Derivation for Marker Code

Yiliang Wan

1 Overview

The problem is to reconstruct the template sequence $\mathbf{t} = \{t_1, \ldots, t_N\} \in \Sigma^N$, given a set of prior probabilities for each template symbol $\mathbf{t}_p = \{p(t_1), \ldots, p(t_N)\}$ and a known sample sequence $\mathbf{s} = \{s_1, \ldots, s_M\} \in \Sigma^M$, where Σ is the alphabet.

In this derivation, we adopt the simple assumption that the probability of each $A \in \Sigma$ is equivalent if there is no other information. Also, the insertion rate ins_p, deletion rate del_p and substitution rate sub_p are known.

To make the following derivation more concise, we first introduce a shorthand:

$$Let \qquad p(A \mid B, \mathbf{t}_p) := \mathbb{E}_{p(\mathbf{t})}\left[p(A \mid B, \mathbf{t})\right] = \sum_{\mathbf{t}} p(A \mid B, \mathbf{t})p(\mathbf{t})$$

To do the reconstruction, we use HMM (Hidden Markov Model), with π_k as the latent variables, which represent the alignment states between certain s_i and t_j. There are three kinds of alignment states:

- **Match** denoted by $\pi^M(i, j)$, which means that s_i and t_j are aligned (including substitution).

- **Insertion** denoted by $\pi^I(i, j)$, which means that s_i is an insertion not being aligned to any symbol in $\mathbf{t}$, with the last aligned or deleted symbol in $\mathbf{t}$ being t_j.

- **Deletion** denoted by $\pi^D(i, j)$, which means there is a deletion for t_j, witch the last aligned or inserted symbol in $\mathbf{s}$ being s_i.

All the alignment states will together form an alignment path $\mathbf{\Pi} = \{\pi_1, \ldots, \pi_K\}$, where K is uncertain because there are insertions and deletions.

For convenience, we further define the start symbols s_0 and t_0 for the sample and the template, respectively, and consider that they are always matched, thus we have

$$p\left(\pi^M(0,0)\right) = 1$$
$$p\left(\pi^I\ (0,0)\right) = 0 \tag{1}$$
$$p\left(\pi^D\ (0,0)\right) = 0$$

In general, there are two steps to decode the marker code. First, calculate $p\left(\pi^S(i, j) \mid \mathbf{s}, \mathbf{t}_p\right)$ for all $S \in \{M, I, D\}$, $i = 0, \ldots, M$ and $j = 0, \ldots, N$, which constitute the belief of the alignment path $p(\mathbf{\Pi} \mid \mathbf{s}, \mathbf{t}_p)$. This belief is later used for inferring the posterior of each t_j by marginalizing all the possible paths

$$\begin{aligned}
\hat{p}(t_j \mid s) &= \mathbb{E}_{p(\mathbf{\Pi} \mid \mathbf{s}, \mathbf{t}_p)} \|p(t_j, \mathbf{\Pi})\| \\
&= \sum_{i=0}^{N} p\left(t_j \mid \pi^M(i, j)\right) p\left(\pi^M(i, j) \mid \mathbf{s}, \mathbf{t}_p\right) + p\left(t_j \mid \pi^D(i, j)\right) p\left(\pi^D(i, j) \mid \mathbf{s}, \mathbf{t}_p\right)
\end{aligned} \tag{2}$$

2 Transition Probabilitiy

Here we use a first order HMM, and the transition probability p_{trans} is the conditional probabilities of an alignment state given the previous alignment state

$$p_{trans} = p(\pi_k \mid \pi_{k-1}) \tag{3}$$

p_{trans} can be represented by a 3×3 matrix T. For example, T_{MI} is the probability of a match following an insertion.

For most of the transitions, we can use a consistent T which is calculated according to the insertion rate and deletion rate of the system. However, the last symbols s_M and t_N should be taken care of. For all $S \in \{M, I, D\}$, we have the following conditional probabilities

$$p\left(\pi^M(M,j) \mid \pi^S(M,j-1)\right) = 0$$
$$p\left(\pi^I\ (M,j) \mid \pi^S(M,j-1)\right) = 0 \qquad for\ j = 1,\ldots,N-1 \tag{4}$$
$$p\left(\pi^D\ (M,j) \mid \pi^S(M,j-1)\right) = 1$$

and

$$p\left(\pi^M(i,N) \mid \pi^S(i-1,N)\right) = 0$$
$$p\left(\pi^I\ (i,N) \mid \pi^S(i-1,N)\right) = 1 \qquad for\ i = 1,\ldots,M-1 \tag{5}$$
$$p\left(\pi^D\ (i,N) \mid \pi^S(i-1,N)\right) = 0$$

