

# Homework Assignment 2

BENG 202/CSE 282

February 13, 2007

This homework is due February, 20 2007 at 3:30 PM. This is an individual assignment. You can discuss the homework with other people in the class, but make sure you document this. Please hand in a hardcopy on the due date.

All problems were taken from N.C. Jones and P.A. Pevzner. Introduction to Bioinformatics Algorithms. The MIT Press. 2004.

## Problem 6.58 (18 Points)

Introns are spliced out of pre-mRNA during mRNA processing and biologists can perform *cDNA sequencing* that provides the nucleotide sequence complementary to the mRNA. The cDNA, therefore, represents the concatenation of exons of a gene. Consequently the exon-intron structure can be determined by aligning the cDNA against the genomic DNA with the aligned regions representing the exons and the large gaps representing the introns. This alignment can be aided by the knowledge of the conserved donor and acceptor splice site sequences ( **GT** at the 5' splice site and **AG** at the 3' splice site).

While a spliced alignment can be used to solve this *cDNA Alignment* problem there exists a faster algorithm to align cDNA against genomic sequence. One approach is to introduce gap penalties that would adequately account for gaps in the cDNA Alignment problem. When aligning cDNA against genomic sequences we want to allow long internal gaps in the cDNA sequence. In addition, long gaps that respect the consensus sequences at the intron-exon junctions are favored over gaps that do not satisfy this property. Such gaps that exceed a given length threshold and respect the donor and acceptor sites should be assigned a constant penalty. This penalty is lower than the affine penalty for long gaps that do not respect the splice site consensus. The input to the cDNA Alignment problem is genomic sequence  $\mathbf{v}$ , cDNA sequence  $\mathbf{w}$ , match, mismatch, gap opening and gap extension parameters, as well as  $L$  (minimum intron length) and  $\delta_L$  (fixed penalty for gaps longer than  $L$  that respect the consensus sequences). The output is an alignment of  $\mathbf{v}$  and  $\mathbf{w}$  where aligned regions represent putative exons and gaps in  $\mathbf{v}$  represent putative introns

Devise an efficient algorithm for the cDNA Alignment problem.

### Problem 7.6 (18 Points)

Develop a linear-space version of global sequence alignment with affine gap penalties.

### Problem 8.25 (18 Points)

Consider two proteins  $P_1$  and  $P_2$ . The combined prefix spectrum of proteins  $P_1$  and  $P_2$  is defined as the union of their prefix spectra. Describe an algorithm for reconstructing  $P_1$  and  $P_2$  from their combined prefix spectrum. Give an example when such a reconstruction is non-unique. Generalize this algorithm for three and more proteins.

### Problem 9.5 (10 Points)

Write an efficient algorithm that will construct a keyword tree given a list of patterns  $\mathbf{p}^1, \mathbf{p}^2, \dots, \mathbf{p}^k$ .

### Problem 9.13 (18 Points)

Design an efficient algorithm that finds a shortest string in text  $\mathbf{t}_1$  that does not appear in text  $\mathbf{t}_2$ .

### Problem 11.6 (18 Points)

Consider a different game where the dealer is not flipping a coin, but instead rolling a three-sided die with labels 1, 2, and 3. (Try not to think about what a three-sided die might look like.) The dealer has two loaded dice  $D_1$  and  $D_2$ . For each die  $D_i$ , the probability of rolling the number  $i$  is  $1/2$ , and the probability of each of the other two outcomes is  $1/4$ . At each turn, the dealer must decide whether to (1) keep the same die, (2) switch to another die, or (3) end the game. He chooses (1) with probability  $1/2$  and each of the other with probability  $1/4$ . At the beginning the dealer chooses one of the two dice with equal probability.

- Give an HMM for this situation. Specify the alphabet, the states, the transition probabilities, and the emission probabilities. Include a start state *start*, and assume that the HMM begins in state *start* with probability 1. Also include an end state *end*.
- Suppose that you observe the following sequence of die rolls: 1 1 2 1 2 2. Find a sequence of states which best explains the sequence of roll. What is the probability of this sequence? Find the answer by completing the Viterbi table. Include backtrack arrows in the cells so you can trace back the sequence of states. Some of the following facts may be useful:

$$\log_2(0) = -\infty$$

$$\log_2(1/4) = -2$$

$$\log_2(1/2) = -1$$

$$\log_2(1) = 0$$

- There are actually two optimal sequences of states for this sequence of die rolls. What is the other sequence of states?