Protein HMM

Danilo Horta

August 2019

1 Model

Definition 1.1. Let $\mathcal{A}$ be the alphabet of amino acids. Let $Q_1, Q_2, \dots$ be a Markov process with (possibly) silent states and let $S_1, S_2, \dots$ be a stochastic process for which $p(S_i \in \mathcal{A} \mid Q_1 = q_1, Q_2 = q_2, \dots) = p(S_i \in \mathcal{A} \mid Q_t = q_t)$ for some t. The pair (Q_t, S_i) is a hidden Markov model with (possibly) silent states (HMM) and alphabet $\mathcal{A}$.

Let (Q_t, S_i) be an amino acid HMM with alphabet $\mathcal{A}$. We want to replace it with a HMM that generates sequences of symbols from the alphabet $\mathcal{B} = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$ of DNA bases and is able to account for frame-shifting. Let M_j be the so-called match state of an amino acid HMM and let $Q_t = \mathrm{M}_j$. From the amino acid emission probabilities and any other relevant source of information (codon usage bias, for instance), one can define the probability $p(X_1 = x_1, X_2 = x_2, X_3 = x_3 \mid Q_t = \mathrm{M}_j)$ of M_j emitting the codon $(x_1, x_2, x_3) \in \mathcal{B}^3$ — one can also write $p(X = x_1x_2x_3 \mid Q_t = \mathrm{M}_j)$, for short. Since measurement errors occur and nature is not perfect, we will replace the codon emission process by one that instead produces base sequences of different lengths to account for base insertions and deletions (indels).

Node M_j in Fig. 2 represents the modified match state. The generated codon will go through four transitions, each one representing one of three possibilities: (i) delete a base; (ii) insert a base; or (iii) do nothing. The deletion can happen in any of the three codon positions with equal probability. If a deletion has already happened, the next deletion can happen in any of the remaining two positions with equal probability. The insertion can happen between any two bases, before the first base or after the last base with equal probability.

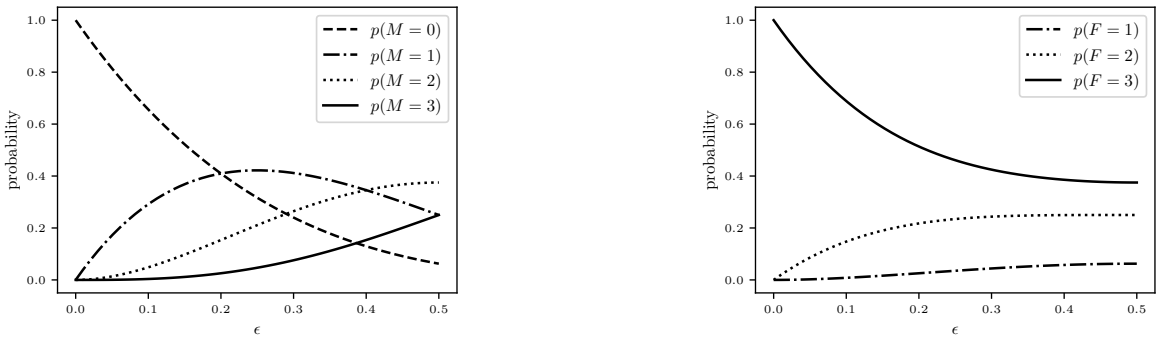
The codon emitted at node M_j can go under $m \in \{0, 1, 2, 3, 4\}$ base indels during the state transitions that end at some leaf-node state. The probability of it undergoing m indels is given by

$$p(M = m) = \binom{4}{4-m}(1-\epsilon)^{4-m}\epsilon^m,$$

where coefficient $\binom{4}{m}$ counts the number of paths corresponding to m base indels. Fig. 1a shows the base indel distributions over different values of ϵ. Let F be a random variable representing the final sequence length generated by the model in Fig. 2. We have the probabilities

$$\begin{aligned} p(F = 1) = p(F = 5) &= \epsilon^2(1-\epsilon)^2, \\ p(F = 2) = p(F = 4) &= 2\epsilon^3(1-\epsilon) + 2\epsilon(1-\epsilon)^3, \text{ and} \\ p(F = 3) &= \epsilon^4 + 4\epsilon^2(1-\epsilon)^2 + (1-\epsilon)^4 \end{aligned}$$

illustrated in Fig. 1b over different values of ϵ.



(a) Base indel distribution.

(b) Sequence length distribution.

Figure 1: Distribution of base indels and sequence length over the transition probability ϵ. It is recommended to choose a value for ϵ that is smaller than 1/5 such that $p(M = m) < p(M = m + 1)$, as per Fig. 1a.