3 Emission Probability

The emission probabilies p_{emis} are the conditional probabilites of a particular sample symbol s_m given the corresponding alignment state and the template's prior $\mathbf{t}_p$

$$p_{emis}^{(S,n)}(s_m) = p\left(s_m \mid \pi^S(m,n), \mathbf{t}_p\right) \tag{6}$$

We only consider the simple condition of that the prior For a match, we have

$$\begin{cases} p_{emis}^{(M,j)}(s_m) = 1/|\Sigma| & if \quad t_j\ is\ not\ a\ marker, \\ p_{emis}^{(M,j)}(s_m) = 1 - sub\_p & if \quad t_j\ is\ a\ marker\ and\ t_j = s_m, \\ p_{emis}^{(M,j)}(s_m) = sub\_p/(1 - |\Sigma|) & if \quad t_j\ is\ a\ marker\ and\ t_j \neq s_m. \end{cases} \tag{7}$$

Here is where the markers provide synchronization information to the decoding process. For an insertion, we have

$$p_{emis}^{(I,j)}(s_m) = p_{emis}^I(s_m) = 1/|\Sigma| \tag{8}$$

since no relevant information is available. The deletion states will alwyas emit a dummy symbol in $\mathbf{s}$, so the emission rate is always 1

$$p_{emis}^{(D,j)}(s_m) = p_{emis}^D(s_m) = 1 \tag{9}$$

4 Belief of Alignment Path

We calculate all possible $p\left(\pi^S(i,j) \mid \mathbf{s}, \mathbf{t}_p\right)$ with forward-backward algorithm.

4.1 Forward Message

The forward messages are defined as

$$\begin{aligned} f^M(i,j) &= P(s_1,\ldots,s_i, \pi^M(i,j) \mid \mathbf{t}_p) \\ f^I\ (i,j) &= P(s_1,\ldots,s_i, \pi^I\ (i,j) \mid \mathbf{t}_p) \\ f^D\ (i,j) &= P(s_1,\ldots,s_i, \pi^D\ (i,j) \mid \mathbf{t}_p) \end{aligned} \tag{10}$$

We will first examine $f^M(i,j)$.

$$\begin{aligned} f^M(i,j) &= P(s_1,\ldots,s_i, \pi^M(i,j) \mid \mathbf{t}_p) = \sum_{\pi_{prev}} P(s_1,\ldots,s_i, \pi^M(i,j), \pi_{prev} \mid \mathbf{t}_p) \\ &= P(s_1,\ldots,s_i, \pi^M(i,j), \pi^M(i-1,j-1) \mid \mathbf{t}_p) + P(s_1,\ldots,s_i, \pi^M(i,j), \pi^I(i-1,j-1) \mid \mathbf{t}_p) \\ &\quad + P(s_1,\ldots,s_i, \pi^M(i,j), \pi^D(i-1,j-1) \mid \mathbf{t}_p) \end{aligned} \tag{11}$$

The first term of (11) can be further decomposed with the conditional independent relationships in the HMM

$$P(s_1, \ldots, s_i, \pi^M(i,j), \pi^M(i-1,j-1) \mid \mathbf{t}_p)$$
$$= P(s_1, \ldots, s_{i-1} \mid s_i, \pi^M(i,j), \pi^M(i-1,j-1)\mathbf{t}_p) \times P(s_i \mid \pi^M(i-1,j-1), \pi^M(i,j), \mathbf{t}_p)$$
$$\times P(\pi^M(i,j) \mid \pi^M(i-1,j-1), \mathbf{t}_p) \times P(\pi^M(i-1,j-1) \mid \mathbf{t}_p)$$
$$= P(s_1, \ldots, s_{i-1} \mid \pi^M(i-1,j-1)\mathbf{t}_p) \times P(s_i \mid \pi^M(i,j), \mathbf{t}_p) \times P(\pi^M(i,j) \mid \pi^M(i-1,j-1))$$
$$\times P(\pi^M(i-1,j-1) \mid \mathbf{t}_p)$$
$$= P(s_1, \ldots, s_{i-1} \mid \pi^M(i-1,j-1)\mathbf{t}_p) \times P(\pi^M(i-1,j-1) \mid \mathbf{t}_p) \times P(\pi^M(i,j) \mid \pi^M(i-1,j-1))$$
$$\times P(s_i \mid \pi^M(i,j), \mathbf{t}_p)$$
$$= P(s_1, \ldots, s_{i-1}, \pi^M(i-1,j-1) \mid \mathbf{t}_p) \times P(\pi^M(i,j) \mid \pi^M(i-1,j-1)) \times P(s_i \mid \pi^M(i,j), \mathbf{t}_p)$$
$$= f^M(i-1,j-1) \times T_{MM} \times p_{emis}^{(M,j)}(s_i)$$
$$(12)$$

