Protein HMM

Danilo Horta

August 2019

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Phasellus eu felis felis. Sed turpis nibh, laoreet a aliquam sit amet, condimentum at leo. Nunc vitae ipsum quis magna mollis lacinia nec a est. Donec aliquam aliquet tortor vel rhoncus. Duis mi erat, vehicula a sem eget, tincidunt malesuada nulla. Aliquam pellentesque posuere nibh nec elementum. Quisque eu accumsan justo, ac viverra nibh.

1 Model

Definition 1.1. A Markov process is a stochastic process $Q_1, \ldots$ for which

$$p(Q_t = q_t \mid Q_1 = q_1, Q_2 = q_2, \ldots, Q_{t-1} = q_{t-1}) = p(Q_t = q_t \mid Q_{t-1} = q_{t-1}).$$

The possible values of Q_t form a finite set $\mathcal{Q}$ called the state space.

Definition 1.2. Let $\mathcal{A}$ be a non-empty finite set of symbols. Let $Q_1, Q_2, \ldots$ be a Markov process and let $S_1, S_2, \ldots$ be a stochastic process for which

$$p(S_t \in \mathcal{A} \mid Q_1 = q_1, Q_2 = q_2, \ldots, Q_t = q_t) = p(S_t \in \mathcal{A} \mid Q_t = q_t).$$

The pair (Q_t, S_t) is a hidden Markov model (HMM) with alphabet $\mathcal{A}$.

The standard HMM definition is often extended to include states that do not emit symbols. Those states are referred to as silent states and are useful to describe a missing alignment position, for example. This section goes a step further by defining a more general hidden Markov model that accounts for states that instead emit sequence of symbols of variable length, including zero-length sequences.

Definition 1.3. Let $\mathcal{A}$ be a non-empty finite set of symbols, $k \in \mathbb{N}_0$, and define $\mathcal{B} = \bigcup_{i=0}^{k} \mathcal{A}^i$. Let $Q_1, Q_2, \ldots$ be a Markov process and let $S_1, S_2, \ldots$ be a stochastic process for which

$$p(S_t \in \mathcal{B} \mid Q_1 = q_1, Q_2 = q_2, \ldots, Q_t = q_t) = p(S_t \in \mathcal{B} \mid Q_t = q_t).$$

The pair (Q_t, S_t) is an invisible Markov model (IMM) with alphabet $\mathcal{A}$ and limit k.

Let $\mathbf{z} = z_1..z_L$ be a sequence emitted from an IMM. The marginal likelihood of $\mathbf{s}$ is given by

$$\mathrm{ML}(\mathbf{z}) = p(V_1 = \mathbf{z}) + p(V_1 \neq \mathbf{z}, V_1||V_2 = \mathbf{z}) + p(V_1 \neq \mathbf{z}, V_1||V_2 \neq \mathbf{z}, V_1||V_2||V_3 = \mathbf{z}) + \ldots, \tag{1}$$

where $||$ denotes sequence concatenation.

Let F_t be the sequence length emitted by S_t, $L_t = 0 + F_1 + \cdots + F_{t-1}$, and $V_{1..t} = V_1||V_2||\ldots||V_t$. We have

$$p(V_{1..1} \neq \mathbf{z}, V_{1..2} \neq \mathbf{z}, \ldots, V_{1..t-1} \neq \mathbf{z}, V_{1..t} = \mathbf{z}) = p(V_{1..t} = \mathbf{z}, L_t < L).$$

Therefore, the marginal likelihood is also given by

$$\mathrm{ML}(\mathbf{z}) = \sum_{t=1}^{\infty} p(V_{1..t} = \mathbf{z}, L_t < L).$$

