

**Problem A2**  
**BLAST**  
**Due: February 9, 2006**

In this assignment, you are given a sequence of a mouse cDNA in the file **ProbA2.txt** on the homeworks web page. Your goal is perform initial sequence analysis and comparison to find similar genes and proteins.

Hand in requested printouts and written answers to the questions. Include justification for your answers (i.e., show basis of calculations). Label your answers clearly.

**Questions (Total of 40 points)**

1. a) Which dinucleotide(s) is(are) the least frequent in the cDNA?  
  
    (b) What is the predicted frequency of the tetranucleotide CCAT using the observed *mononucleotide* frequencies (show your calculations)?  
  
    (c) What is the predicted frequency of CCAT using the observed *dinucleotide* frequencies (show your calculations)?
2. (a) What locations in the sequence match the consensus sequence "TRGCYA"?  
  
    (b) Using the observed mononucleotide frequencies, what is the expected number of occurrences of this consensus sequence in the cDNA?
3. Do a BLAST Search (use blastn) either from within MacVector or using the BLAST web page (<http://www.ncbi.nlm.nih.gov/BLAST/>) to find related nucleotide sequences.
  - a) Print and submit the first two pages of the BLAST results.
  - b) Is the cDNA sequence present in the Genbank database? If so, what is the name of the protein encoded by it?
4. Find the protein sequence that this cDNA encodes and also find protein sequences homologous to this protein. Using PubMed, read the abstracts for some of the papers linked to the protein sequences homologous to the ProbA2 protein sequence. Briefly describe the conclusions you reached from interpreting the abstracts and the results of the BLAST searches. For example, what proteins are related to the protein encoded by the ProbA2 cDNA? Are there any distinguishing features of these proteins?

**Extra credit (5 points)**

5. Using the observed dinucleotide frequencies for ProbA2.txt cDNA, what is the predicted frequency of occurrence of the tetranucleotide CWSG (show calculations)?

6. Using the BLOSUM62 matrix, which is more likely for an average protein: that an arginine residue could be replaced with a lysine or that a tryptophan could be replaced by a phenylalanine? Would the answer be different using the PAM250 matrix?