Similar procedures can be applied to the second and the third term of (11)

$$P(\pi^I(i-1,j-1), \pi^M(i,j), s_1, \ldots, s_i \mid \mathbf{t}_p) = f^I(i-1,j-1) \times T_{IM} \times p_{emis}^{(M,j)}(s_i) \qquad (13)$$

$$P(\pi^D(i-1,j-1), \pi^M(i,j), s_1, \ldots, s_i \mid \mathbf{t}_p) = f^D(i-1,j-1) \times T_{DM} \times p_{emis}^{(M,j)}(s_i) \qquad (14)$$

Put (12) (13) and (14) into (11), and we can obtain the recursion of f^M

$$f^M(i,j) = (f^M(i-1,j-1) \times T_{MM} + f^I(i-1,j-1) \times T_{IM} + f^D(i-1,j-1) \times T_{DM}) \times p_{emis}^{(M,j)}(s_i) \qquad (15)$$

Similarly, we can obtain the recursion formula of f^I and f^D

$$f^I(i,j) = \sum_{\pi_{prev}} P(s_1, \ldots, s_i, \pi^I(i,j), \pi_{prev} \mid \mathbf{t}_p)$$
$$= (f^M(i-1,j) \times T_{MI} + f^I(i-1,j) \times T_{II} + f^D(i-1,j) \times T_{DI}) \times p_{emis}^I(s_i) \qquad (16)$$

$$f^D(i,j) = \sum_{\pi_{prev}} P(s_1, \ldots, s_i, \pi^I(i,j), \pi_{prev} \mid \mathbf{t}_p)$$
$$= f^M(i,j-1) \times T_{MD} + f^I(i,j-1) \times T_{ID} + f^D(i,j-1) \times T_{DD} \qquad (17)$$

4.2 Backward Messages

The forward messages are defined as

$$b^M(i,j) = P(s_{i+1}, \ldots, s_M \mid \pi^M(i,j), \mathbf{t}_p)$$
$$b^I(i,j) = P(s_{i+1}, \ldots, s_M \mid \pi^I(i,j), \mathbf{t}_p) \qquad (18)$$
$$b^D(i,j) = P(s_{i+1}, \ldots, s_M \mid \pi^D(i,j), \mathbf{t}_p)$$

Similarly, we will first examine $b^M(i,j)$.

$$b^M(i,j) = P(s_{i+1}, \ldots, s_M \mid \pi^M(i,j), \mathbf{t}_p) = \sum_{\pi_{next}} P(s_{i+1}, \ldots, s_M, \pi_{next} \mid \pi^M(i,j), \mathbf{t}_p)$$
$$= P(s_{i+1}, \ldots, s_M, \pi^M(i+1,j+1) \mid \pi^M(i,j), \mathbf{t}_p) + P(s_{i+1}, \ldots, s_M, \pi^I(i+1,j) \mid \pi^M(i,j), \mathbf{t}_p) \qquad (19)$$
$$+ P(s_{i+1}, \ldots, s_M, \pi^D(i,j+1)) \mid \pi^M(i,j), \mathbf{t}_p)$$

$b^M(i,j)$ is the sum of three terms which correspond to three possible next states. The first term of (19) can be further decomposed

$$P(s_{i+1}, \ldots, s_M, \pi^M(i+1,j+1) \mid \pi^M(i,j), \mathbf{t}_p)$$
$$= P(s_{i+2}, \ldots, s_M \mid s_{i+1}, \pi^M(i+1,j+1), \pi^M(i,j), \mathbf{t}_p) \times P(s_{i+1}, \pi^M(i+1,j+1) \mid \pi^M(i,j), \mathbf{t}_p)$$
$$= P(s_{i+2}, \ldots, s_M \mid s_{i+1}, \pi^M(i+1,j+1), \pi^M(i,j), \mathbf{t}_p) \times P(s_{i+1} \mid \pi^M(i+1,j+1)\pi^M(i,j), \mathbf{t}_p)$$
$$\times P(\pi^M(i+1,j+1) \mid \pi^M(i,j), \mathbf{t}_p)$$
$$= P(s_{i+2}, \ldots, s_M \mid \pi^M(i+1,j+1), \mathbf{t}_p) \times P(s_{i+1} \mid \pi^M(i+1,j+1), \mathbf{t}_p) \times P(\pi^M(i+1,j+1) \mid \pi^M(i,j))$$
$$= b^M(i+1,j+1) \times p_{emis}^{(M,j+1)}(s_{i+1}) \times T_{MM}$$
$$(20)$$

