Due Date: November 29, 2016 (23:55)
Tuesday

CENG 465
Introduction to Bioinformatics
Fall 2016-2017

Assignment #2

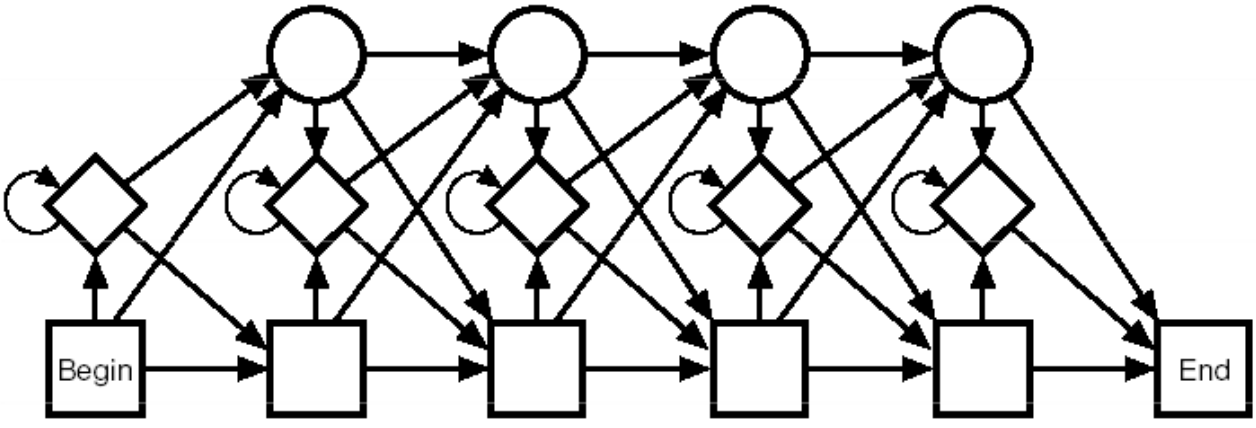
Profile Hidden Markov Models

Consider the following multiple alignment of a family of protein sequences.

```
---SHPTYSEMIAAAIRAEKSRGGSS-----V
---AHPSSSEMVLAAITALKERGGSS--NYTV
---AHPPVIDMITAAIAAQKERRGSS------
---AHPPVATM---AILGLKERKGSSAANYRV
--------TELIVKAVSSSKERSGVSA---YD
---SHPPTLSMVVEVLKKNTERKGTSLS--PT
---THPPYIEMVKDAITTLKERNGSS------
RGSALSDTERAQLDVMKLLN----VSS-----
-MRSSAKQEELVKAFKALLK--KFSSQEQ-GF
----KPSTLSMIVAAITAMK-RKGSSL---KG
    **   **   **   *** ***
```

Conserverd colums, which correspond to match states, are indicated by an asterisk at the bottom of the alignment.

Use the following profile hidden Markov model structure to construct a profile HMM for these sequences.



Your pHMM will have **12** *match* states, **13** *insert* states and **12** *delete* states in addition to the Begin end End states. Some of the *insert* or *delete* states may not be visited at all in your pHMM, so these unvisited states may be deleted from the final structure. When constructing the pHMM, find the emission probabilities at match states by computing the frequencies of amino acids at

match columns. Do not use pseudocounts. Use 1/20 as the emission probabilities of all amino acids at insertion states. Determine the transition probabilities between the states of the profile HMM by using the sequence of states visited by the protein sequence in the given multiple alignment.

After you construct the profile HMM for these sequences, determine the most likely sequence of states visited by the following sequence:

```
HPSWTEMEDRAVYQAKRLGNS
```

Note that this sequence is a new sequence which is not in the alignment. So, you will have to use the Viterbi algorithm to align this sequence to your constructed pHMM.

As your solution, report the sequence of states visitied by this query protein seuqnece and also the probabilty associated with these sequence of states as found by the Viterbi algorithm, i.e. ($V_{End}(\text{“S”})$). You may indicate the probability in scientific notation, e.g. 0.123e-30. Also indicate which tool you used to solve this problem, e.g. Excel, Matlab, R, C, Java, or manually with a calculator (!).

Submission

Submit your solutions as a single PDF/DOC/ODT document (scanned copies of handwritten solutions are accepted as well) via ODTU-Class before the deadline. Late submission is -20 pts per day.