CS 262: Homework 1

Due 1/26/2016 at the beginning of class

Collaboration is allowed in groups of at most three students, but you must submit separate writeups. Please write the names of all your collaborators on your solutions. You are not allowed to copy group work. If you are working alone, we will drop the problem with the lowest score. If you submit your solutions after the submission deadline, you must write the date and time of submission on your writeup. Under no circumstances will a homework be accepted more than three days after its due date.

Problem 1 (10 points)

NOTE: Feel free to use external sources (e.g. Wikipedia) to find out more about the biological concepts in this question but use your own words for the answers in your write-up.

(a) (4 points) Identify the type of the following mutation events (e.g. point mutation, insertion, deletion, inversion, translocation, etc) and specify what type of effect they are likely to have on the resulting gene product. Be as specific as possible. Assume that this sequence corresponds to an exon and that the translation reading frame starts with the first nucleotide of each mutated sequence.

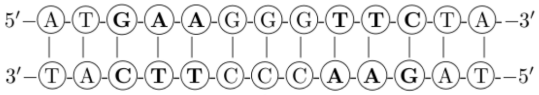
$5' - \text{CCTGCCTTT}\textbf{C}\text{GACGT} - 3' \quad \underset{\text{(i)}}{\overset{↯}{\rightsquigarrow}} \quad 5' - \text{CCTGCCTTT}\textbf{T}\text{GACGT} - 3'$

$5' - \text{CATGCCTTTCGACGT} - 3' \quad \underset{\text{(ii)}}{\overset{↯}{\rightsquigarrow}} \quad 5' - \text{CAT}\textbf{AT}\text{GCCTTTCGACGT} - 3'$

(b) (6 points) Find the reverse complement of the following DNA strand:

```
5′–(G)-(G)-(A)-(T)-(C)-(C)–3′
    |   |   |   |   |   |
3′–(?)-(?)-(?)-(?)-(?)-(?)–5′
```

You should notice something interesting about this sequence and its reverse complement. This type of sequence is a DNA palindrome. Palindromic sequences have a range of important biological properties. For example, palindromic sequences are a common target of restriction endonucleases (enzymes that cut DNA near a specific site); the above sequence is recognized by the restriction endonuclease BamH1 in the Bacillus amyloliquefaciens bacteria, which cleaves the DNA right after the 5' guanine. Using the information about nucleotide binding affinity and the special property of DNA palindromes, find a plausible alternative (non-linear) structure for the following DNA molecule shown in its standard B-DNA (linear) conformation (hint: some of the nucleotides will be left unpaired):


5′–A-T-G-A-A-G-G-G-T-T-C-T-A–3′
3′–T-A-C-T-T-C-C-C-A-A-G-A-T–5′

Serving as mutational hotspots, such alternative DNA structures can have significant consequences on an organism; in particular, they can lead to chromosomal translocations, which can result in cancer or intellectual disability.

Problem 2 (15 points)

We have discussed several different alignment algorithms. The algorithms differ in scope (global, local), end- treatment (ends-free, ends-full), and gap-penalty (constant, linear, affine, convex, custom). We discussed overlap detection, bounded dynamic programming, linear space dynamic programming, and other variants. We also talked about BLAST-based sequence homology search. Each algorithm is appropriate for different biological settings. For each of the following settings, describe which algorithm you would use, and justify your answer. You are also free to design a new variant or extension of these algorithms, or a new algorithm, if you think that your algorithm would be more appropriate than one covered in class. In this case, please describe your algorithm/variant/extension in broad terms.

(a) (5 points) You are given the sequence of a mature mRNA from a given organism and you would like to find its corresponding location(s) in the organisms genome.

(b) (5 points) You want to find single nucleotide polymorphisms (SNPs) or small insertions and deletions between the two parental haplotypes of the same gene in the same individual, given the sequence from the genes two haplotypes.

(c) (5 points) You are given a protein sequence and would like to find proteins in the Protein Data Bank (which contains thousands of proteins) with which it shares domains. Assume that a typical domain that you wish to find is a subsequence of at least 10 amino acids and you are only interested in matches with at least 75% sequence identity to your query sequence.

Problem 3 (25 points)

(a) (5 points) Prove that the number of optimal global alignments between sequences x and y can be exponential in $|x|$ and $|y|$ (where $|x|$ is the length of x), by constructing an example of sequences x and y that have exponentially many optimal (global) alignments, for fixed scoring parameters m, s, and d.

