Note that instead of explicitly labeling the data and then counting state transitions as we do with labeled data, the association of symbols and states is implicit in the re-estimation process in the inner loop of the algorithm.

$\mathcal{E}_i(\sigma)$ is the expected number of times that σ is emitted from state E_i:

$$\begin{aligned}\mathcal{E}_i(\sigma) &= \sum_d P(q_t^d = E_i, O_t^d = \sigma | O^d, \lambda) &(5.10)\\ &= \sum_d \frac{\sum_{\{t|O_t^d=\sigma\}} \alpha(t,i)\,\beta(t+1,i)}{P(O^d)}\,. &(5.11)\end{aligned}$$

Again, the quantities on the right hand side can be calculated using the Forward and Backward algorithms. Finally, the initial probability π_i is given by

$$\begin{aligned}\pi_i &= \sum_{d=1}^{k} p(q_1^d = S_i | O_d, \lambda) &(5.12)\\ &= \sum_{d=1}^{k} \frac{I^d(i)}{P(O^d)} &(5.13)\end{aligned}$$

It can be proven that if current estimate is replaced by these new estimates then the likelihood of the data will not decrease (i.e. will increase unless already at a local maximum/critical point). See Durbin, Section 11.6 for discussion of avoiding local maxima and other typical pitfalls with this algorithm.

The *Baum-Welch* algorithm estimates the values of the parameters from training data and, thus, implicitly discovers the motif. *Baum-Welch* does output an explicit description of the motif. To determine the motif explicitly, the Viterbi or posterior decoding are used to label each of the input sequences.

5.8 Profile HMMs

In 1994, Krogh, Haussler[2] and colleagues introduced a generic HMM topology specifically designed to model conserved sequence motifs. It captures the propensity to observe specific amino acids or nucleotides at each position in a pattern and allows for insertions and deletions. This topology, called a *Profile HMM*, can be customized for a broad range of conserved motifs by selecting the appropriate length for a given motif and initializing the parameters to capture the specific properties of the motif.

Here, we introduce the features of the Profile HMM model by showing how it could be used to model the `WEIRD` motif based on the following alignment, which has no gaps and no positional dependencies:

[2] Krogh, A., Brown, M., Mian, I. S., Sjolander, K., and Haussler, D. (1994). Hidden Markov models in computational biology: Applications to protein modeling. J. Mol. Biol., 235:1501-1531.

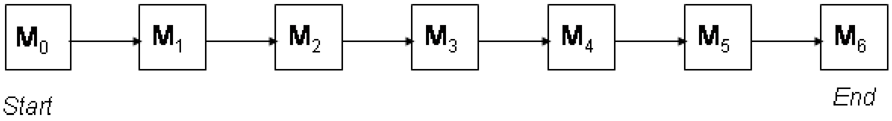


Figure 5.10: A HMM for modeling fixed length motifs. Note that this HMM is equivalent to a Position Specific Scoring Matrix.

```
WEIRD
WEIRD
WEIRE
WEIQH
```

We can recognize the `WEIRD` motif using an HMM with the simple topology shown in Fig. 5.10, where the transitions probabilities are

$$a_{M_i,M_j} \begin{cases} 1, & \text{if } j = i+1 \\ 0, & \text{otherwise} \end{cases}.$$

The emission probabilities can be estimated from labeled training sequences. Given an ungapped multiple alignment of k sequences, the emission probabilities are

$$e_{M_i}(\sigma) = \frac{c[\sigma,i]+b}{k+b|\Sigma|}, \tag{5.14}$$

where $c[\sigma,i]$ is the number of σ's at position i in the training motif and b is a pseudocount. The Start and End states (M_0 and M_6 in Fig. 5.10) are silent. Note that $e_{M_i}(\sigma)$ is equivalent to $q[\sigma,i]$, the frequency matrix that we derived for the PSSM example using pseudocounts. Moreover, when $\mathcal{E}_{M_i}(\sigma) = c[\sigma,i]$, Equation 5.14 is equivalent to the general definition of emission probabilities for an HMM given in Equation 5.6.

In order to assess whether a new sequence, O, contains an instance of the `WEIRD` motif, we calculate a likelihood ratio:

$$\frac{P(O|H_A)}{P(O|H_O)}.$$

Our alternate hypothesis, H_A, is that O is an instance of the motif represented by the HMM in Fig. 5.10. In order to obtain a likelihood ratio, we also need a model of the null hypothesis, H_0, that the amino acids in O were sampled from the background distribution. Fig. 5.11 shows a very simple null model. In this model, all transition probabilities are equal to one. The emission probability $e(\sigma)$ is defined to be $p(\sigma)$, where $p(\sigma)$ is the background frequency of residue σ.