3

Similar procedures can be applied to the second and the third term of (19)

$$P(s_{i+1},\ldots,s_M,\pi^I(i+1,j)\mid\pi^M(i,j),\mathbf{t}_p) = b^I(i+1,j)\times p^I_{emis}(s_{i+1})\times T_{MI} \tag{21}$$

$$P(s_{i+1},\ldots,s_M,\pi^D(i,j+1))\mid\pi^M(i,j),\mathbf{t}_p) = b^D(i,j+1)\times T_{MD} \tag{22}$$

Put (20) (21) and (22) into (19), and we can obtain the recursion of b^M

$$
\begin{aligned}
b^M(i,j) &= b^M(i+1,j+1)\times p^{(M,j+1)}_{emis}(s_{i+1})\times T_{MM} + b^I(i+1,j)\times p^I_{emis}(s_{i+1})\times T_{MI}\\
&\quad + b^D(i,j+1)\times T_{MD}
\end{aligned} \tag{23}
$$

Similarly, we can obtain the recursion formula of b^I and b^D

$$
\begin{aligned}
b^I(i,j) &= b^M(i+1,j+1)\times p^{(M,j+1)}_{emis}(s_{i+1})\times T_{IM} + b^I(i+1,j)\times p^I_{emis}(s_{i+1})\times T_{II}\\
&\quad + b^D(i,j+1)\times T_{ID}
\end{aligned} \tag{24}
$$

$$
\begin{aligned}
b^D(i,j) &= b^M(i+1,j+1)\times p^{(M,j+1)}_{emis}(s_{i+1})\times T_{DM} + b^I(i+1,j)\times p^I_{emis}(s_{i+1})\times T_{DI}\\
&\quad + b^D(i,j+1)\times T_{DD}
\end{aligned} \tag{25}
$$

4.3 Scaling Factors

To prevent underflowing, we define scaled message based on the forward and backward messages. The scaled forward messages are defined as

$$
\begin{aligned}
\widehat{f}^M(i,j) &= P(\pi^M(i,j)\mid s_1,\ldots,s_i,\mathbf{t}_p) = \frac{f^M(i,j)}{p(s_1,\ldots,s_i\mid\mathbf{t}_p)}\\
\widehat{f}^I(i,j) &= P(\pi^I(i,j)\mid s_1,\ldots,s_i,\mathbf{t}_p) = \frac{f^I(i,j)}{p(s_1,\ldots,s_i\mid\mathbf{t}_p)}\\
\widehat{f}^D(i,j) &= P(\pi^D(i,j)\mid s_1,\ldots,s_i,\mathbf{t}_p) = \frac{f^D(i,j)}{p(s_1,\ldots,s_i\mid\mathbf{t}_p)}
\end{aligned} \tag{26}
$$

Correspondingly, we can define scaled backward messages

$$
\begin{aligned}
\widehat{b}^M(i,j) &= \frac{P(s_{i+1},\ldots,s_M\mid\pi^M(i,j),\mathbf{t}_p)}{p(s_{i+1},\ldots,s_M\mid s_1,\ldots,s_i,\mathbf{t}_p)} = \frac{b^M(i,j)}{p(s_{i+1},\ldots,s_M\mid s_1,\ldots,s_i,\mathbf{t}_p)}\\
\widehat{b}^I(i,j) &= \frac{P(s_{i+1},\ldots,s_M\mid\pi^I(i,j),\mathbf{t}_p)}{p(s_{i+1},\ldots,s_M\mid s_1,\ldots,s_i,\mathbf{t}_p)} = \frac{b^I(i,j)}{p(s_{i+1},\ldots,s_M\mid s_1,\ldots,s_i,\mathbf{t}_p)}\\
\widehat{b}^D(i,j) &= \frac{P(s_{i+1},\ldots,s_M\mid\pi^D(i,j),\mathbf{t}_p)}{p(s_{i+1},\ldots,s_M\mid s_1,\ldots,s_i,\mathbf{t}_p)} = \frac{b^D(i,j)}{p(s_{i+1},\ldots,s_M\mid s_1,\ldots,s_i,\mathbf{t}_p)}
\end{aligned} \tag{27}
$$

