

1. (7 points each) Outline the algorithm used by

(a) the BLAST server to find protein sequences homologous to a query protein sequence

(b) MacVector to generate PCR primer pairs in two flanking regions

2. You are given two protein sequences, “rakym” and “rrsfpl”.

(a) (4 points) Use the blanks below to carry out a dynamic programming alignment using the PAM250 similarity matrix and no gap penalties.

(b) (2 points) What is the optimal alignment between the two sequences according to the assumptions above?

3. (3 points each) List the inputs to and outputs from

(a) a module to perform Chou-Fasman prediction of secondary structure

(b) a module to generate a dot matrix of protein sequences according to the method of Pustell

4. (3 points) You are given two nucleotide sequence files in MacVector format, **UnknownA** and **UnknownB** (located on the Comp. Biol. Appleshare server in the folder called "Midterm"). Generate a dot matrix plot of the two sequences using MacVector. What are your conclusions from viewing the plot?

5. (2 points each) **(a)** Generate a codon preference (codon bias) plot for the yeast actin-2 gene (EMBL accession number V01289). From this plot, are any forward reading frames likely to contain a long translated region? If so, which reading frame?

(b) Is this the same reading frame as that of the major exon of the gene? Explain.

6. (2 points each)

(a) Find the sequence of the human Rho-GAP hematopoietic protein C1 (it is thought to play a role in regulating cytoskeletal structure). What are the first five amino acids of its sequence?

(b) What chromosome is the gene located on?

(c) Find the DNA sequence that encodes the protein (Hint: Find literature articles related to the protein sequence, follow the link to the article by Tribioli, and select the link to DNA sequence). What is its EMBL accession number?

(d) Find the yeast DNA sequence with the closest homology to the human DNA sequence (Note: the homology is weak, yielding a probability of random occurrence of 0.75). What is its EMBL accession number?

(e) What chromosome is the yeast gene located on?

(f) Find the sequence of the yeast protein encoded by this DNA sequence. What are the five amino acids of its sequence?

(g) Download the 3D structure of the yeast protein. What is its PDB name?

(h) Which of the following secondary structures is more common in the protein: α -helix, β -sheet or β -turn?

(i) Display the structure using a spacefilling model and rotate it to gain an appreciation of its properties. Describe any feature or features that reflect on the protein's suspected role as part of a complex with other proteins in a signalling pathway.

Circle the correct answer (1 point each)

7. An amphiphilic protein domain (a) has a low hydrophobic moment, (b) has a high hydrophobic moment, (c) has a high content of α -helix, or (d) is rarely found in membrane-associated proteins.

Questions 8 and 9 refer to the following matrix that describes a possible eukaryotic termination signal.

	1	2	3	4	5	6	7	8
A	0.61	0.21	-1.18	-1.18	-0.49	-2.28	-0.50	0.98
C	0.39	-2.09	-2.79	-2.79	-1.40	0.90	0.61	0.61
G	-0.55	0.93	-1.32	1.08	-0.81	-1.17	-1.03	-0.55
T	0.04	-1.75	0.59	-2.85	0.47	0.22	0.12	-0.04

8. It is (a) a dot matrix, (b) similarity matrix, (c) a dynamic programming matrix, (d) a frequency matrix, (e) a weight matrix derived from a similarity matrix, (f) a weight matrix derived from a frequency matrix, or (g) a log-odds matrix derived from a dot matrix.

9. According to this matrix, a eukaryotic termination signal is least likely to have (a) a G in position 4, (b) a T in position 1, (c) a T in position 4, or (d) a C in position 6.

10. An ORF (a) always starts with an AUG and ends with a stop codon, (b) is always translated into protein, (c) can have a high codon preference score and not be translated into protein, or (d) cannot be interrupted by introns.

11. All exons in a gene that gives rise to a spliced mRNA (a) are in the same reading frame, (b) contain splice donor, acceptor and branch point sites that match well with consensus sequences, (c) contain splice donor, acceptor and branch point sites that match at least to a statistically-significant degree to consensus sequences, or (d) can contain stop codons in at least one reading frame.

12. You can display three-dimensional models of proteins whose sequence is similar to a previously-unknown protein that you have cloned and sequenced by using only (a) the BLAST server and MacVector, (b) the Entrez VAST neighbors feature and RasMol, (c) the BLAST server and RasMol, or (d) AssemblyLIGN and RasMol.

13. Only the backbone of a protein is displayed using (a) a ball-and-stick model, (b) a space-filling model, (c) a ribbon model, or (d) a wireframe model.

14. To obtain the best possible alignment between two previously-unknown proteins, we can use (a) MacVector, (b) the BLAST server, (c) the Entrez server, (e) RasMol, or (e) AssemblyLIGN.

15. Taking into account the observed statistics of occurrence of dinucleotides in a sequence, the probability of occurrence of the trinucleotide AYR in that sequence is

(a) $p_A(p_{AC} + p_{AT})(p_{CA} + p_{CG})$, (b) $p_{AC}p_{CA} + p_{AC}p_{CG} + p_{AT}p_{TA} + p_{AT}p_{TG}$
 (c) $p_A(p_{AC}(p_{CA} + p_{CG}) + p_{AT}(p_{TA} + p_{TG}))$ (d) $f_{AC}(f_{CA} + f_{CG}) + f_{AT}(f_{TA} + f_{TG})$