1.1 Viterbi

Let $i \leq L$ indicate the prefix $z_{1:i}$. We define

$$\mathcal{V}_{q_t,f_t}(i) = \max_{q_{t-1},f_{t-1}} \left\{\mathcal{V}_{q_{t-1},f_{t-1}}(i - f_t)p(V_t = z_{i-f_t+1:i} \mid Q_t = q_t)p(Q_t = q_t \mid Q_{t-1} = q_{t-1})\right\},$$

for $0 < i - f_t$, and

$$\mathcal{V}_{q_t,f_t}(i) = \max\Big\{ \max_{q_{t-1}} \left\{\mathcal{V}_{q_{t-1}}(0)p(V_t = z_{1:i} \mid Q_t = q_t)p(Q_t = q_t \mid Q_{t-1} = q_{t-1})\right\},$$
$$p(V_t = z_{1:i} \mid Q_t = q_t)p(Q_t = q_t)\Big\},$$

for $0 = i - f_t$.

The maximal probability over all state paths $q_1, q_2, \ldots$ and emission lengths $f_1, f_2, \ldots$ is given by

$$\max_{q_t,f_t} \mathcal{V}_{q_t,f_t}(L).$$

1.2 Not sure

Let $q_{1..t}$ be a sequence of states (also known as state path). The likelihood of sequence $\mathbf{z}$ for state path $q_{1..t}$ is given by

$$\mathrm{L}(\mathbf{z}, q_{1..t}) = p(V_{1..t} = \mathbf{z}, Q_{1..t} = q_{1..t}).$$

Note that

$$\begin{aligned}
\mathrm{L}(\mathbf{z}, q_{1..t}) &= \sum_{l_t=0}^{L} p(V_{1..t} = \mathbf{z}, L_t = l_t, Q_{1..t} = q_{1..t}) \\
&= \sum_{l_t=0}^{L} p(V_{1..t-1} = z_{1..l_t}, V_t = z_{l_t+1..L}, Q_{1..t-1} = q_{1..t-1}, Q_t = q_t) \\
&= \sum_{l_t=0}^{L} p(V_t = z_{l_t+1..L} \mid Q_t = q_t)p(Q_t = q_t \mid Q_{t-1} = q_{t-1})\mathrm{L}(z_{1..l_t}, q_{1..t-1})
\end{aligned}$$

1.3 Discussion

For computational reasons, it would be useful to have an upper bound on the summation of the marginal likelihood. We will define a type of IMM that has such a feature.

Definition 1.4. A cycle is any probable sequence of states that starts and ends with the same state.

Definition 1.5. A quiet state is a state that has a non-zero probability of emitting an empty sequence.

Definition 1.6. A quiet cycle is any cycle having only quiet states.

Corollary 1.1. Let M be the number of states of the IMM. If it has no quiet cycles, any sequence of M states will have emitted at least one symbol.

If IMM has no quiet cycles, there is always a $N \leq L \cdot M$ such that

$$\mathrm{ML}(\mathbf{z}) = \sum_{t=1}^{L \cdot M} p(V_{1..t} = \mathbf{z}, L_t < L).$$

2 Frame

Let (Q_t, S_t) be a HMM with the amino acid 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 $p\left(X_1 = x_1, X_2 = x_2, X_3 = x_3 \mid Q_t = \mathrm{M}_j\right)$ of M_j emitting the codon $(x_1, x_2, x_3) \in \mathcal{B}^3$ — one can also write $p\left(X = x_1x_2x_3 \mid Q_t = \mathrm{M}_j\right)$, 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. 4 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.

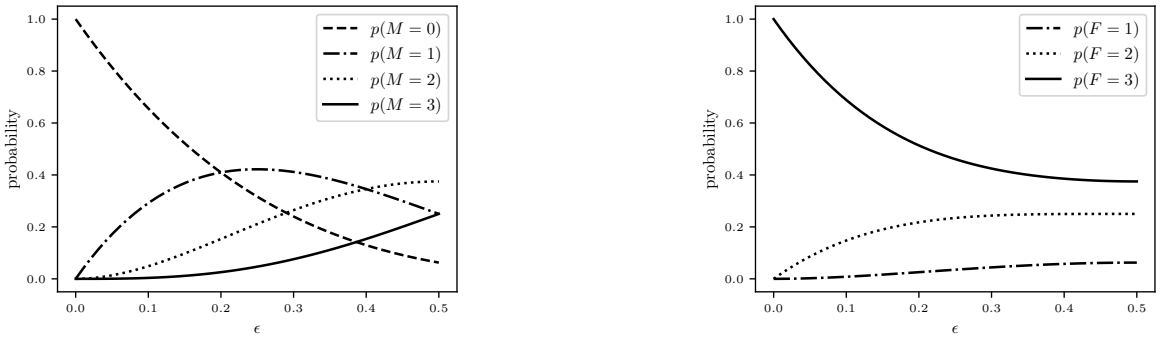
The codon emitted at node M_j can go under $m \in \{0, 1, 2, 3, 4\}$ base indels during the state transitions that end at some leaf-node state. The probability of it undergoing m indels is given by