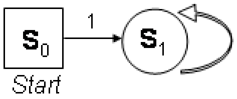


Figure 5.11: An HMM representing the null model. Each symbol, σ, is emitted with probability $p(\sigma)$, the background frequency of σ.

Given these two models, we can score O by calculating the probability that O was emitted by the Profile HMM in Fig. 5.10 and comparing it with the probability that O was emitted by the background model in Fig. 5.11. For example, if $O = O_1O_2O_3O_4O_5$, then $P(O|H_A)$ is equal to

$$\pi_{M_O} \cdot e_{M_1}(O_1) \cdot a_{M_O M_1} \cdot e_{M_2}(O_2) \cdot a_{M_2 M_3} \cdot e_{M_3}(O_3) \cdot a_{M_3 M_4} \cdot e_{M_4}(O_4) \cdot a_{M_4 M_5} \cdot e_{M_5}(O_5) \cdot a_{M_5 M_6} \cdot$$

Since the initial and transition probabilities are equal to one ($\pi_{M_O} = 1; a_{M_i M_{i+1}} = 1, 0 \leq i \leq 5$), this reduces to

$$P(O|H_A) = e_{M_1}(O_1) \cdot e_{M_2}(O_2) \cdot e_{M_3}(O_3) \cdot e_{M_4}(O_4) \cdot e_{M_5}(O_5)$$

or

$$P(O|H_A) = \prod_{t=1}^{5} e_{M_t}(O_t) \ .$$

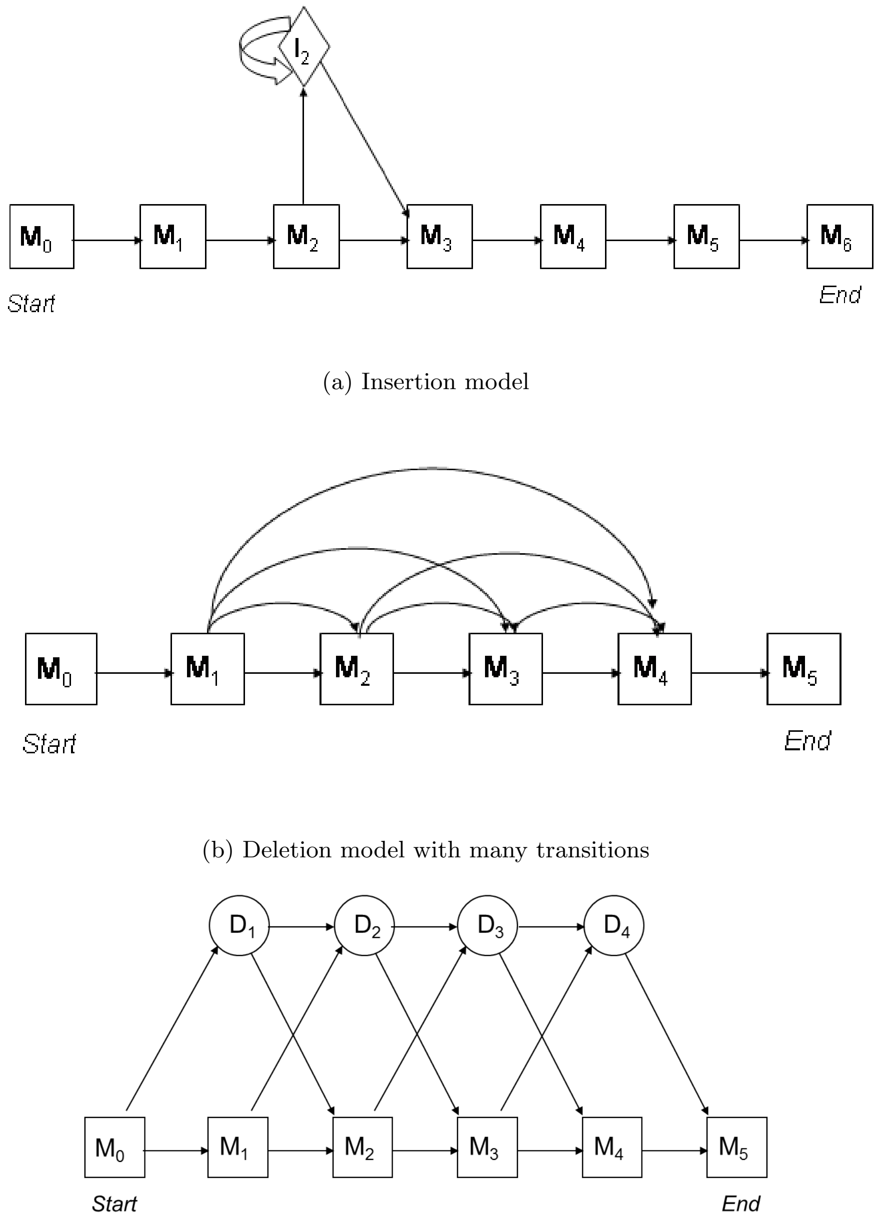
The probability that O was emitted by the background model is $\prod_{t=1}^{5} p(O_t)$. The score of sequence O is the log likelihood ratio

