Policy on collaboration

When solving your homework problems and working on your project, you may discuss HIGH-LEVEL approaches to the homework problems with your classmates, HOWEVER, you are to work out all details of any solutions discussed and write up the solution completely on your own. In particular, when working with a student on an assigned homework problem you should do so verbally – Nothing should be written. This is aimed at keeping your discussion at a high level so everyone can work out the details on their own. Please follow the spirit of this rather than working to find ways to share details verbally. Also you must clearly acknowledge anyone (except the instructor) with whom you discussed any problem and say briefly what you discussed.

COPYING MATERIAL FROM WEB SITES, BOOKS, OR OTHER SOURCES IS CONSIDERED EQUIVALENT TO COPYING FROM ANOTHER STUDENT. INFORMATION FROM THESE SOURCES MAY BE USED IF THE SOURCES ARE ACKNOWLEDGED AND YOU HAVE WRITTEN DOWN THE IDEAS ON YOUR OWN.

I have read the above policy and accept the terms.

Print Your Name __________________________

Your Signature __________________________
Homework 1
Due: Oct 4, before class starts

Problem 1 (15 points)
You are given the following DNA sequence, which is believed to contain a small protein-coding gene.
GGAGGC\text{GTAA AATCGGTACT GGTAATGCAA ACTATGG}

• If this sequence is fully transcribed (used as a coding strand), what is the corresponding mRNA sequence?

• Which region of the mRNA do you think can be translated into a protein (hint: Can you identify the start codon and stop codon from the mRNA sequence?)

• What is the protein sequence encoded by the gene?

• If the reverse-complementary strand of the DNA sequence is also transcribed, what will be the mRNA sequence?

• Do you think the reverse-complementary strand can encode a protein?

Problem 2 (20 points)
Consider the sequences $v = \text{TACGGGTAT}$ and $w = \text{GGACGTACG}$. Assume that the match score is +1, and the mismatch and gap penalties are -1.

• Fill out the dynamic programming table for a global alignment between $v$ and $w$. Draw arrows in the cells to store traceback information. What is the score of the optimal global alignment and what alignment(s) achieve this score?

• Fill out the dynamic programming table for a local alignment between $v$ and $w$. Draw arrows in the cells to store traceback information. What is the score of the optimal local alignment in this case and what alignment(s) achieve this score?

Problem 3 (10 points)
Solve problem 6.35 on pages 3-4 of this document (the text within the rectangle region).
Problem 4 (10 points) KMP algorithm and Failure Links

Compute SP(i) for the following pattern: actaactc. Show the failure links that are constructed using the SP values. Illustrate how the failure links can be used to speed up the search of the pattern in the following text: actaactaactc. How many comparisons are needed by the naive algorithm to find the pattern? How many comparisons are needed by the KMP algorithm?

Problem 5 (15 points) Suffix Tree

1. Draw a suffix tree for the string taataaataa. Label the edges and terminal nodes explicitly.

2. Briefly describe an efficient suffix tree-based algorithm (with $O(n)$ running time) for finding the shortest non-repeated string in a text, that is, a shortest string that appears in the text only once. Some informal description of the algorithm is good enough, as long as I can understand your idea and why it is efficient. You can use graphs.

3. Apply your algorithm to the example above.

Bonus (5 points)

How much time did you spend on this homework? Who did you discuss with and what was the discussion about? How is the difficulty level? Do you have any comments about the lectures and the homework?
A tandem repeat $P^k$ of a pattern $P = p_1 \ldots p_n$ is a pattern of length $n \cdot k$ formed by concatenation of $k$ copies of $P$. Let $P$ be a pattern and $T$ be a text of length $m$. The Tandem Repeat problem is to find a best local alignment of $T$ with some tandem repeat of $P$. This amounts to aligning $P^k$ against $T$ and the standard local alignment algorithm solves this problem in $O(km^2)$ time.

**Problem 6.33**

Devise a faster algorithm for solving the tandem repeat problem.

An alignment of circular strings is defined as an alignment of linear strings formed by cutting (linearizing) these circular strings at arbitrary positions. The following problem asks to find the cut points of two circular strings that maximize the alignment of the resulting linear strings.

**Problem 6.34**

Devise an efficient algorithm to find an optimal alignment (local and global) of circular strings.

The early graphical method for comparing nucleotide sequences—dot matrices—still yields one of the best visual representations of sequence similarities. The axes in a dot matrix correspond to the two sequences $v = v_1 \ldots v_n$ and $w = w_1 \ldots w_m$. A dot is placed at coordinates $(i, j)$ if the substrings $s_i \ldots s_{i+k}$ and $t_j \ldots t_{j+k}$ are sufficiently similar. Two such substrings are considered to be sufficiently similar if the Hamming distance between them is at most $d$.

When the sequences are very long, it is not necessary to show exact coordinates; figure 6.29 is based on the sequences corresponding to the $\beta$-globin gene in human and mouse. In these plots each axis is on the order of 1000 base pairs long, $k = 10$ and $d = 2$.

**Problem 6.35**

Use figure 6.29 to answer the following questions:

- How many exons are in the human $\beta$-globulin gene?
- The dot matrix in figure 6.29 (top) is between the mouse and human genes (i.e., all introns and exons are present). Do you think the number of exons in the $\beta$-globulin gene is different in the human genome as compared to the mouse genome?
- Label segments of the axes of the human and mouse genes in figure 6.29 to show where the introns and exons would be located.

A local alignment between two different strings $v$ and $w$ finds a pair of substrings, one in $v$ and the other in $w$, with maximum similarity. Suppose that we want to find a pair of (nonoverlapping) substrings within string $v$ with maximum similarity (Optimal Inexact Repeat problem). Computing an optimal local alignment between $v$ and $v$ does not solve the problem, since the resulting alignment may correspond to overlapping substrings.

**Problem 6.36**

Devise an algorithm for the Optimal Inexact Repeat problem.
In the chimeric alignment problem, a string $v$ and a set of strings $\{w_1, \ldots, w_N\}$ are given, and the problem is to find $\max_{1 \leq i, j \leq N} s(v, w_i \circ w_j)$ where $w_i \circ w_j$ is the concatenation of $w_i$ and $w_j$ ($s(\cdot, \cdot)$ stand for the score of optimal global alignment).

**Problem 6.37**
Devise an efficient algorithm for the chimeric alignment problem.

A virus infects a bacterium, and modifies a replication process in the bacterium by inserting

- at every A, a polyA of length 1 to 5.
- at every C, a polyC of length 1 to 10.
- at every G, a polyG of arbitrary length $\geq 1$.
- at every T, a polyT of arbitrary length $\geq 1$.

No gaps or other insertions are allowed in the virally modified DNA. For example, the sequence AAATAAAAGGGCCCCCTTTTTTTTCC is an infected version of ATAGCTC.

**Problem 6.38**
Given sequences $v$ and $w$, describe an efficient algorithm that will determine if $v$ could be an infected version of $w$. 