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

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. 4. 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 $\mathbf{z} = z_1z_2\ldots$ of finite but variable length will emerge at the end of the process represented in Fig. 4. 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

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

3 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, $p(X = x_1\_\_) = \sum_{x_2,x_3} p(X = x_1x_2x_3)$.

3.1 Sequences of length 1

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

3.2 Sequences of length 2

$$\begin{aligned}
p(Z^2 = z_1z_2, F = 2) = {} & 2\epsilon(1-\epsilon)^3(p(X = _z_1z_2) + p(X = z_1_z_2) + p(X = z_1z_2_))/3 \\
& + \epsilon^3(1-\epsilon)(p(X = z_1__) + p(X = _z_1_) + p(X = __z_1))p(z_2)/3 \\
& + \epsilon^3(1-\epsilon)(p(X = z_2__) + p(X = _z_2_) + p(X = __z_2))p(z_1)/3
\end{aligned}$$

3.3 Sequences of length 3

$$\begin{aligned}
p(Z^3 = z_1z_2z_3, F = 3) = {} & (1-\epsilon)^4p(X = z_1z_2z_3) \\
& + 4\epsilon^2(1-\epsilon)^2(p(X = _z_2z_3) + p(X = z_2_z_3) + p(X = z_2z_3_))p(z_1)/9 \\
& + 4\epsilon^2(1-\epsilon)^2(p(X = _z_1z_3) + p(X = z_1_z_3) + p(X = z_1z_3_))p(z_2)/9 \\
& + 4\epsilon^2(1-\epsilon)^2(p(X = _z_1z_2) + p(X = z_1_z_2) + p(X = z_1z_2_))p(z_3)/9 \\
& + \epsilon^4(p(X = z_3__) + p(X = _z_3_) + p(X = __z_3))p(z_1)p(z_2)/9 \\
& + \epsilon^4(p(X = z_2__) + p(X = _z_2_) + p(X = __z_2))p(z_1)p(z_3)/9 \\
& + \epsilon^4(p(X = z_1__) + p(X = _z_1_) + p(X = __z_1))p(z_2)p(z_3)/9
\end{aligned}$$

3.4 Sequences of length 4

$$\begin{aligned}
p(Z^4 = z_1z_2z_3z_4, F = 4) = {} & \epsilon(1-\epsilon)^3(p(X = z_2z_3z_4)p(z_1) + p(X = z_1z_3z_4)p(z_2) \\
& + p(X = z_1z_2z_4)p(z_3) + p(X = z_1z_2z_3)p(z_4))/2 \\
& + \epsilon^3(1-\epsilon)(\\
& + p(X = _z_3z_4)p(z_1)p(z_2) + p(X = _z_2z_4)p(z_1)p(z_3) \\
& + p(X = _z_2z_3)p(z_1)p(z_4) + p(X = _z_1z_4)p(z_2)p(z_3) \\
& + p(X = _z_1z_3)p(z_2)p(z_4) + p(X = _z_1z_2)p(z_3)p(z_4) \\
& + p(X = z_3_z_4)p(z_1)p(z_2) + p(X = z_2_z_4)p(z_1)p(z_3) \\
& + p(X = z_2_z_3)p(z_1)p(z_4) + p(X = z_1_z_4)p(z_2)p(z_3) \\
& + p(X = z_1_z_3)p(z_2)p(z_4) + p(X = z_1_z_2)p(z_3)p(z_4) \\
& + p(X = z_3z_4_)p(z_1)p(z_2) + p(X = z_2z_4_)p(z_1)p(z_3) \\
& + p(X = z_2z_3_)p(z_1)p(z_4) + p(X = z_1z_4_)p(z_2)p(z_3) \\
& + p(X = z_1z_3_)p(z_2)p(z_4) + p(X = z_1z_2_)p(z_3)p(z_4))/9
\end{aligned}$$

