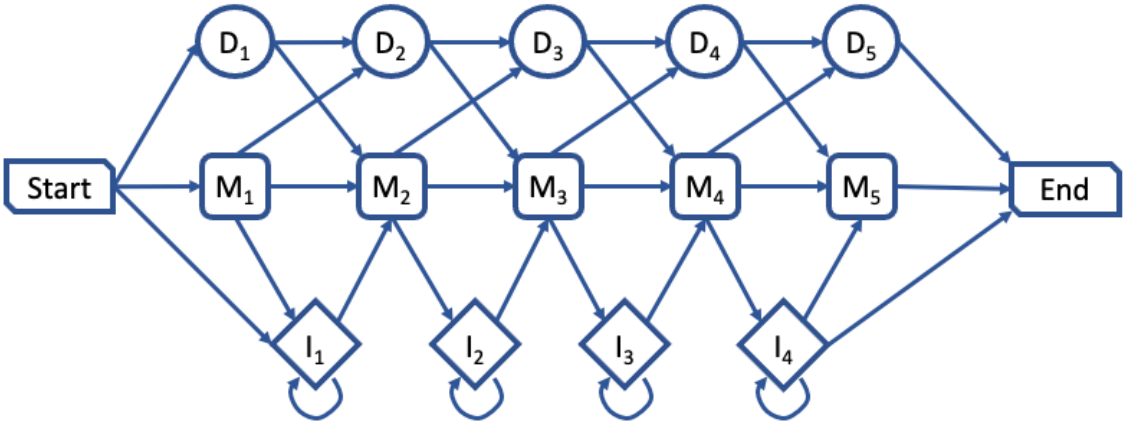
Profile HMMs for protein sequence alignment

We have a hidden Markov model (HMM) that describes the alignment of five positions in the sequences of protein family A



Seq1	H N Y - H S
Seq2	H H Y - H G
Seq3	N H Y - - S
Seq4	T N Y g F S
Seq5	N G Y - H G
Seq6	H N Y - F S
Seq7	H N Y - H S
Seq8	N H Y – G G
Seq9	H H Y - G S
Seq10	H N Y g F S

Every match (M_i) and insertion (I_i) has emission probabilities for 20 amino acids (though some of these probabilities may be zero). The deletion states (D_i) do not emit observations and are called 'silent' or 'non-emitting'.

Additionally, we have the ten protein sequences that were aligned to train this HMM. Note that though the data contains six positions for amino acid residues, the HMM describes five match states, and we encode additional residues as insertions. It would be possible to create an HMM describing six match states with missing residues encoded as deletions. In practice, the rule of thumb is to take the average of the length of the training sequences.

Exercises

Calculations

1.1 Parameterize the HMM by calculating the starting probabilities π and the matrices A and B describing the transition and emission probabilities, respectively.
Hints:
(1) A reasonable estimate for π is the frequency of each amino acid in the training data.
(2) The transition probabilities can be estimated by counting the types of transitions in the training data. A similar approach can be taken to estimate the emission probabilities.

1.2 Use the Viterbi algorithm to infer the hidden state sequences that emitted Seq1, Seq3, and Seq4. For 1.2 and 1.3, the probabilities you get may be quite small. That is expected and in practice, sequence alignment of proteins using HMMs involves a log odds formulation to prevent underflow issues in computing.

1.3 Calculate the probability of the emitted sequences Seq2, Seq3, and Seq10.

Bioinformatics

HMMER is a popular biosequence analysis tool. It uses hidden Markov models to provide insights on protein sequence input such as homolog detection and protein family identification. HMMER is frequently used with profile databases such as Pfam for protein family matching. HMMER is available for installation as a command line function, but can also be accessed on a web interface. This exercise will use the web version of HMMER.

HMMER is hosted at https://www.ebi.ac.uk/Tools/hmmer/
HMMER was recently described in the *Nucleic Acids Research* article, *HMMER web server: 2018 update* https://nar.oxfordjournals.org/content/46/W1/W200

You have isolated and sequenced a fragment of protein from a mystery tissue sample and want to learn more about both the protein and the tissue sample.

2.1 Use HMMER to search this protein sequence

VQFKSKMPIQPDLLVFLFSNDVSAPVRCQWNFHLSVEPLTANNAKMRESLLRS

How many significant matches did HMMER produce? What does HMMER define as a 'significant result'? What is the interpretation of the E-value? Some target matches will have come from the same species; why would this be?

2.2 Identify the protein that produced this fragment. Describe its function.

2.3 Did any Pfam matches come up? Why or why not?

2.4 Based on your assessment of the match evidence, what is the most likely species that this tissue sample is from?