ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

D-BSSE
Department of Biosystems
Science and Engineering

Statistical Models in Computational Biology

Jack Kuipers
David Dreifuss
Xiang Ge Luo
Rudolf Schill

Due date: 23th March 2023

Membership detection with profile HMMs

The aim of this project is to parametrise profile hidden Markov models for two protein families, and use the models to determine the family membership of unclassified protein sequences. For this you will use the forward algorithm to determine the probability of a sequence being related to a given protein family irrespective of the alignment.

	M	M	I	I	I	M
bat	A	G	-	-	-	C
rat	A	-	A	G	-	C
cat	A	G	-	-	-	C
gnat	-	G	A	A	A	C
goat	A	G	-	-	A	C
	1	2	.	.	.	3

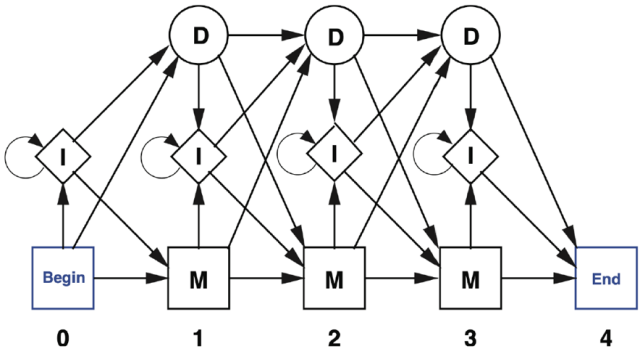

Figure 1: Example of a profile HMM (adapted from Figure 5.7 of [1]). The multiple alignment is shown on the left. Columns with more than 50% gaps (–) are considered insert states with respect to the consensus. The rest are considered consensus match states, which yields a model length of 3. The corresponding model is shown on the right, where 'M' denotes 'Match' (square), 'I' denotes 'Insertion' (diamond), and 'D' denotes 'Deletion' (circle).

Problem 8: Estimating match emission probabilities **(1 points)**

Figure 1 shows an example of a profile HMM for a multiple alignment of five DNA sequences. Columns in the multiple alignment with less than 50% gaps (–) are considered match states (labelled by 1, 2, 3). The rest are considered insert states (labelled by dots). We can estimate the match emission probability of symbol $a \in \mathcal{A} = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$ at position $i \in \{1, 2, 3\}$ by

$$e_i(a) = \frac{E_i(a) + 1}{\sum_{a' \in \mathcal{A}} \big(E_i(a') + 1\big)}$$

where $E_i(a)$ is the number of observations of the symbol a at position i across all sequences in the multiple alignment. We have added the pseudo-count 1 in order to avoid estimating some probabilities as 0.

What are the estimated match emission probabilities of the profile HMM in Figure 1?

Problem 9: Estimating insert emission probabilities **(2 point)**

Columns in the multiple alignment with more than 50% gaps (–) are considered to be insert states. Symbols appearing whilst in an insert state are considered insert emissions. Note that contiguous insert states have the same position in the model (see figure).

What are the estimated insert emission probabilities of the profile HMM in Figure 1?

Problem 10: Estimating transition probabilities **(3 points)**

Let $\mathcal{S} = \{\mathrm{M}, \mathrm{I}, \mathrm{D}\}$ denote the states 'Match', 'Insertion', and 'Deletion', respectively. The transition probability from state $k \in \mathcal{S}$ at position $i \in \{0, 1, 2, 3\}$ to state $\ell \in \mathcal{S}$ at position $i + 1$ can be estimated by

$$t_i(k \to \ell) = \frac{T_i(k \to \ell) + 1}{\sum_{\ell' \in \mathcal{S}} \left(T_i(k \to \ell') + 1\right)}$$

where $T_i(k \to \ell)$ is the observed number of transitions from state k at position i to state ℓ at position $i + 1$. We adopt three conventions:

- A profile HMM is assumed to begin and end in match states
- In marked columns (M), gaps are assigned to delete states (i.e., a gap is the result of a deletion)
- In unmarked columns (I), symbols are assigned to insert states, and gaps are *ignored*

What are the estimated transmission probabilities in the profile HMM in Figure 1?

Hint: For each sequence (bat, rat, etc.), draw a diagram of its path through the profile HMM from beginning to end.

Problem 11: Protein family membership classification **(4 points)**

You are given multiple alignments of protein sequences for two protein families: GTP binding proteins, and a family of ATPases. The task is to determine to which family certain unclassified proteins belong.

1. Run `source("profileHMM.R")` to import functions which you will use below.
2. Read the two alignments 'GTP_binding_proteins.txt' and 'ATPases.txt' into memory using the function `parseAlignment()`.
3. Use the function `learnHMM()` to parametrise two profile HMMs: one for each protein family (multiple alignment).
4. Identify the position(s) with the highest match and with the highest insert emission frequencies over all symbols. Plot the respective match and insert emission frequencies for the identified positions. (1 point)
5. The file `Unclassified_proteins.txt` contains 31 protein sequences from unknown families. Load the protein sequences into a list using the `parseProteins()` function.
6. The function `forward()` takes as input a profile HMM $\mathcal{M}$ and a sequence x. It returns the log odds ratio

$$\log \frac{P(x \mid \mathcal{M})}{P(x \mid \mathcal{R})}$$

of the probability of observing the sequence x given the model $\mathcal{M}$ versus the probability of observing the sequence x given the random model $\mathcal{R}$. For each unclassified protein $x^{(i)}$ in the list, apply the forward algorithm for both models M_1 and M_2 to obtain the log odds ratio

$$q(x^{(i)}) := \log\left(\frac{P(x^{(i)} \mid M_1)}{P(x^{(i)} \mid M_2)}\right).$$

Plot the values $q(x^i)$ and include this in your report. Which proteins in the list belong to which family? Can you clearly decide for each protein? (3 points)

References

[1] Durbin, R. et al. *Biological Sequence Analysis*. Cambridge University Press, 1998, https://doi.org/10.1017/CBO9780511790492