3.5 Sequences of length 5

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

4 Model generalisation

Fig. 2 illustrates a probabilistic graphical representation of the IMM.

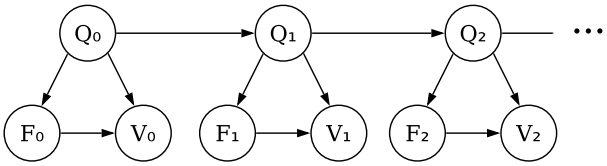


Figure 2: Invisible Markov model.

5 Old model generalisation

Let $L_t = F_0 + F_1 + \cdots + F_{t-1}$ be the random variable describing the number of emitted symbols right before the sequence emission by Q_t. Let $S_1, S_2, \ldots$ a stochastic process for symbol emission such that $S_t \in \mathcal{A}$ and

$$p(s_{l_t+1}..s_{l_t+f_t}, L_t = l_t, F_t = f_t, V_t = \mathbf{v}_t) = p(s_{l_t+1}..s_{l_t+f_t} \mid L_t = l_t, F_t = f_t, V_t = \mathbf{v}_t).$$

Let us define an additional stochastic process associated with a given IMM (random variables are omitted for didactic reasons whenever appropriate):

$$\begin{aligned}
p(S_0 = \emptyset \mid Q_0 = q_0) &= 1 \\
p(s_1..s_{f_1}, f_1, q_1) &= p(s_1..s_{f_1} \mid f_1, q_1)p(f_1, q_1) = p(s_1..s_{f_1} \mid f_1, q_1)p(f_1 \mid q_1)p(q_1) \\
p(s_1..s_{f_1+f_2}, f_1, f_2, q_1, q_2) &= p(s_{f_1+1}..s_{f_1+f_2} \mid f_1, f_2, q_2)p(s_1..s_{f_1}, f_1, f_2, q_1, q_2) \\
&= p(s_{f_1+1}..s_{f_1+f_2} \mid f_1, f_2, q_2)p(s_1..s_{f_1} \mid f_1, q_1)p(f_1, f_2, q_1, q_2) \\
&= p(s_{f_1+1}..s_{f_1+f_2} \mid f_1, f_2, q_2)p(s_1..s_{f_1} \mid f_1, q_1)p(f_1 \mid q_1)p(f_2 \mid q_2)p(q_2 \mid q_1)p(q_1)
\end{aligned}$$

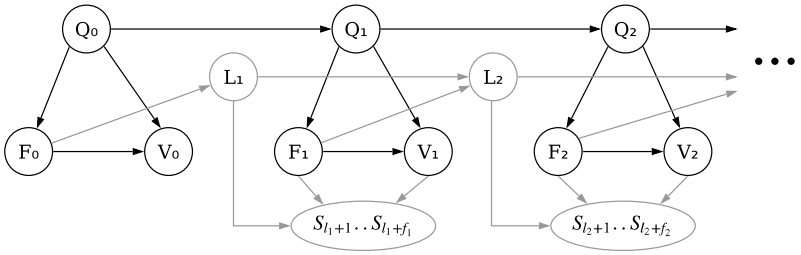


Figure 3: Invisible Markov model with symbol emission stochastic process.

