Algorithms for Molecular Biology
© Ron Shamir
School of Computer Science
Tel Aviv University
January 23, 2002

Algorithms for Molecular Biology: Assignment No. 5
due on February 6th

General Guidelines:
This assignment is part of your test in the course. It should be done **INDEPENDENTLY**, without any help from others. If you use articles or books, please include the reference in your relevant answer.

Credit: Solve 6 items out of 8 for 12/75.

1 Phylogeny

1. You are the computational biologist of the *Jurassic Park*. You are given a phylogenetic tree, with branch lengths, including the park animals (and only them) as internal nodes, and contemporary sequences of all their present day descendants. The sequences of the park animals are not available. You also have the *PAM* matrix, and you are working only on one specific protein, which is multiply aligned without gaps for all the input sequences.

   (a) Reconstruct the most plausible set of sequences of the ancient creatures, i.e, the set whose probability given the data is maximized.

   (b) In contrast to item 1a, assume all you care about is one specific creature (*T-Rex*, for instance). Fortunately, T-Rex has no known descendants, and its position in the tree is known. Reconstruct the most plausible T-Rex sequence given the contemporary sequences.

2 Physical Mapping

2. Prove that in the Lander-Waterman model given in class the expected length of an apparent island is $\frac{e^{R(1-\theta)}-1}{R} + \theta$.

3. Some physical mapping experiments have all of their clones roughly of the same length, motivating interest in *unit interval graphs* which are interval graphs which have a realizations by a set of equal length intervals. Prove that for a finite interval graph G, the following are equivalent:

(a) G is unit interval.

(b) G has a realization in which no interval strictly contains another.

(c) G has no induced subgraph isomorphic to $K_{1,3}$ ($K_{1,3}$ is the graph $G'(V = \{a, b, c, d\}, E = \{ab, ac, ad\})$).

3 Genome Rearrangements

4. The (i, j, k) *transposition* on a permutation $\sigma_1, \dots, \sigma_n$ is defined for $i < j < k$ as the operation of removing the block $\sigma_i, \dots, \sigma_{j-1}$ and re-inserting it before σ_k. The resulting permutation is $\sigma_1, \dots, \sigma_{i-1}, \sigma_j, \sigma_{j+1}, \dots, \sigma_{k-1}, \sigma_i, \sigma_{i+1}, \dots, \sigma_{j-1}, \sigma_k, \sigma_{k+1}, \dots, \sigma_n$. Augment σ by defining $\sigma_0 = 0, \sigma_{n+1} = n + 1$. Define the breakpoint graph $G(V, E_{red} \cup E_{blue})$ for this operation, by $V = \{0, \dots, 2n + 1\}$, $E_{red} = \{(2i, 2i + 1)\}$, $E_{blue} = \{(2\sigma_i, 2\sigma_{i+1} - 1)\}$.

   (a) Prove that G is a union of disjoint cycles, and every transposition affects at most three cycles.

   (b) Prove that when applying a transposition the number of cycles changes by $0, +2$, or -2, and so does the number of cycles of odd-length (i.e., cycles with $4k + 2$ edges: $2k - 1$ of each color).

5. Recall that the score of a reversal according to Bergeron's method is the number of oriented pairs in the resulting permutation. In the overlap graph each oriented pair is translated into an oriented vertex.

   (a) Prove that the score of a reversal corresponding to an oriented vertex v is $T + U - O - 1$, where T is the total number of oriented pairs, U is the number of unoriented pairs adjacent to v and O is the number of oriented vertices adjacent to v.

   (b) Use the previous item to prove that a reversal with maximal score is safe.

4 Gene Expression and Network Reconstruction

6. Consider sequencing a target of length l by hybridization to an array containing all possible k-mers. Prove that sequence reconstruction from the hybridization experiment is ambiguous whenever there exists a quadruple of indices $1 \leq i_1 < i_2 \leq j_1 < j_2 \leq l$, such that the $(k - 1)$-mer in position i_r equals that in position j_r, and $i_2 - i_1 \neq j_2 - j_1$. Compute the expected number of such quadruples, assuming a uniformly random sequence.

7. Suppose we wish to reconstruct a boolean network whose indegree is bounded by d.

   (a) Prove the upper bound of $O(n^{2d})$ on the required number of experiments sketched in class.

   (b) Improve the latter bound to $O(n^{d+1}2^{2d})$.

8. Given a hierarchical clustering of gene expression data in the form of a binary tree T devise an algorithm that orders the leafs of the tree for optimal visualization such that:

   (a) The order respects the tree topology, i.e., the leafs of any subtree appear consecutively.

   (b) The order maximizes $\sum_{i=1}^{n} S(v_i, v_{i+1})$ where v_i is the expression vector of i-th leaf in the ordering, and S is some similarity function between vectors.