i.e. algoritmos de string-matching) que faciliten su trabajo. Además, dicha interfaz gráfica tendrá la facilidad de ampliarse a través de componentes de software conocidos como plugins.

Prototipo:

Firmas: El documento será firmado al calce por el solicitante y el asesor.