Computational Genomics
© Benny Chor and Amos Tanay
School of Computer Science
Tel Aviv University
November 12, 2002

Computational Genomics: Assignment No. 2
due on November 24th, 2002

General Guidelines:
This assignment is part of your final grade in the course. It should be done *independently*, either individually or in *pairs*, without any help from others. Duplicated and copied works will be given zero grade. Using articles or books is perfectly acceptable as long as you include the reference in your relevant answer.
Credit: Solve all items for 120 points + one golden star. Starred (*) sections are harder, but you can accumulate the full 100% grade even without solving them.

1. **Local Alignment Heuristics** (20 pts)

   Describe an example where the FASTA heuristic *misses* the optimal local alignment. For simplicity, you may assume that both sequences S and T are over the alphabet $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$, `ktup`=6, and the similarity score gives each match +4, each mismatch −2, and each indel −3. (If you prefer, you may use any other reasonable scoring matrix.)

2. **Alignment of Alignments** (20 pts)

   In the approximation algorithm for MSA described in class, we have aligned each of the sequences against a single, carefully selected "center" sequence (S_1 in the notation used on the board). Show how, alternatively, we can *efficiently* align two given multiple alignments S_1, S_2 for optimal score under the SP model. The rules are that we cannot change the "internals" of the two original alignments except for adding gaps in some positions of all of the sequences in S_1 (or in S_2).

3. **Multiple Sequence Alignment** (40 pts)

   The multiple alignment approximation algorithm shown in class aligns all sequences to a well-chosen "center" sequence. Finding this "center" sequence requires most of the running time of the algorithm. In this exercise we (you ?) will investigate how essential the choice of the "center" sequence is. ($d(M)$ denotes the SP score of a multiple alignment M, and as in class we will assume the cost matrix δ is a metric)

   (a) Show that choosing a different sequence and aligning all sequences to it may yield an infinitely high approximation ratio. Do this by showing that for every ratio r,

there exists a set of sequences, a cost matrix, and a "bad" sequence b, such that the SP score of the multiple alignment M_b found using b is worse then the optimal alignment M_{opt} by at least r: $d(M_b)/d(M_{\text{opt}}) \geq r$. (**Bonus:** what's the worst ratio you can get with k sequences ?)

(b) Show that despite this, for any set of sequences, the mean ratio over all "center sequence" choices is at most 2: $E_b\left(\frac{d(M_b)}{d(M_{\text{opt}})}\right) \leq 2$.

(c) Use this to devise an $O(kn^2)$ randomized algorithm (instead of the $O(k^2n^2)$ deterministic algorithm shown in class) that finds a multiple sequence alignment M_{Rand}, such that $P\left(\frac{d(M_{\text{Rand}})}{d(M_{\text{opt}})} \leq 3\right) \geq \frac{1}{2}$ for all input sets of sequences.

(d*) **Bonus:** For any constant $r > 2$ and $p < 1$, describe an $O(kn^2)$ randomized algorithm that finds a multiple sequence alignment $M_{r,p}$ such that

$$P\left(\frac{d(M_{r,p})}{d(M_{\text{opt}})} \leq r\right) \geq p \ .$$

4. **Center Star MSA Approximation** (20 pts)

Show that the approximation algorithm for multiple alignment shown in class does not achieve a constant approximation ratio better than 2. Do this by showing how to construct, for every $r < 2$, a set of sequences and a metric such that $d(M_c)/d(M_{\text{opt}}) > r$, where $d(M_c)$ is SP score of the alignment found by the algorithm. (Note that this does *not* prove that this problem cannot be approximated by a constant ration better then 2, but simply that this specific algorithm does not achieve a better ratio).

5. **Consensus Sequences** (20 pts)

Let $\mathcal{S}$ be a set of sequences $S_1, \ldots, S_k$. The *consensus sequence* is a sequence that represents the common features of all sequences in $\mathcal{S}$. Formally, we define the *consensus error* of a given sequence, T, as $E(T) = \sum_{i=1}^{k} D(T, S_i)$. Let C be the sequence that minimizes the consensus error. As can be expected, computing the minimum consensus error is NP-Hard. Show that if the cost matrix δ is a metric, then there is a sequence $S_i \in \mathcal{S}$, such that $E(S_i) \leq 2E(C)$ (namely, S_i is a 2-approximation for the consensus string).

6. **And the golden star question is:**

The game of ChocBar! is a two players game, played on an chocolate bar of size M-by-N, with a *poisonous* leftmost/topmost square. Players take turns picking chocolate squares. In her (or his) turn, a player must pick one of the remaining squares, and eat it along with all the squares that are "below it and to its right". Using matrix-notation, the poisonous square is denoted by entry $(1,1)$, and the initial "state" of the brand new bar consists of the whole bar $\{(i,j)|1 \leq i \leq M, 1 \leq j \leq N\}$. Picking the square (i_0, j_0) means that one has to eat all the remaining squares (i,j) for which *both* $i \geq i_0$ and $j \geq j_0$ hold. The player that eats the poisonous (topmost/leftmost) square dies in excruciating pains, and consequently loses the game. Picking $(1,1)$ to be your move kills you (in pains...), so a player who is non-suicidal will not play that move (unless she/he is forced to).

For example, if $M = 5$ and $N = 3$, then the initial game-position is

$$\begin{array}{ccc} X & X & X \\ X & X & X \\ X & X & X \\ X & X & X \\ X & X & X \end{array} .$$

The first player may choose to play $(5,3)$, in which case the chocolate bar (game) state becomes

$$\begin{array}{ccc} X & X & X \\ X & X & X \\ X & X & X \\ X & X & X \\ X & X & \end{array} ,$$

or she may choose to play $(2,2)$, which shrinks the chocolate bar to

$$\begin{array}{ccc} X & X & X \\ X & X & X \\ X & & \\ X & & \\ X & & \end{array} ,$$

and so on.

A *strategy* for a player can be thought of as a table (possibly huge) which tells the player which move to make in each given position (state of the game). A well known result from game theory states that in every two person game like the one we deal with (*i.e.* finite and deterministic), either the first player or the second one has a *winning*

strategy. This theorem applies, for example, to the game of chess, provided a draw is counted as a win for one side (but we are still a long way from knowing which player has a winning strategy, let alone *describing* such winning strategy).

(a) Suppose $N > 1$ and our chocolate bar is a square of size N-by-N. Describe (in words) a simple winning strategy for the first player in a game of N-by-N ChocBar!

(b*) Suppose $M, N > 1$ and our bar is a rectangle of size M-by-N. Prove the *existence* of winning strategy for the first player in an a game of M-by-N ChocBar!

The existence proof we know of is very simple, short, and extremely elegant. Unfortunately, the proof gives *no clue* what that winning strategy might be.

Remark: Constructing such winning strategy for the general case (each value of M and N, $M \neq N$) will make a very impressive Ph.D. thesis (though not necessarily in computational biology...). Constructing a winning strategy for some *specific* values of $M > 3$ and general N *may* make a decent M.Sc. thesis (again, not in computational biology).