Algorithms for Molecular Biology
© Ron Shamir
School of Computer Science
Tel Aviv University
Fall 2001
December 27, 2001

Algorithms for Molecular Biology: Programming Assignment
due on January 24, 2002

1 HMM Profiling

General Guidelines: This assignment is part of your test in the course. You may team up in pairs but teams should work **INDEPENDENTLY**. Please submit your work by sending an e-mail to `roded@post.tau.ac.il` with a compressed archive attachment, containing:

1. All of your source code.
2. Instructions how to compile it (i.e. `Makefile` on unix)
3. I/O files for test runs.
4. Any intermediate files or tables your application needs.
5. Time/memory requirements of your application.
6. A short README file, listing what is each of your files, and with your names and IDs.

In your application, pay attention to:

1. Creating something practical, i.e., a program that runs within minutes using the memory of a standard PC, and avoiding problems like float precision.
2. Writing readable code. Use only C/C++/Java.
3. The code should compile and will be tested on Unix.

Credit: 20/75. Exceptional work will be given extra credit.

Description: In this project you will create a program which learns HMMs for protein families, uses these HMMs for identifying family members, learns subfamilies within the families, and performs a multiple alignment of sequences to a given profile.

You will work on two protein datasets: globins and sodium-channel scorpion toxins. Both datasets can be obtained by searching for appropriate keywords using Entrez. For each dataset collect as many members as you can, and divide them randomly into two groups - a training set containing 80% of the sequences and a test set containing the rest 20%. Also prepare a list of 10,000 arbitrarily selected Swiss-Prot proteins, including the test sets in this list. This list is referred below as *db*.

1. Write an application in which you implement the following options:
   (a) Train an HMM for each family. Use at least one strategy for avoiding local maxima.
   (b) Score each entry in the db for membership in each family (using the HMMs saved). Record the number of false positives and false negatives you obtain.
   (c) Devise a new HMM for each family which will sub-divide it into k sub-families. Play with $k = 4, 5, 6, 7$. What are the characteristics of the proteins in each sub-family? (Hint: The new HMM should contain k copies of the old one, and one new state.)
   (d) Perform a profile alignment for the whole family and for each sub-family. Compare the obtained multiple-alignment of the whole family to any relevant alignment from the literature.
   (e) Check the sub-family classification on the globin set by comparing the obtained sub-families to the major known ones - alpha, beta and myoglobin (use text search to obtain the known classification).
2. Submit documentation of the resulting HMMs in each stage, the multi-alignments obtained and the classification results.
3. Your program should support the following I/O:
   - INPUT: A file of sequences in FASTA format (first line - name, second line - sequence), and a parameter $k \geq 1$.
   - OUTPUT: A k-sub-family HMM along with k profile alignments for each sub-family. Any output sequence should be accompanied by its name.