(b) (7 points) A traceback path is the backtracking path that follows the pointers to reconstruct the alignment in the dynamic programming matrix. Since the traceback paths in a dynamic programming matrix correspond one-to-one with optimal alignments, the number of distinct optimal alignments can be obtained by computing the number of distinct traceback paths. Give an algorithm to compute this number in time $O(|x||y|)$ for sequences x and y.

(c) (7 points) Derive the number of different alignments between two n-letter sequences. The answer may be left as a summation.

(d) (6 points) There are 4^n different RNA molecules of length n. How many different double-stranded DNA molecules of length n are there? Recall that a double-stranded DNA molecule is equivalent to its reverse complement.

Problem 4 (20 points)

(a) Recall the linear-space global alignment algorithm which reconstructs the optimal alignment. In this problem, you will be asked to consider possible extensions of it to other alignment problems. For each part below, either briefly describe an extension of the linear-space algorithm to the problem that preserves the space-time complexity, or discuss why this is not possible.

   (i) (7 points) Local alignment (Smith-Waterman), where you are asked to find a single optimal highest scoring local alignment.

   (ii) (7 points) Alignment under affine gap scores.

(b) (6 points) In the linear-space alignment algorithm, the original problem of size $m{\cdot}n$ is reduced to two subproblems of sizes $\frac{k \cdot m}{2}$ and $\frac{(n-k)\cdot m}{2}$. In a parallel implementation of sequence alignment, it is desirable to have a balanced partitioning that breaks the original problem into two subproblems of (almost) equal sizes. Design a linear space alignment algorithm with balanced partitioning.

Problem 5 (10 points)

A string $x = x_1 \ldots x_j$ is a substring of a string $y = y_1 \ldots y_n$, if for some $0 \leq i \leq n-j$, $y_{i+1} \ldots y_{i+j} = x_1 \ldots x_j$. y is then a superstring of x.

A string $x = x_1 \ldots x_j$ is a subsequence of a string $y = y_1 \ldots y_n$, if for some $1 \leq i_1 < \ldots < i_j \leq n$, $y_{i_1} \ldots y_{i_j} = x_1 \ldots x_j$. y is then a supersequence of x.

(a) (6 points) The scoring function of an alignment problem plays an important role in determining the resulting alignment. Given two sequences x and y, describe what the Smith-Waterman algorithm finds for the following scoring schemes:

   (i) $m = 1, s = 0, d = 0$.

   (ii) $m = 1, s = \infty, d = \infty$.

(b) (4 points) Given the strings x and y, devise an algorithm to find the shortest supersequence for both x and y in $O(|x||y|)$ time.

Problem 6 (20 points)

For this problem, let B be the Burrows-Wheeler Transform (BWT) of the word $X = X_1 \ldots X_n$ where $X_n = \$$ is a special symbol that marks the end of a string and is considered lexicographically smaller than any other symbol in the alphabet used to create X. Let also S be the suffix array of X, so that $B_i = X_{S(i)-1}$, when $S(i) > 1$ and $B_i = \$$ otherwise. Finally, let α be a single character of the alphabet used for X and W be a (possibly empty) string from that alphabet.

(a) (5 points) Give the BWT of the word $X = MAMALIGA\$$, together with the corresponding BWT matrix and suffix array S.

(b) Let $C(\alpha)$ be the number of symbols of X that are lexicographically smaller than α and $F(\alpha, i)$ be the number of occurrences of α in $B_1 \ldots B_i$. Finally, let $L(W)$ and $U(W)$ be respectively the indices of the first and last row of the BWT matrix that start with the prefix W.

(i) (4 points) Prove that the first row of the BWT matrix that starts with α is the row indexed $C(\alpha) + 1$. Recall that \$ is lexicographically smaller than any other symbol.

(ii) (6 points) Prove that $L(\alpha W) = C(\alpha) + 1 + F(\alpha, L(W) - 1)$ and $U(\alpha W) = C(\alpha) + F(\alpha, U(W))$.

(iii) (5 points) Assuming that B, S, C, and F are pre-computed, describe how you would use the equations above to compute the indices of all occurrences of W in X in time $O(|W|)$.