

BiMM 140

Spring 2003

Quiz 2

Name \_\_\_\_\_

1. In a “search by signal” approach,  
a) (9) Give 3 examples of signals?

promoters, terminators, splice junctions, cap sites, ribosome binding sites, etc

- b) (11) Briefly describe the most common method used to detect such signals?

signals are generally represented as log-odds weight matrices which are slid along the sequence one base at a time. The score is the sum of the scores for each base of the sequence at each position of weight matrix at each position as it slides.

2. (5) .What are the most difficult features to get right in gene modeling? Why?

Very short introns/exons because they are too short to establish the long term features needed for search by content approaches.

3. (5) Do all “search by content” approaches require a set of known genes (with known correct exons) to train the method?

No. Test code or positional base preferences, for example, do not.

4. (25) What are the steps used in FASTA and where are the init1, initn, and opt scores calculated?

- 1) determine initial diagonals using lookup table
- 2) rescore all positions in top 10 diagonals using (init1 score)
- 3) connect diagonals into longer regions (initn score)
- 4) calculate optimal SW alignment in band around diagonals (opt score)
- 5) fit statistical distribution and calculate statistics

5. (25) Describe the differences in how BLAST and FASTA calculate significance? What are the advantages of each approach?

Blast calculates statistics in advance using Karlin-Altschul equation and random sequence model. This allows more efficient screening because thresholds for neighborhood matches are determined in advance. FASTA calculates statistics based on fitting the observed scores to an EVD at the last stage. The significance is based on unrelated sequences and therefore includes all of the subtle similarities that real proteins (but not random sequences) share.

6. (10) What is a log-odds scoring system and how does it relate to the Dayhoff PAM250 scoring table?  
a log-odds scoring system compares two models, the foreground and background models, which describe alternate possibilities.  $S = \log ( f / b )$ . In the case of the PAM250 matrix, the foreground model is the frequencies of the amino acid residues expected under the Dayhoff Markov model of evolution, and the background model is a random sequence model.

7. (10) What is an extreme value distribution and why is it used to evaluate the significance of alignments?  
an EVD describes the distribution of scores expected for processes that involve optimization. It resembles a normal distribution with a larger positive tail. Sequence alignments give the optimum score out of many many possibilities and thus their scores are an EVD. Correct evaluation of significance therefore requires one to use the EVD not normal distributions.