$$\sum_{t=1}^{5} \log \frac{e_{M_t}(O_t)}{p(O_t)} \ ,$$

which is equivalent to $\sum_{t=1}^{5} S[O_t, i]$, the score we would have obtained with the PSSM for the `WEIRD` motif.

We now have an HMM that is equivalent to a PSSM for a conserved motif. It can be used to identify motifs of a fixed size, but not cannot handle variations in length. We next extend this model to accommodate insertions and deletions. We can modify the basic HMM to recognize motif instances with insertions, such as $O =$ `WECIRD`, by adding an insertion state between any two match states, M_i and M_{i+1}, as shown in Fig. 5.12a. The emission probabilities for the insertion state are the background frequencies.

We also wish to recognize motif instances with deletions, e.g., $O =$ `WERD`. One approach to capturing such deletions would be to add edges allowing us to jump over any set of match states. An example of this is shown in Fig. 5.12b. The disadvantage to this approach



(a) Insertion model

(b) Deletion model with many transitions

(c) Deletion model with fewer transitions

Figure 5.12: (a) Additional insertion states enable recognition of pattern instances with insertions. This example allows for the insertion of one or more symbols between positions 2 and 3 in the pattern. (b) Adding an arc between every pair of sequences allows for deletions, but the number of transitions grows rapidly with the number of Match states. (c) In this topology, the number of transitions grows linearly with the number of Match states.

is that the number of transitions grows rapidly as the number of match states increases. To infer the transition probabilities, we would need a very large set of training data, in which all deletions of all possible sizes were represented. An alternative approach that requires fewer parameters is to model long deletions as sequences of short ones, as seen in the HMM in Fig. 5.12c.

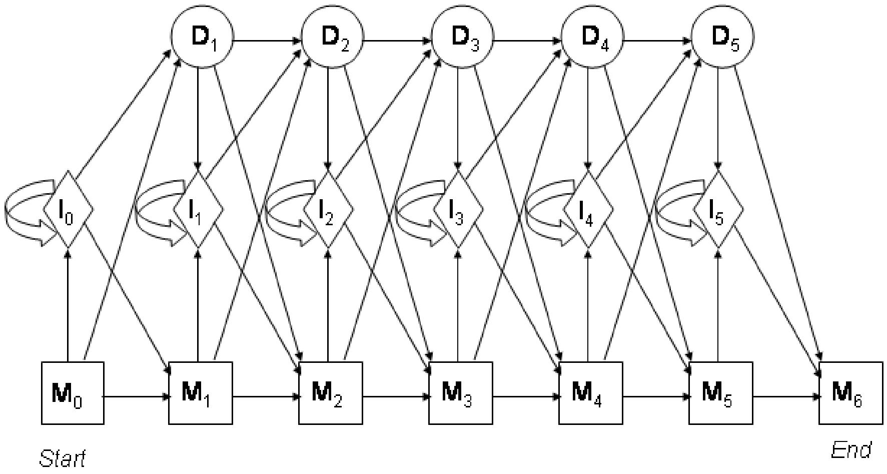


Figure 5.13: A profile HMM of length 5

The canonical Profile HMM, shown in Fig. 5.13, combines these features. A Profile HMM has a column containing a Match, an Insertion, and a Deletion state for each position in the conserved pattern. States M_i, I_i, and D_i correspond to the ith position in the pattern. We refer to the number of Match states, not including the silent Start and End states, as the *length* of a Profile HMM. A leading insertion state, I_0, allows for patterns that occur in the middle of a longer sequence. If the pattern ends before the end of the sequence, the remaining sequence is emitted by the insertion state I_n, where n is the last position in the pattern.

Note that in a Profile HMM there is a path from the Start state, M_0, to the End state, M_{n+1}, that passes only through Insertion and Deletion states. Thus, a Profile HMM can emit a sequence that does not contain an instance of the pattern. Such a sequence would have a low probability, compared with a sequence generated by the Match states.

