Due Date: Monday, May 9, 2011, 11:55PM

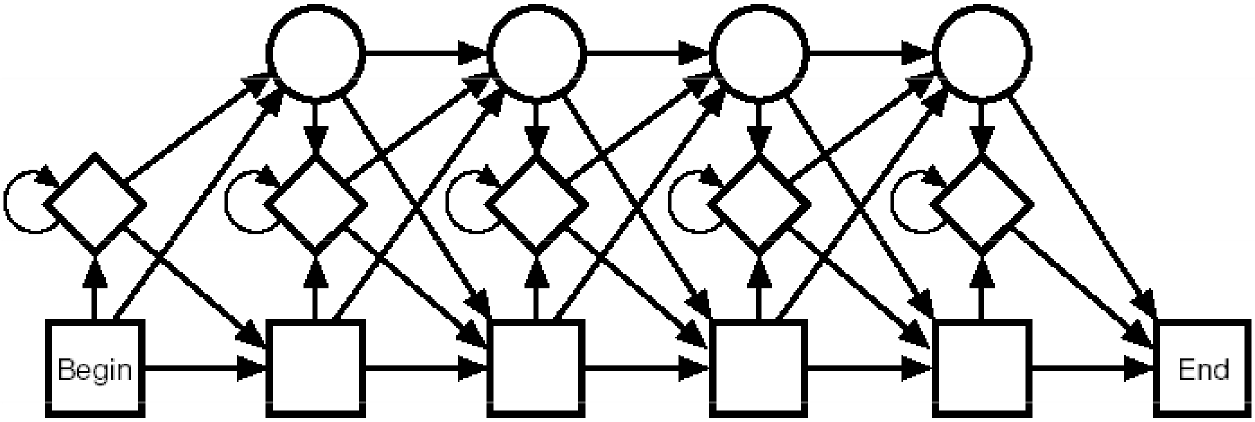
CENG 465
Spring 2010-2011

Assignment #3

Part 1 of Programming Assignment on profile HMMs
Implementing the Forward Algorithm

Scoring a Sequence against a profile HMM using the Forward Algorithm

In this assignment, you are going to implement a program to score a protein sequence against profile hidden Markov model. The score is going to be the overall probability that the profile HMM has generated the given sequence. This probability will be computed using the Forward algorithm as a summation over all possible paths that could have generated that sequence. You will compute this probability by filling in a Dynamic Programming partial probability table. You will assume that the input profile HMM has the same profile hidden Markov model structure used in Assignment #2.



For this assignment, your program will take a profile HMM of a family of protein sequences and a single protein sequence as two separate inputs, and compute the probability that the profile HMM generated the protein sequence. The format of the profile HMM file is the same as the output format specified in Assignment #2.

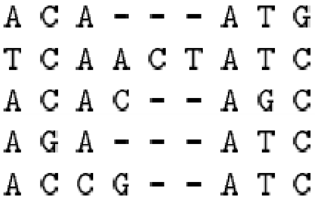
```
M<tab><number of match states>
S<tab><total number of states with unvisited states removed from the HMM>

#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 input protein sequence is provided as a file with a single line containing the protein sequence with upper case one letter amino acid codes. Your program should output a single floating point number which is the probability found by the forward algorithm.

The emission probabilities at Insert states are same for each amino acid type, i.e., 1/20.

For example for the following profile HMM and the DNA sequence given as input (emission at Insert states are ¼ for this example):



```
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
```

Profile HMM:

```
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
```

Input sequence:

```
AGCCATG
```

The output of your program should be 4.608E-4.

For the example profile HMM provided at:

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

and the following protein sequence as input:

```
GIHPKMISDLQVIPAGPQCSKAEVIATLKNGKEVCL
```

Your output should be 2.0411134874896203E-31.

After you write your program test all of the following sequences against the profile HMMs of four protein families provided as test cases for Assignment #2.

http://www.ceng.metu.edu.tr/~tcan/ceng465_s1011/Assignments/hw3_test_sequences.zip http://www.ceng.metu.edu.tr/~tcan/ceng465_s1011/Assignments/hw3_profile_HMMs.zip

(The profile HMMs are going to be provided after May 3 because of possible late submissions to Assignment #2. However, you may start coding your Assignment #3 as soon as possible, and test it by using the hw2_example_hmm.txt)

After testing all 10 input sequences against all 4 profile HMMs, you will obtain a probability matrix of size 4x10. Each input sequence is going to be assigned to the family which results in the highest probability for that sequence.

Submit your 4x10 probability matrix and the family id assigned to each sequence as a report (.txt, .doc, or .pdf format) along with your source code.

NOTE:

Since you do not use pseudocounts in the profile HMM, it is possible that some of the sequences will have 0.0 probability to be emitted from a given profile HMM. If a sequence gets 0.0 probability for all the four profile HMMs you will not be able to predict its family.

Deliverables:

1. The source code of your program. You may use any programming language of your choice.
2. A short report containing your results described above.

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.