CENG 465 – Introduction to Bioinformatics
Spring 2008-2009

Assignment #3 (Programming Assignment)
Generating HMMs from Multiple Sequence Alignments and scoring a sequence against an HMM using the Forward algorithm

Due Date: April 29, 2009, 11:59PM

Programming Assignment about Hidden Markov Models

In this assignment, you are going to implement the simple technique we have discussed in class to generate profile Hidden Markov Models from multiple sequence alignments. (The technique is explained in slides 21-30 of Week #5 lecture notes and also on pages 185-187 of the textbook). After generating the HMM for a given multiple sequence alignment you will score each sequence of the respective alignment against the HMM and find the sequence which is most distant (i.e., divergent) in the protein family. This sequence is also called an "orphan" sequence if it is highly diverged from the rest of the family members. To compute the sequence vs. HMM score, you will implement the Forward algorithm discussed in class (slides 40-41 of Week #5 lecture notes and on pages 187-191 in the textbook). Alternatively, instead of implementing the Forward algorithm, you may output your HMMs in SAM[1] or HMMER[2] format and use SAM or HMMER tools to score your sequence against the profile HMM you generated. Below are the step by step details of the assignment.

1) You are going to retrieve the input multiple sequence alignments from the BAliBASE 2.0 benchmark. You will use the "short" subset of the Reference 2 alignments at: http://bips.u-strasbg.fr/fr/Products/Databases/BAliBASE2/align_index.html#ref2 . There are 9 multiple sequence alignments in that set. For example the first alignment is the multiple alignment of the SH3 family and identified by the first sequence in the alignment: 1aboA. There are 15 sequences in the first alignment. The alignment can be viewed at: http://bips.u-strasbg.fr/fr/Products/Databases/BAliBASE2/ref2/test/1aboA_ref2.html or downloaded in RSF and MSF formats (if you want to use these formats you need to learn about them by yourself).
2) In the alignments, there are certain columns which are displayed as underlined capital letters. These are "core blocks" of the alignment. In other words, these are regions that are conserved well in that protein family. These core blocks will be the

[1] http://compbio.soe.ucsc.edu/sam.html
[2] http://hmmer.janelia.org/

"match states" of the profile HMM you will generate. In other words, if the total number of underlined columns is *n*, you will have *n* match states in your HMM. For example, for the 1aboA alignment, there will be 36 match states in the HMM that you generate. (Note that the column which mostly contains gaps in the final core block is also considered a match state).

3) Using the technique described in class, you will generate a profile HMM for each of the 9 reference alignments. Use the simple global profile HMM model (with no support for flanking) which is given in Figure 6.7 (A) of the textbook. Compute the emission probabilities using pseudocounts (EQ 6.27 in the textbook).
4) The Reference 2 alignment set of BAliBASE2 contains protein families aligned with a highly divergent "orphan" sequence. In other words, one of the sequences in the alignment is not so similar to the rest of the sequences. Find the "orphan" sequence by computing the probability of each sequence to be emitted by the profile HMM that you have generated. The sequence with the lowest probability should be the "orphan" sequence. (However, note that since the reference MSA is manually created to match all the core blocks in all of the sequences, the "orphan" sequence may look similar and score similar to the other sequences. It is OK that if the sequence you report is not the real orphan sequence). Implement the Forward algorithm described in class to compute the probability of a sequence to be emitted from a given HMM. Instead of actual probabilities you may compute the log-odds score (EQ 6.32 in the textbook).
5) Instead of implementing the Forward algorithm you are free to use SAM or HMMER tools to score a sequence against an HMM. However, note that, you need to output your models in a format that is compatible with SAM or HMMER. Refer to the web sites of these tools to learn more about profile HMM formats that they use.

In your report you should give the orphan sequences you have found for each of the 9 reference alignments in the "short" Reference 2 subset of BAliBASE2.

You may use your own judgment for any issue that is not specified clearly in this text.

Deliverables:

- The source code of your program(s). You may use any programming language.
- A short report which contains a step by step description of the tasks that you have performed, your results, and your interpretation of the results.

Submission:
Submit the deliverables as a zip bundle or as a tarball using the COW system.

Late Submission Policy:
Your final assignment grade will be penalized 20 points per late day.

CHECK THE NEWSGROUP REGULARLY FOR POSSIBLE UPDATES ON THE ASSIGNMENT.