The scaling factors are

$$c_i = P(s_i\mid s_1,\ldots,s_{i-1},\mathbf{t}_p) \tag{28}$$

and we can get (29) with the chain rule.

$$P(s_1,\ldots,s_m\mid\mathbf{t}_p) = \prod_{i=1}^m c_i \tag{29}$$

The scaled version of (15), (16) and (17) can be derived

$$
\begin{aligned}
c_i\widehat{f}^M(i,j) &= (\widehat{f}^M(i-1,j-1)\times T_{MM} + \widehat{f}^I(i-1,j-1)\times T_{IM} + \widehat{f}^D(i-1,j-1)\times T_{DM})\\
&\quad \times p^{(M,j)}_{emis}(s_i)
\end{aligned} \tag{30}
$$

$$c_i\widehat{f}^I(i,j) = (\widehat{f}^M(i-1,j)\times T_{MI} + \widehat{f}^I(i-1,j)\times T_{II} + \widehat{f}^D(i-1,j)\times T_{DI})\times p^I_{emis}(s_i) \tag{31}$$

$$\widehat{f}^D(i,j) = \widehat{f}^M((i,j-1)\times T_{MD} + \widehat{f}^I((i,j-1)\times T_{ID} + \widehat{f}^D((i,j-1)\times T_{DD} \tag{32}$$

Since the sum of the probabilities for all possible paths is 1, we have

$$\sum_{j=0}^N \left(\pi^M(i,j) + \pi^I(i,j)\right) = 1 \tag{33}$$

4

Similarly, we also have (34) because it is also the case for a conditional probability.

$$\sum_{j=0}^{N} \left(\widehat{f}^M(i,j) + \widehat{f}^I(i,j) \right) = 1 \tag{34}$$

So, if we denote the right sides of (30) and (31) as $\widetilde{f}^M(i,j)$ and $\widetilde{f}^I(i,j)$, repectively, we can calulate c_i with

$$c_i = \sum_{j=0}^{N} \left(\widetilde{f}^M(i,j) + \widetilde{f}^I(i,j) \right) \tag{35}$$

Similarly, we can get the scaled version of (22), (23) and (24)

$$
\begin{aligned}
c_{i+1}\widehat{b}^M(i,j) &= \widehat{b}^M(i+1,j+1) \times p_{emis}^{(M,j+1)}(s_{i+1}) \times T_{MM} + \widehat{b}^I(i+1,j) \times p_{emis}^I(s_{i+1}) \times T_{MI} \\
&\quad + c_{i+1} \times \widehat{b}^D(i,j+1) \times T_{MD} \\
c_{i+1}\widehat{b}^I(i,j) &= \widehat{b}^M(i+1,j+1) \times p_{emis}^{(M,j+1)}(s_{i+1}) \times T_{IM} + \widehat{b}^I(i+1,j) \times p_{emis}^I(s_{i+1}) \times T_{II} \\
&\quad + c_{i+1} \times \widehat{b}^D(i,j+1) \times T_{ID} \\
c_{i+1}\widehat{b}^D(i,j) &= \widehat{b}^M(i+1,j+1) \times p_{emis}^{(M,j+1)}(s_{i+1}) \times T_{DM} + \widehat{b}^I(i+1,j) \times p_{emis}^I(s_{i+1}) \times T_{DI} \\
&\quad + c_{i+1} \times \widehat{b}^D(i,j+1) \times T_{DD}
\end{aligned}
\tag{36}
$$

It is not difficuly to verify that the probabilities of all alignments can be calculated with correspoinding scaled forward and backward messages

$$p\left(\pi^S(i,j) \mid \mathbf{s}, \mathbf{t}_p\right) = \widehat{f}^S(i,j)\, \widehat{b}^S(i,j) \tag{37}$$

4.4   Boundary Condition

For all invalid i and j, we set $p\left(\pi^S(i,j) \mid \mathbf{s}, \mathbf{t}_p\right)$ to 0.

For the forward messages, we can initilize the recursion with (1) and $c[0]=1$.

For backward message passing, we initilize the last row with

$$\widehat{b}^S(M,j) = 1 \qquad for\ all\ S \in \{M,I,D\}\ and\ j = 1,\dots,N \tag{38}$$

which can be verified by replacing the definition of $\widehat{f}^S(M,j)$ into (37).

Also, the message passing on the boundary should be adapted according to (4) and (5).