Due Date: Monday, April 25, 2011, 11:55PM

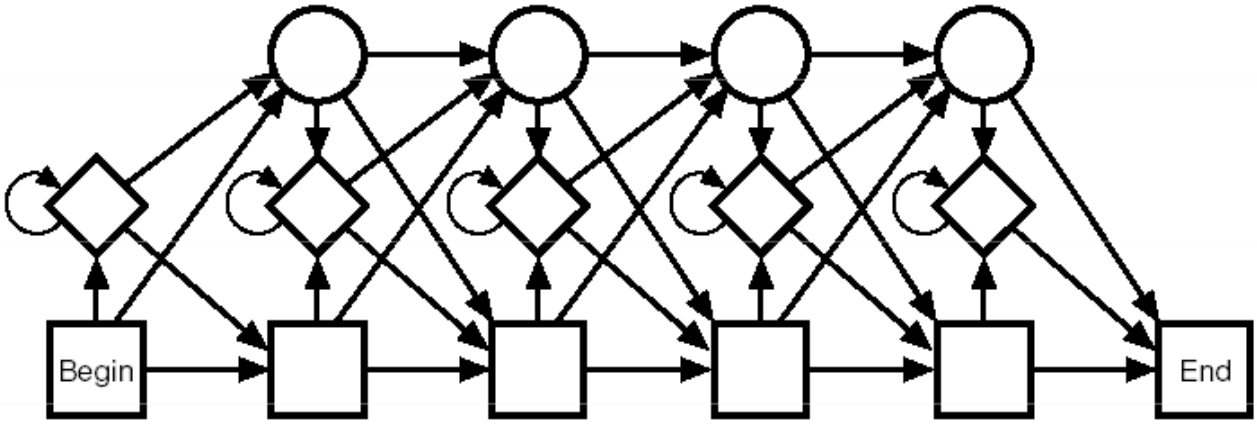
CENG 465
Spring 2010-2011

Assignment #2

Part 1 of Programming Assignment on profile HMMs
Construction of profile HMMs from MSAs

Construction of a Profile Hidden Markov Model from a Multiple Sequence Alignment

In this assignment, you are going to implement a program to build a profile hidden Markov model given a multiple sequence alignment of a family of sequences. You will use the following state structure for the profile hidden Markov model.



In this structure if there are ***N*** *match* states, there are ***N***+1 *insert* states and ***N*** *delete* states in addition to the Begin end End states. So, a profile HMM with ***N*** match states will have 3****N***+3 states in total. Some of the *insert* or *delete* states may not be visited at all in a given multiple sequence alignment, so these unvisited states may be deleted from the final structure.

For this assignment, your program will take a multiple sequence alignment of a family of protein sequences as input, construct a profile HMM for these sequences using the following steps:

1. Determine the number of match states by finding the columns of the alignment which correspond to conserved regions,
2. Find the emission probabilities at match states by computing the frequencies of amino acids at match columns
3. Find the transition probabilities between the states of the profile HMM.
4. Delete the states in the standard profile HMM structure which are not visited by the given MSA

In this assignment, you will NOT use pseudocounts when determining the emission probabilities.

A column of a multiple sequence alignment is considered a conserved column (i.e., match state) if the majority of that column is non-gap (i.e., <0.50 gap frequency) AND at least half of the sequences contain the same amino acid type in that column (i.e., the highest frequency amino acid at that column has frequency >= 0.50).

The states of the profile HMM are named as the following:

BEGIN, I0, M1, D1, I1, M2, D2, I2, ….., M_N, D_N, I_N, END.

The input multiple sequence alignment for ***K*** protein sequences will be given as a file with ***K*** lines each line containing the same number of characters. The gaps are indicated by a dash (‘-‘) and the amino acids are indicated by one letter amino acid codes in uppercase.

After constructing the profile HMM you will output your profile HMM in an output file with the following format:

```
M<tab><number of match states>
S<tab><total number of states with unvisited states removed from the HMM>

#Emission probabilities
M1<tab><number of amino acid types emitted at match state M1>
<amino acid type 1><tab><emission probability for amino acid type 1>
<amino acid type 2><tab><emission probability for amino acid type 1>
…
<amino acid type x><tab><emission probability for amino acid type x>
…
…
MN<tab><number of amino acid types emitted at match state MN>
<amino acid type 1><tab><emission probability for amino acid type 1>
<amino acid type 2><tab><emission probability for amino acid type 1>
…
<amino acid type x><tab><emission probability for amino acid type x>

#Transition Probabilities (only outgoing arcs are reported)
BEGIN<tab><number of outgoing arcs>
<destination state of outgoing arc 1><tab><transition probability>
<destination state of outgoing arc 2><tab><transition probability>
…
<destination state of outgoing arc y><tab><transition probability>

<State Name><tab><number of outgoing arcs>
<destination state of outgoing arc 1><tab><transition probability>
<destination state of outgoing arc 2><tab><transition probability>
…
<destination state of outgoing arc y><tab><transition probability>
```

The example DNA sequence alignment given on lecture slides and the corresponding output is given below:

```
A C A - - - A T G
T C A A C T A T C
A C A C - - A G C
A G A - - - A T C
A C C G - - A T C
```

```
M     6
S     9
#Emission Probabilities
M1    2
T     0.2
A     0.8
M2    2
G     0.2
C     0.8
M3    2
C     0.2
A     0.8
M4    1
A     1.0
M5    2
T     0.8
G     0.2
M6    2
G     0.2
C     0.8
#Transition Probabilities (only outgoing arcs are reported)
BEGIN 1
M1    1.0

M1    1
M2    1.0

M2    1
M3    1.0

M3    2
I3    0.6
M4    0.4

I3    2
I3    0.4
M4    0.6

M4    1
M5    1.0

M5    1
M6    1.0

M6    1
END   1.0
```

When outputting emission or transmission probabilities, the ordering of amino acids or the ordering of states is insignificant. You may output them in any order in the output file.

A more realistic example and its output is provided at the following links:

http://www.ceng.metu.edu.tr/~tcan/ceng465_s1011/Assignments/hw2_example.txt

http://www.ceng.metu.edu.tr/~tcan/ceng465_s1011/Assignments/hw2_example_hmm.txt

Test your program with the test data provided at:

http://www.ceng.metu.edu.tr/~tcan/ceng465_s1011/Assignments/hw2_test_data.zip

In the zip file provided at the above URL, there are four multiple sequence alignments for four protein families.

Submit the results of your program for the test data as a report (.txt, .doc, or .pdf format) along with your source code.

Hint:

When computing the transition probabilities process the MSA sequence by sequence (i.e., row by row) and find the states visited by each sequence. Then use these visited states to compute the transition probabilities.

Deliverables:

1. The source code of your program. You may use any programming language of your choice.
2. A short report containing the resulting profile HMMs for the test data.

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.