Name: _____ Student ID: _____

CS482 Midterm Exam, Spring 2011

(2:30-3:50pm)

***Attention**: There are in total 10 questions (2 pages).*
***Attention**: Some questions may have multiple correct answers. You may not get marks if your answer is correct but not optimal.*
***Attention**: Answer all questions within the space provided in the questions.*

Question 1 (10 marks): The two types of DNA base pairs are _____(A/C/G/T – A/C/G/T) and _____ (A/C/G/T – A/C/G/T). For a segment of DNA sequence ATTATTACG (from 5' to 3') on one strand, the sequence on the other strand at the same location is _____(from 5' to 3').

Question 2 (8 marks): The dynamic programming algorithm for the multiple alignment of *n* length-*m* sequences takes O(_____) time.

Question 3 (16 marks): For two input sequences with lengths *m* and *n*, respectively, the best algorithm for computing optimal pairwise alignment score takes O(_____) time and O(_____) space. The best algorithm for computing the actual optimal alignment takes O(_____) time and O(_____) space.

Question 4 (10 marks): Find all the hits of the following alignment with the spaced seed 111*1**1. Put all the starting positions of the hits in the following blank: _____. The index of the string positions start with 1.

```
GAGTACTCAACACCA
||  |||||||||  ||
GAATACTCAACAGCA
```

Question 5 (10 marks): Briefly (in a couple of sentences) explain why spaced seed is more sensitive than the consecutive seed with the same weight for sequence homology search.

Question 6. (10 marks) Log likelihood ratio has been widely used to define score functions in bioinformatics, including the BLOSUM matrix. In the computing of BLOSUM matrix, suppose the background frequencies of two different letters a and b are p(a) and p(b), respectively. Suppose the frequency of observing a and b together in the same column of the sequence alignments in the BLOCK database is p(a,b). What is the similarity score between a and b?

Answer: _____

Question 7 (8 marks): In order to use dynamic programming to solve an optimization problem, the general requirement for the sub-solution of the optimal solution is:

_____

Question 8 (10 marks): In the skew algorithm for suffix sorting. The n suffixes are divided into two sets with size $\frac{2n}{3}$ and $\frac{n}{3}$, respectively. The $\frac{2n}{3}$ part is sorted recursively. The $\frac{n}{3}$ part is sorted in linear time. And the merging will take linear time. Thus the total running time is bounded by $T(n) \leq T\left(\frac{2n}{3}\right) + c \cdot n$ for a constant c. Please prove that the running time is linear to n.

Question 9 (10 marks): Draw a suffix tree of the string "banana$". Use the actual substring, instead of the start/end positions of the substring, to label each edge of the suffix tree.

Question 10 (10 marks): Write out the recurrence relation for the local alignment of two sequences S[1..m] and T[1..n]. Assume the score scheme is match=1, mismatch=-1, indel=-2. No need to include the initialization and the backtracking step.