A sequence $\mathbf{z} = z_1 z_2 \dots$ of finite but variable length will emerge at the end of the process represented in Fig. 2. Let $\mathcal{Q}_f$ be the set of hidden paths, starting with $Q_t = \mathrm{M}_j$ and ending at some leaf-node state, that generate sequences of

length f. Let $Z^f = (Z_1^f, \ldots, Z_f^f)$ be a f-tuple of random variables that generates such sequences of length f. We have

$$p(Z^f = z_1 \ldots z_f, F = f) = \sum_{\mathbf{q} \in \mathcal{Q}_f} p\left(Z^f = z_1 \ldots z_f \mid Q_{t..t+4} = \mathbf{q}\right) p(Q_{t..t+4} = \mathbf{q}).$$

2 Final sequence distribution

We will write $p(X = x_1x_2x_3) = p(X = x_1x_2x_3 \mid Q_t = \mathrm{M}_j)$ for brevity. Underscore _ denotes a summation over the corresponding random variables. For example, $p(X = x_1\_\_) = \sum_{x_2,x_3} p(X = x_1x_2x_3)$.

2.1 Sequences of length 1

$$p(Z^1 = z_1, F = 1) = \epsilon^2(1-\epsilon)^2(p(X = z_1\_\_) + p(X = \_z_1\_) + p(X = \_\_z_1))/3$$

2.2 Sequences of length 2

$$\begin{aligned} p(Z^2 = z_1z_2, F = 2) = {} & 2\epsilon(1-\epsilon)^3(p(X = \_z_1z_2) + p(X = z_1\_z_2) + p(X = z_1z_2\_))/3 \\ & + \epsilon^3(1-\epsilon)(p(X = z_1\_\_) + p(X = \_z_1\_) + p(X = \_\_z_1))p(z_2)/3 \\ & + \epsilon^3(1-\epsilon)(p(X = z_2\_\_) + p(X = \_z_2\_) + p(X = \_\_z_2))p(z_1)/3 \end{aligned}$$

2.3 Sequences of length 3

$$\begin{aligned} p(Z^3 = z_1z_2z_3, F = 3) = {} & (1-\epsilon)^4 p(X = z_1z_2z_3) \\ & + 4\epsilon^2(1-\epsilon)^2(p(X = \_z_2z_3) + p(X = z_2\_z_3) + p(X = z_2z_3\_))p(z_1)/9 \\ & + 4\epsilon^2(1-\epsilon)^2(p(X = \_z_1z_3) + p(X = z_1\_z_3) + p(X = z_1z_3\_))p(z_2)/9 \\ & + 4\epsilon^2(1-\epsilon)^2(p(X = \_z_1z_2) + p(X = z_1\_z_2) + p(X = z_1z_2\_))p(z_3)/9 \\ & + \epsilon^4(p(X = z_3\_\_) + p(X = \_z_3\_) + p(X = \_\_z_3))p(z_1)p(z_2)/9 \\ & + \epsilon^4(p(X = z_2\_\_) + p(X = \_z_2\_) + p(X = \_\_z_2))p(z_1)p(z_3)/9 \\ & + \epsilon^4(p(X = z_1\_\_) + p(X = \_z_1\_) + p(X = \_\_z_1))p(z_2)p(z_3)/9 \end{aligned}$$

2.4 Sequences of length 4

$$\begin{aligned} p(Z^4 = z_1z_2z_3z_4, F = 4) = {} & \epsilon(1-\epsilon)^3(p(X = z_2z_3z_4)p(z_1) + p(X = z_1z_3z_4)p(z_2) \\ & + p(X = z_1z_2z_4)p(z_3) + p(X = z_1z_2z_3)p(z_4))/2 \\ & + \epsilon^3(1-\epsilon)(\\ & + p(X = \_z_3z_4)p(z_1)p(z_2) + p(X = \_z_2z_4)p(z_1)p(z_3) \\ & + p(X = \_z_2z_3)p(z_1)p(z_4) + p(X = \_z_1z_4)p(z_2)p(z_3) \\ & + p(X = \_z_1z_3)p(z_2)p(z_4) + p(X = \_z_1z_2)p(z_3)p(z_4) \\ & + p(X = z_3\_z_4)p(z_1)p(z_2) + p(X = z_2\_z_4)p(z_1)p(z_3) \\ & + p(X = z_2\_z_3)p(z_1)p(z_4) + p(X = z_1\_z_4)p(z_2)p(z_3) \\ & + p(X = z_1\_z_3)p(z_2)p(z_4) + p(X = z_1\_z_2)p(z_3)p(z_4) \\ & + p(X = z_3z_4\_)p(z_1)p(z_2) + p(X = z_2z_4\_)p(z_1)p(z_3) \\ & + p(X = z_2z_3\_)p(z_1)p(z_4) + p(X = z_1z_4\_)p(z_2)p(z_3) \\ & + p(X = z_1z_3\_)p(z_2)p(z_4) + p(X = z_1z_2\_)p(z_3)p(z_4))/9 \end{aligned}$$

