Introduction to Computational Biology
Homework 2
2/26/2014

Due 3/5/2014

Problem 1: Concave gap penalty function

Let γ be a gap penalty function defined over non-negative integers. The function γ is called sub-additive iff it satisfies the following: $\gamma(k_1 + k_2 + ... + k_n) \leq \gamma(k_1) + \gamma(k_2) + ... + \gamma(k_n)$.

(a) Show that a concave γ, i.e. one that satisfies $\gamma(x+1)-\gamma(x) \leq \gamma(x)-\gamma(x-1)$, is sub-additive if $\gamma(0) \geq 0$. *Hint*: it is sufficient to show that $\gamma(k_1 + k_2) \leq \gamma(k_1) + \gamma(k_2)$ and express $\gamma(k_1 + k_2)$ as $\gamma(k_1)$ plus the increments up to k_2.

The next set of questions are intended to help you understand why the DP algorithm we saw in class requires γ to be concave. Here's the algorithm again:

$$A(i,j) = \max \begin{cases} A(i-1,j-1) + s(i,j) & (1) \\ A(i,j-k) - \gamma(k),\ k = 1...j & (2) \\ A(i-k,j) - \gamma(k),\ k = 1...i & (3) \end{cases}$$

Without loss of generality, we restrict our attention to alignments that end with a gap in x. Call such an alignment of $x_1...x_i$ and $y_1...y_j$ "good" if it ends with a gap of length k in x for some $k > 0$ and **optimally** aligns $x_1...x_i$ to $y_1...y_{j-k}$.

(b) Show that a "good" alignment of $x_1...x_i$ and $y_1...y_j$ is not necessarily optimal. It is enough to give a counter example.

(c) Show that if γ is concave, then an optimal alignment (that ends with a gap in x) of $x_1...x_i$ and $y_1...y_j$ is a "good" alignment.

(d) Show that if γ is concave, then for any given alignment of $x_1...x_i$ and $y_1...y_j$ with score S, if we split the alignment in two parts with scores S_1 and S_2, then $S_1 + S_2 \leq S$.

(e) Put parts (b) and (c) and (d) together to argue that step (2) must check all $k = 1...j$, and it computes the optimal alignment of $x_1...x_i$ and $y_1...y_j$ that ends with a gap in x.

(f) Construct an instance of an optimal alignment (that ends with a gap in x) that is not "good". (you cannot use a concave gap function according to part (c)). Argue that the algorithm above will not work properly if such an instance can be constructed.

Problem 2: Random star alignment

Let $M = \sum_{S_i} D(S_c, S_i)$ as defined in the star alignment algorithm. Suppose that instead of S_c, we choose a string at random to be the center of the star. Let M_R be defined in an analogous way when S_c is replaced by a random string.

(a) Show that $E[M_R] \leq 2M$ and argue that the expected score of the alignment is at most a factor of 4 of optimal.

(b) Next, show that the median of M_R is at most $3M$.

(c) Finally, argue that the value of the multiple alignment is at most a factor of 6 of optimal with probability at least 1/2.

Problem 3: Consensus string

Given a set of strings S, let S_c be the center string as defined for the star alignment of S, and assume that the scoring scheme (distance) satisfies the triangular inequality. In addition, for any string T, define the consensus error $E(T) = \sum_{S_i \in S} D(T, S_i)$.

(a) Let S^*, not necessarily in S, be the string that minimizes $E(S^*)$. Show that for any string T in S:

$$E(T) \leq (|S| - 2)D(T, S^*) + E(S^*)$$

(b) Show that if $T \in S$ is the closest to S^*, then:

$$\frac{E(T)}{E(S^*)} < 2$$

(c) Show that $E(S_c)/E(S^*) < 2$.

Problem 4: Example substitution matrix

Let's say we would like to build a DNA substitution matrix (4x4 matrix) optimized for finding 88% identity alignments.

- assume the background frequencies are identical, i.e. $p_i = 0.25$ for each nucleotide i
- assume that all matches are equally probable
- assume that all mismatches are equally probable

(a) Compute q_{ij} for all i and j.

(b) Construct the matrix using the log-likelihood ratio.

(c) Choose a scaling factor λ to make the substitution matrix close to an integer matrix.

Problem 5: Unrevealing BLOSUM62

(a) Find the 20x20 BLOSUM62 substitution matrix online. BLOSUM62 has the property that the background probabilities and the observed probabilities are consistent, i.e. $p_i = \sum_j q_{ij}$.

(b) Given a symmetric and consistent substitution matrix S, with a scaling factor λ, let M be the matrix $e^{\lambda S}$. Note $M_{ij} = \frac{q_{ij}}{p_i p_j}$. Let Y be the inverse of M (assuming M is invertible). Show that the sum of the i^{th} column (or row) of Y must be equal to the background probability p_i. *Hint*: Consider the vector $p = [p_1, \ldots, p_n]$. Show that $pM = [1, \ldots, 1]$. Use this result to compute pMY in two ways.

(c) Using the above strategy, and knowing that $\lambda = 0.3176$ for BLOSUM62, find the background probabilities p_i for BLOSUM62.

(d) Using part (c), find the observed set of probabilities q_{ij}.