

MATH 5610, Take-Home Final Exam  
(due Thursday, 5/12/05 at 2 p.m.)

**Instructions:** This is an open book exam. You may use any resources you like to answer these questions except another person. **You may not discuss the test with anyone except your instructor.** The exam is due at 2 p.m. on Thursday, May 12. No late exams will be accepted unless arrangements are made with the instructor by 1 p.m. on Friday, May 6.

**Name:** \_\_\_\_\_

**Scores:**

1. (25 pts) \_\_\_\_\_

2. (25 pts) \_\_\_\_\_

3. (25 pts) \_\_\_\_\_

4. (25 pts) \_\_\_\_\_

Total \_\_\_\_\_

- Let  $s$  and  $t$  be two DNA sequences with  $|s| = n$  and  $|t| = m$ . Earlier in the semester you developed an algorithm to compute the optimal global alignment of two sequences  $s$  and  $t$  using an affine gap penalty  $g(k) = \alpha k + \beta$  and a scoring function  $w : \{A, T, C, G\} \times \{A, T, C, G\} \rightarrow \mathbb{R}$ .

In this problem, you will be asked to describe a modified algorithm that will also take into account secondary structure elements. The modified algorithm will accept as input two “annotation” functions  $f : s \rightarrow \{h, r, b, \cdot\}^n$  and  $g : t \rightarrow \{h, r, b, \cdot\}^m$ , which assign to each position of  $s$  and  $t$  information about the secondary structure of the associated protein. Here ‘ $h$ ’ denotes helix, ‘ $b$ ’ denotes beta sheet, ‘ $r$ ’ denotes turn, and ‘ $\cdot$ ’ denotes unknown. For example, for  $s = \text{‘ACGTAACGTGGCATATAC’}$  we might have  $f(s) = hhh \dots rrrbbb$ .

We would like to find the maximum scoring global pairwise alignment between  $s$  and  $t$  using the affine gap penalty  $g$  and scoring function  $w$ . However, for any symbol  $s_i \in s$ , it must be the case that either

- $s_i$  is aligned against a gap in  $t$ , or
- if  $s_i$  is aligned against  $t_j \in t$ , then either
  - $f(s)_i = g(t)_j$  or
  - $f(s)_i = \cdot$  or
  - $g(t)_j = \cdot$ .

Similar rules apply to each symbol  $t_j$  in  $t$ . Specifically, either

- $t_j$  is aligned against a gap in  $s$ , or
- if  $t_j$  is aligned against  $s_i \in s$ , then either
  - $g(t)_j = f(s)_i$  or
  - $f(s)_i = \cdot$  or
  - $g(t)_j = \cdot$ .

Show how to compute this alignment in  $O(nm)$  time.

- Design a hidden Markov model (HMM) to detect open reading frames in DNA. Draw out your HMM including all states, arcs, and emission and transition probabilities. Describe how you would use this HMM to identify all ORFs in a DNA sequence.

3. Consider the following set of DNA fragments:

$$\mathcal{F} := \{\text{TACCA}, \text{GTAC}, \text{AGTACC}, \text{ACTGA}\}.$$

- (a) Construct the overlap graph  $\mathcal{OG}(\mathcal{F})$  corresponding to this set of fragments.
  - (b) Using the greedy algorithm, construct the max-weight Hamiltonian path for  $\mathcal{OG}(\mathcal{F})$  and state the corresponding superstring for  $\mathcal{F}$ .
  - (c) Note that the superstring you found is longer than the superstring AGTACCACTGA. Describe a simple change to the method above that will yield shorter superstrings.
4. A gene expression array experiment was performed to identify genes that are differentially expressed when alcohol is present in the blood. In this experiment, data was collected from two groups of 4 mice. The first group was the control group. Alcohol was administered to the second group so that each rat reached a blood alcohol level of 10%. Gene expression data was measured from a sample of brain tissue from each mouse. The data was normalized and the logarithm of the data is posted at

[www-math.cudenver.edu/~billups/courses/ma5610/finaldata.txt](http://www-math.cudenver.edu/~billups/courses/ma5610/finaldata.txt).

In this file, each row corresponds to a gene. The first column is the gene number. The next four columns are the log expression values from the control group. The last four columns are the log expression values for the second group.

For your convenience, the data is stored in MATLAB format in the same directory in the file `finaldata.mat` file. This file can be loaded into MATLAB using the `load` command. This will load a matrix `X`, whose  $i$ th row contains the 8 expression values for the  $i$ th gene.

Your job is to identify as many differentially expressed genes as possible from the above data. However, because expensive followup studies will be performed for each identified gene, it is important that no more than 10% of the genes you identify be false positives.

Using whatever method you choose, identify a list of genes that you would recommend for further study. Justify your recommendation.