

**Introduction to Computational Biology**  
**Homework 3**  
**10/28/09**

**Due 11/11/09**

**Problem 1: Database Search**

The purpose of this problem is to make you experiment “a little bit” with BLAST.

- (a) Go to <http://www.ncbi.nlm.nih.gov/> and click on Genomic Biology in the left margin.
- (b) Click on Homo Sapien in the middle of the page. This will display a the set of human chromosomes. Choose your favorite chromosome.
- (c) Click on a gene under symbol (you can scroll up and down on the chromosome and you can zoom in).
- (d) Scroll down to NCBI Reference Sequences (RefSeq) Database and find the accession numbers for the mRNA sequence and the protein sequence. These numbers uniquely identify the sequences. Click on the accession number of the mRNA. This will display information about that gene like name, references, authors, etc...
- (e) Scroll down until you see the translated amino acid sequence. This is the amino acid sequence that is produced by the mRNA.
- (f) Cut and paste the translated amino acid sequence in any text editor.
- (g) Pick a substring of that sequence of 30-50 amino acids. Go the <http://www.ncbi.nlm.nih.gov/blast>. Choose protein blast. Paste the substring in the query box.
- (h) Blast it! Verify that your original sequence is one of the sequences that score highly with the query. Note the score in bits and the e-value.

**Problem 2: 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  $\lambda$  to make the substitution matrix close to an integer matrix.

**Problem 3: 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  $\lambda$ , 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  $i^{\text{th}}$  column (or row) of  $Y$  must be equal to the background probability  $p_i$ . *Hint*: Consider the vector  $p = [p_1, \dots, p_n]$ . Show that  $pM = [1, \dots, 1]$ . 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  $q_{ij}$ .