Exam in TDT4287 Algorithms for bioinformatics

Monday 5th of December 2011
Time: 09:00-13:00

Allowed aids:
D - No printed or handwritten aids allowed. Given, simple calculator allowed (HP30S).

All sub-problems count equally unless stated otherwise.

Problem 1  String indexes  (a: 10%, b: 10%)

a) Draw the keyword tree for the set of strings $T=\{"ACGT", "CAT", "CACG"\}$. Add the tree’s non-trivial failure links.

b) Draw the generalized suffix tree for the set of strings $T=\{"ACGT", "CAT", "CACG"\}$ and add the suffix links.

Problem 2  Alignment  (a: 10%, b: 15%)

Assume you have sequences $v = v_1 \ldots v_n$ and $w = w_1 \ldots w_m$.

a) Given $v = CTATCGAA$ and $w = TAACC$, find the best global alignment. Also, find the best local alignment(s). Note that you only have to show the alignments themselves. In the alignments, matches are given a score $+1$, and mismatches and gaps score $-1$.

b) Assume that we want to find the best alignment of a suffix of $v$ to a prefix of $w$. This is like a global alignment, except that gaps in the end of $v$ and in the beginning of $w$ are not penalized. Give an $O(nm)$ algorithm for finding the best such alignment and show how the algorithm finds the best fit. Use the same scoring scheme as above.
Problem 3  Hidden Markov model  (a: 5%, b: 5%, c: 10 %, d: 10%)

The following state diagram represents a Hidden Markov Model (HMM) generating DNA sequences.

```
1 0.9
   / 0.1
  / 1   1
  | 1
  | 3
  | 2
  | 4
```

The initial state distribution is \( \pi = [0.7, 0.1, 0.1, 0.1] \).

a) Write down the transition probability matrix \( P \) for this HMM. States 2–4 model a motif, while state 1 represents the background sequence. List the possible motifs that this HMM can generate and the corresponding emission probabilities.

b) The observed sequence is \( O = TATA C \). List all possible paths \( Q \) that can generate \( O \). Calculate the joint probability \( Pr(Q \& O) \) for two of the paths.

c) Apply the Viterbi algorithm to find \( Pr(Q \& O) \) the most probable path. Write down the recurrence equation and the computed values. Compare with your result in b.

d) Show how the same motifs can be represented by a Position Weight Matrix (PWM). Then, show how you can modify the HMM for each of the following changes in the motif model. Draw a state diagram in each case.

   i) allowing a deletion of the second symbol
   ii) allowing an insertion of one or more symbols between the first and second symbols
   iii) allowing the motifs in the original model, except TCC.

Can these modified models be represented by PWMs?

Problem 4  Search performance  (a: 5%, b: 5%)
a) You have a protein sequence $s$, a database $D$ that contains both homologous $H$ and non-homologous $N$ sequences, and a search method $f(s, D)$ (such as BLAST). Describe what a true positive (TP), false positive (FP), true negative (TN), and false negative (FN) search result is in relation to $f(s, D), H,$ and $N$.

b) What is a ROC graph? Draw ROC-graphs that shows random and perfect classification. Define the x-axis and y-axis in the graph in terms of TP, FP, TN, and FN.

**Problem 5  Shortest unique substring  (15%)**

Design an efficient algorithm that finds the shortest substring occurring only once in a given string $s = s_1...s_n$. What is the running time of your algorithm?