CSE 311 Lab Project

Online sequence alignment for string matching to reference genomes.

Objectives
Create an online tool that allows users search the genomes of various organisms.

Outline
April 1 - Sequence Alignment Revisited
April 8 - HTML basics, forms, Web.py
April 15 - Integrating the algorithms
April 22 - Help Session
April 29 - Project Due/Review for Final

Algorithms
Local Sequence Alignment
Global Sequence Alignment
Rabin-Karp algorithm
BLAST
BLAT, Bowtie

As learned in class, searching, sorting, and matching short sequences are fairly straightforward, given the right algorithms, such as those using dynamic programming (**Figure 1**). However, when operating on large sequences, such as the human genome or genomes of other organisms, with over 3 billion characters, the task becomes much harder. A simple operation of string matching would take exceedingly long. Blast has been designed to address these issues (**Figure 2).**

Next-generation sequencing technology brings more challenges in that we need to align trillions of short sequences (32-150bps) to the reference genome which is extremely time expensive.

For this project, a new algorithm will be introduced that is similar to how modern sequence alignment tools such as BLAT and Bowtie work (**Figure 3**). In addition to implementing this algorithm, you must also develop a web page to allow users to find a matching strand of DNA against various genomes. As this is not a web development course, not much emphasis will be placed on the inner workings of a web server or the interface, but you will gain a very brief introduction into Python web development using web.py.

You may choose any of the implementations listed under algorithms, but make sure your tool will at least support aligning one query sequence (or a set of sequences) to a large genome sequence in an efficient manner.

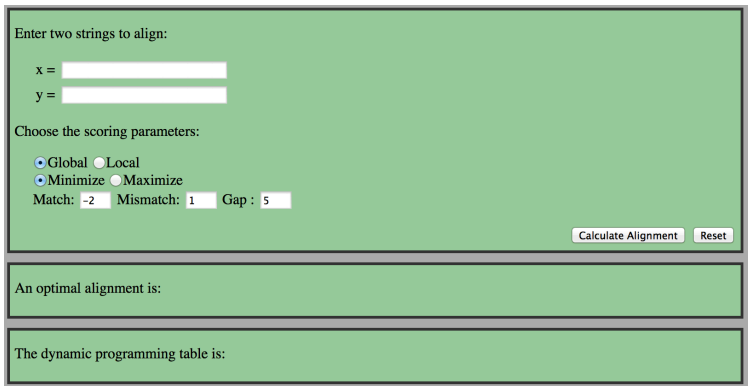

Figure 1. A sequence alignment tool interface which supports both global and local alignment between two given sequences.

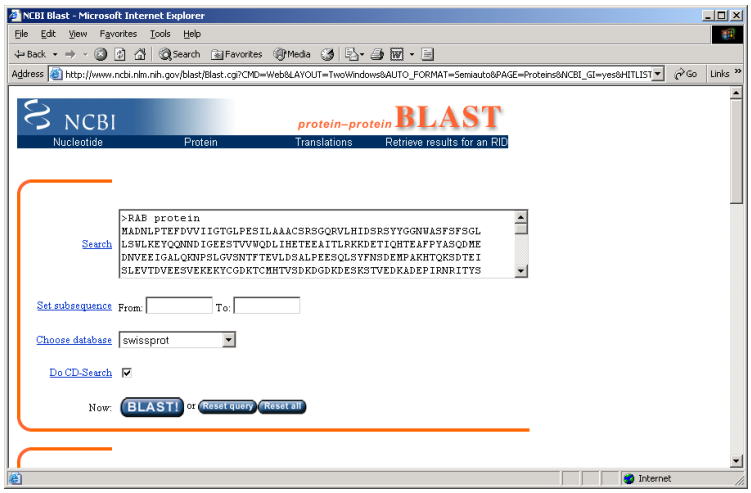

Figure 2. Blast interface

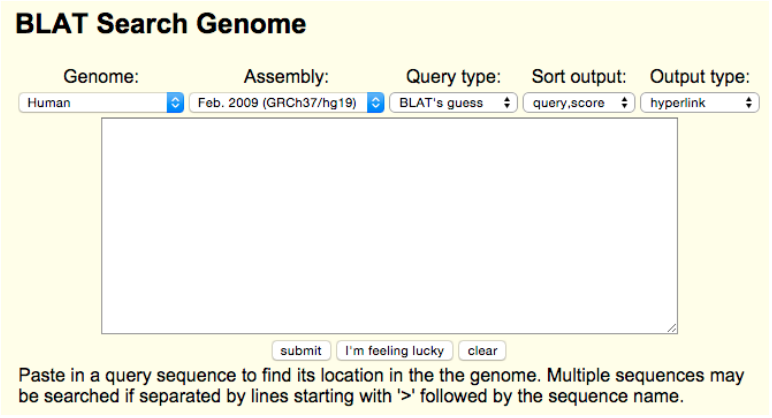

Figure 3: A very basic interface for the UCSC Blat Genome Search