2.5 Sequences of length 5

$$\begin{aligned} p(Z^5 = z_1z_2z_3z_4z_5, F = 5) = {} & \epsilon^2(1-\epsilon)^2(\\ & + p(z_1)p(z_2)p(X = z_3z_4z_5) + p(z_1)p(z_3)p(X = z_2z_4z_5) \\ & + p(z_1)p(z_4)p(X = z_2z_3z_5) + p(z_1)p(z_5)p(X = z_2z_3z_4) \\ & + p(z_2)p(z_3)p(X = z_1z_4z_5) + p(z_2)p(z_4)p(X = z_1z_3z_5) \\ & + p(z_2)p(z_5)p(X = z_1z_3z_4) + p(z_3)p(z_4)p(X = z_1z_2z_5) \\ & + p(z_3)p(z_5)p(X = z_1z_2z_4) + p(z_4)p(z_5)p(X = z_1z_2z_3))/10 \end{aligned}$$

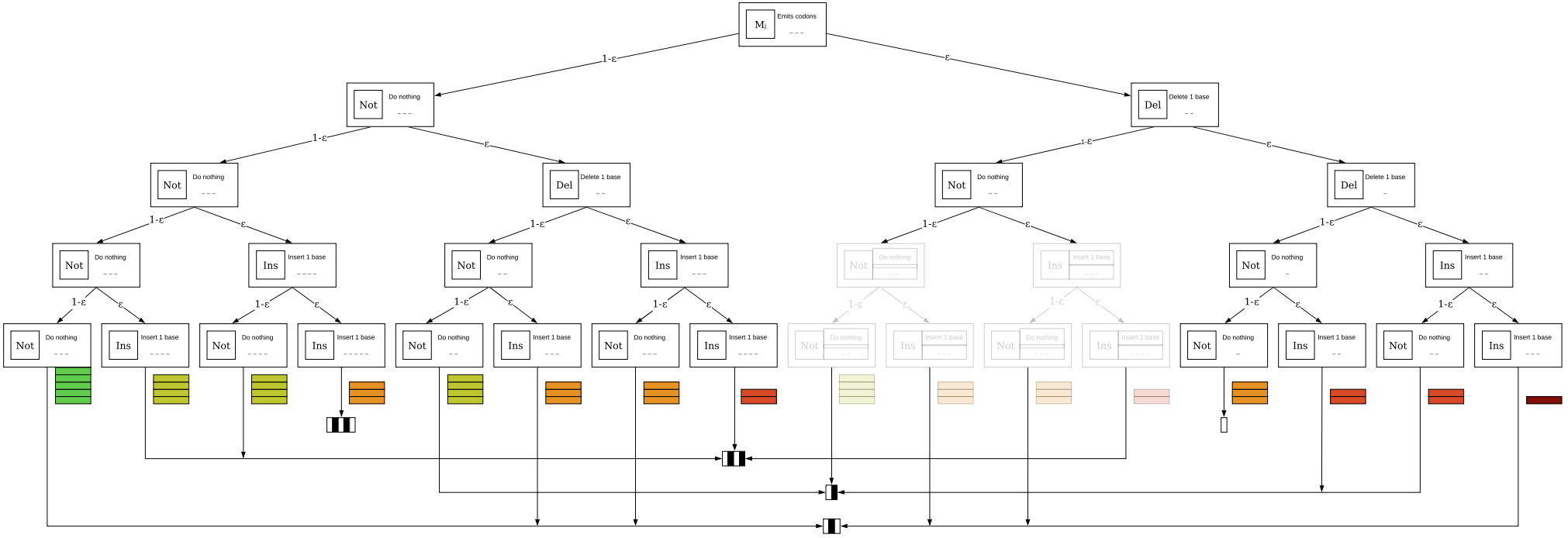


Figure 2: Matched codon HMM tree. The ϵ-transitions occur infrequently and exist to account for sequence errors. The most probably path ends at the first leaf-node from left to right.