Course: Algorithms for Biological Sequence Analysis Midterm exam. Nov. 9, 2005
Instructor: Kun-Mao Chao TA: Yao-Ting Huang

Note: Both the correctness and efficiency of your algorithms will be evaluated. You are required to justify your answer.

1. (10%) Given a real number sequence *A*, please describe an algorithm for computing the maximum-average segment ending at each index of *A*. You can use the following sequence to describe your algorithm.

   3 5 6 8 3 6 7 3

2. (10%) You are given a real number sequence *A*. Design a linear-time algorithm for computing the nearest larger elements for all elements of *A*. You can use the following sequence to illustrate your algorithm.

A	3	5	6	8	3	6	7	3
nearest larger element	5	6	8	x	8	7	8	7

3. (10%) What is a "better partner" of each index defined in RMSQ? Explain why it is better.

4. (15%) Given two sequences *A* and *B*, we say that a common substring-subsequence is a substring, defined as a contiguous subsequence, of *A* which is also a subsequence of *B*. Give an efficient algorithm for finding the longest common substring-subsequence between *A* and *B*.

5. (15%) Assume the following scoring scheme:
   A match is given a bonus +8;
   A mismatch is penalized by –5;
   A gap of length *k*, where *k* is between 1 and 50, is given a constant penalty 6+3*k*;
   A gap of length more than 50 is given a constant penalty 156;
   Give the recurrences for computing the score of an optimal global alignment. Explain why they work.

6. (a) (5%) Describe Hirschberg's linear-space idea for delivering an optimal alignment. Explain why the time complexity remains the same as that of merely computing the score of an optimal alignment. Show that directly applying this approach to a band might cause an additional log *n* factor in time, where *n* is the sequence length.
   (b) (5%) Describe a linear-space alignment method for aligning two sequences in a band.

7. (15%; 3% each) (a) What is haplotype inference? (b) Give an Integer Quadratic Programming formulation for haplotype inference. (c) What are tag SNPs? (d) What is an LD bin? (e) Explain why the problem of finding a minimum set of LD bins is related to the minimum clique cover problem

8. Consider the problem of computing all Δ-points of two sequences of lengths *M* and *N*, where $M \le N$. (a) (10%) Describe a method that works in $O(MN)$ time and $O(N^{\frac{1}{4}}M+N)$ working space; (b) (5%) Describe a method that works in $O(MN)$ time and $O(M^{1+\frac{1}{4}}+N)$ working space.