

ECE-S690-501 Genomic Signal Processing: Sequence Alignment and BLAST Homework Assignment 2

Due on:

May 12th, 2009

Introduction

You must turn in your Matlab code in your solutions, or you will receive a 0.

1

Please do the following problem by hand. Chapter 2 of Durbin et al. will be helpful

Calculate the Dynamic Programming matrix and an optimal alignment for the DNA sequences GAATTC and GATTA, scoring +2 for a match, -1 for a mismatch, and with a linear gap penalty of $d = 2$. (Hint: There are two possible solutions).

2

Please do the following problem by hand. Chapter 2 of Durbin et al. will be helpful

Pairwise alignment:

(a) Compute the global (N-W algorithm) and local (Smith-Waterman) alignment of SCIONS with CONSCIENCES, using the following scoring system: matches = 4, mismatches = -3, indels = -3. **Show your alignment matrix with scores and traceback on the attached alignment template.**

(b) **What is the score of the global alignment? What is the score of the optimal local alignment?** There is more than one optimal local alignment. **Show all optimal local alignments.**

3

Inspired from Exercise 3.9 from Cristian:

- a. Obtain two highly similar E. Coli bacterial genomes NC_002655 and NC_002695. Is it possible to take the global alignment of them with Matlab (**show code and results**)? Can we BLAST them (show BLAST results)?
- b. Take two similar genes of E. Coli sequences, GeneID: 4187696 and GeneID: 1038700. Align them as nucleotide sequences (print out the score and alignment). Where is the region of dense dissimilarity?
- c. Now align the Genes as Amino Acid sequences. Do these nucleotide regions of dissimilarity yield a high difference in the amino acid sequence?
- d. Use ClustalW (<ftp://ftp.ebi.ac.uk/pub/software/clustalw2> where clustalx is a graphical interface) to perform multiple alignment of CAA56038.1, CAC80518.1, CAA09227.1, AAL67846.1, and AAM33145.1 .