BioE/MCB/PMB C146/246, Spring 2005

Problem Set 7

Due 11 April 2005 by midnight to lfl@compbio.berkeley.edu. Turn in a hard copy in class.

1. 20 points

Given the following multiple sequence alignment:

```
MMKE
MKKE
IKIE
MEME
MKME
IMKI
MKME
IKKE
MKME
IKKE
```

Build a profile from this alignment using the PSI-BLAST method. Use the BLOSUM62 matrix and $\beta = 1$.

2. 10 points

Transmembrane proteins can be modeled as sequences with stretches of hydrophobic residues and stretches of unrestricted residues. Draw a hidden Markov model to represent this, and write out the recursion relationships to find the best hidden state path for a given sequence.

2. 30 points

Write a program that generates random DNA sequences of chosen length *N* defined by a two-component mixture model. The parameters are the width of the motif w, a position weight matrix θ (4 rows x w columns) describing the motif, a column vector θ_0 describing the probability of finding a given base in background sequence, and a parameter λ describing the probability that a motif begins at a given base. At each position in your generated sequence, use λ to decide whether to emit a motif of width w or to emit a background base. Document and submit your code.

3. 40 points

Choose a transcription factor from http://jaspar.cgb.ki.se/cgi-bin/jaspar_db.pl and download its weight matrix. Plug the matrix values into θ in your program above, and using uniform background values θ_0, generate 25 sequences of length 1000 with $\lambda = 0.1$. Repeat with $\lambda = 0.01$.

Using the web version of MEME at http://meme.sdsc.edu or a local implementation, run MEME (using the two-component mixture model) on these sequences.

(A) (points) Compare the MEME output matrices with the starting matrix from JASPAR. How do the results change with the different values of λ? Include your starting matrix and the MEME matrices in your answer.

(B) (points) For $\lambda = 0.01$, what is the log likelihood of the background model, given the sequences? Show the formulas you use to compute the likelihood.

(C) (points) Report the number of binding sites that MEME detects in the sequences generated for $\lambda = 0.01$. Assuming that all the binding sites were accurately detected, evaluate the log likelihood that these 25 sequences were generated by the mixture model. Show the formulas you use to compute the likelihood.