Let $L_t = F_0 + F_1 + \cdots + F_{t-1}$, be the random variable describing the number of emitted symbols right before the sequence emission by Q_t. We define

$$p(S_{l_t+1} = s_{l_t+1}, \ldots, S_{l_t+f_t} = s_{l_t+f_t} \mid L_t = l_t, F_t = f_t, Q_t = q_t) = p(V_t = (s_{l_t+1}, \ldots, s_{l_t+f_t}) \mid F_t = f_t, Q_t = q_t).$$

Given an IMM, we define the marginal likelihood of a sequence $\mathbf{z} = z_1 z_2 \ldots z_L$ as being

$$p(S_1 = z_1, \ldots, S_L = z_L) = \sum_{\substack{q_1..q_N \\ f_1..f_N}} p(S_1 = z_1, \ldots, S_L = z_L, Q_1 = q_1, F_1 = f_1, \ldots, Q_N = q_N, F_N = f_N),$$

subject to $L = l_N + f_N$.[1]

Silent and non-silent HMM states are particular cases of IMM states that emit sequences of length zero and one, respectively.

[1] It might be a good moment to talk about degenerate IMM: when there is silent states cycle and/or there are states with unbounded sequence lengths.

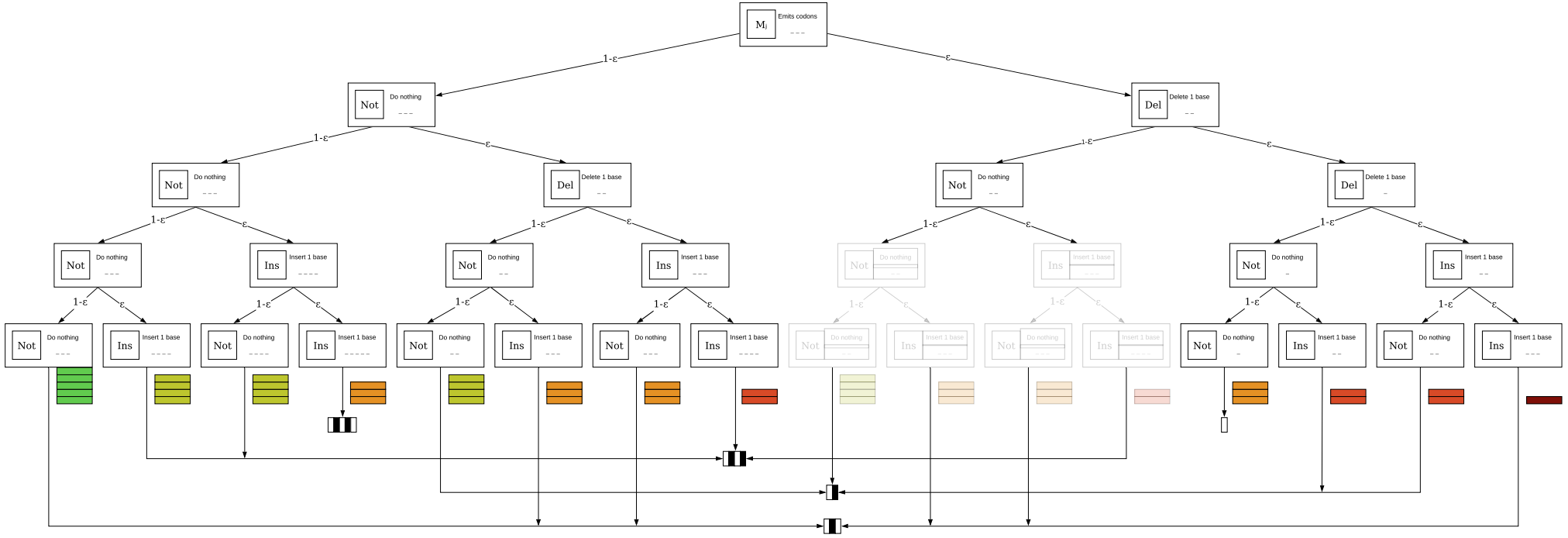


Figure 4: Matched codon HMM tree. The ϵ-transitions occur infrequently and exist to account for sequence errors. The most probably path ends at the first leaf-node from left to right.