CENG 465
Spring 2020-2021

Due Date: May 2, 2021, 23:59 via ODTU Class

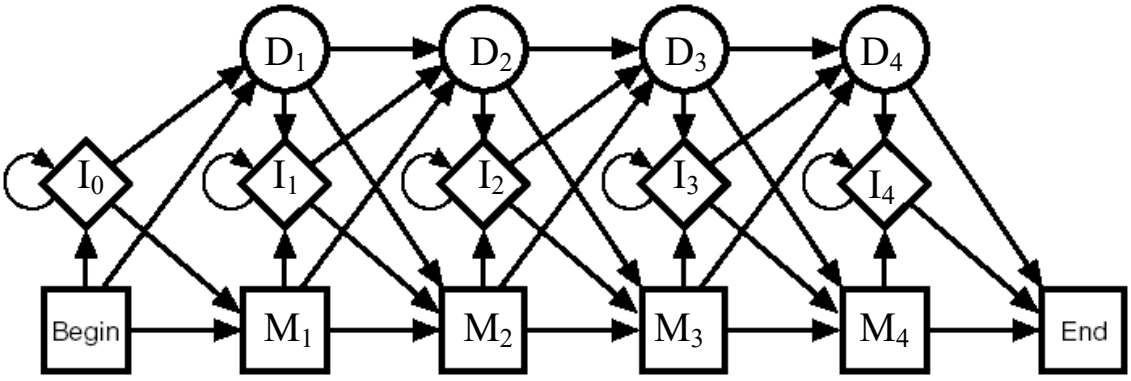
Assignment #2

Written Assignment

In this assignment, your goal is to construct a profile Hidden Markov Model (pHMM) for the following 8 DNA sequences using their multiple alignment given below and then align a new sequence to the constructed model using the Viterbi algorithm.

```
--TT--GA-G—
TATT--CAC-A
-CATGGCA-—-
GC-T---ACGA
--GT--CT-—-
AC-TA-CA--A
GC-GT-CT-—-
-C-T--CA-G—
 *  *   **
```

Part A: Use the following pHMM structure template. The columns with an * at the bottom indicate the conserved columns, which should be modeled by match states. You may omit some insert/delete states of the pHMM if they are not visited by any of the 8 sequences. Give the emission probabilities at match states and the transition probabilities between applicable states. Emission probabilities at insertion states are 1/4 for all nucleotides. Do not use pseudocounts when computing the emission or transition probabilities.



Part B: Align the following sequence to the profile HMM using the Viterbi algorithm. In other words, find the sequence of states which is most likely to emit that sequence. Show the contents of the partial probability table you construct. What is the probability associated with the best path?

AATGAC

Note: Initialize the partial probability (or likelihood) table with $v_{Begin}(\text{“”})=1.0$. Do not use log of probabilites. During multiplications you may use scientific notation and round to 2 digit after the decimal point to represent small numbers, e.g., 3.42E-7 or 2.07E-3 to indicate 0.000000342321 and 0.00206784, respectively. Use the rounded numbers in the subsequent computations to make your computations easier.

Submission:

Submit your solution as a single PDF document via ODTU-Class before the deadline. Your solution may be a scanned copy of a handwritten solution or a document written on computer. Late submissions will be penalized 15 points per day.