Marcus Bamberger

Bio 131 Final Project

The purpose of this project was to construct a function to determine global alignment with affine gaps. When comparing a raw section of DNA to one that has been spliced to remove introns and exons, it is necessary to recognize that the correct alignment minimizes the number of gaps, not the number of individual deletions. This program incorporates that prioritization into its alignment-finding.

This program, gapAlign, used toy sequences constructed by hand and data obtained from the UCSC genome browser of human chromosome 1. In order to test the influence of the affine gap prioritization, I tested this program against a more basic alignment-finding function from lab six with the same sequences. The toy sequences used were generally a repeated motif or motifs, such as GCAT, repeated with no intervening base pairs in one string – the “short sequence” and junk, random base pairs in the second – the “full sequence”. In this way, I tested whether the alignment would minimize gaps and match motif to motif or align with no regard for the number of gaps. The UCSC genome sequences used were trimmed to a usable length, as aligning an entire chromosome was extremely impractical. One sequence, the “full sequence”, was prepared by capitalizing all letters in the sequence, eliminating the distinction between introns and exons. The second sequence, the “short sequence”, was prepared by removing the introns entirely, creating the same setup as the toy sequences with authentic data.

gapAlign takes as input two strings, to compute an alignment between, and several values, which determine how it scores the alignments. The values are for opening a gap, continuing a gap, a mismatch, and a match. Different values provide different optimal

a

alignments. This is valuable for locating smaller repeated motifs concealed within larger sequences.

The scoring system used to find this particular alignment was as follows: Opening a gap and mismatching two base pairs were both penalized by 1 point. Matching two base pairs rewarded one point. Continuing an already open gap penalized the score by -0.1 points. The