Parameter estimation

The emission and transition probabilities of a Profile HMM must be estimated from data. If the sequences are aligned and the position of the motif in the alignment is known, then we have labeled training data. In other words, it is possible determine from the alignment which state is associated with each symbol in each sequence. In that case, all we need to do is determine the number of Match states in the Profile HMM, set up the topology, and calculate the parameter values from the labeled data.

Given unlabeled sequences that are known to have a common pattern, we can use the Profile HMM to discover the pattern using the Baum-Welch algorithm to infer the values of the parameters. Once the parameters have been estimated, we use the Profile HMM to label the data. Finally, we construct a multiple sequence alignment by planning symbols with the same label in the same column of the alignment. We give an example of each case below.

Constructing a profile HMM for a variable length motif with labeled data: Profile HMMs, like the one show in Fig. 5.13, can be used to model variable length motifs. We illustrate this process with this example:

```
VG--H
V---N
VE--D
IAADN
```

The length of the Profile HMM should correspond to the typical length of the motif. A rule of thumb is to use the average of the length of the training sequences. The lengths of the above sequences are 3, 2, 3, and 5 (before the gaps were inserted), respectively, yielding an average length of 3.25. This suggests that a Profile HMM of length 3 is appropriate for modeling this pattern. Our HMM will have a silent Start state M_0, Match states M_1, M_2, M_3, Insertion states, I_0, I_1, I_2, I_3, Deletion states, D_1, D_2, D_3, and a silent End state, M_4.

The Insertion and Match states emit the 20 amino acids (for protein motifs), or the four nucleotides (for DNA and RNA motifs). Deletion states emit the indel symbol, e.g. "–". For our Profile HMM, the emission probabilities of the Insertion and Deletion states might look like this:

$$e_{D_j}(\sigma) = 0, \quad e_{D_j}(-) = 1$$

$$e_{I_j}(\sigma) = p(\sigma), \ \forall I_j$$

where $p(\sigma)$ is the background probability of residue σ. In order to estimate the parameters for the Match states, we assign labels to the data using the multiple alignment as a guide.

Columns in the alignment that have gaps in less than half of the rows correspond to Match states. Those with more gaps in than half of the rows correspond to Insertion states:

V	G	-	-	H
V	-	-	-	N
V	E	-	-	D
I	A	A	D	N
M_1	M_2	I_2	I_2	M_3

This yields the following labeled sequences:

V	G	H
M_1	M_2	M_3

V	_	H
M_1	D_2	M_3

V	E	D
M_1	M_2	M_3

I	A	A	D	N
M_1	M_2	I_2	I_2	M_3

Note that when a gap ('_') appears in a Match column (M_i), it is labeled as a deletion (D_i). For example, the first gap in the second sequence is labeled D_i.

From these labeled sequences, we can estimate the emission and transition probabilities from Equations 5.6 and 5.7. For example, using $b = 1$ as a pseudocount, we obtain

$$e_{M_1}(V) = \frac{3+1}{4+20}.$$

Similarly, the probability of a transition from M_2 to I_2 is

$$a_{M_2 I_2} = \frac{1+1}{(2+1)+(1+1)+(0+1)} \, .$$

The three sums in parentheses in the denominator correspond to all possible transitions out of state M_2. The first term in each sum in the number of transitions observed in the training data; the second term is a pseudocount. In the training sequences in our example, there are two transitions from M_2 to M_3, one transition from M_2 to I_2 and no transitions from M_2 to D_3. The other emission and transition probabilities are calculated the same way. The model always starts in M_0, so $\pi_{M_j} = 0$, when $j > 0$.

Modeling unlabeled data with a Profile HMM: To discover a pattern in unlabeled data requires the following steps:

1. **Estimating the length:** Given a set of unaligned sequences, where each sequence is an instance of the pattern, we set the length of the HMM (i.e., the number of non-silent Match states) to L, where L is the average sequence length. An example of this type of input would be sequences approximately 50 residues long, where each sequence corresponds to an instance of the Ig domain.
   Alternatively, we might be given sequences that contain a pattern, but are much longer than the pattern. In this case, we must rely on biological knowledge to obtain an initial estimate of the pattern length. The initial estimate of the pattern length can be adjusted later using model surgery (Step 6).
   An example of this type of input would be a set of protein sequences, typically several hundred residues in length, each of which contains an instance of an unknown domain. In this case, you might estimate the length of the pattern to be approximately 100, since that is the length of a typical protein domain.

2. **The topology:** Construct a Profile HMM with $L + 2$ Match states, $L + 1$ Insertion states, and L Deletion states. M_0 and M_{L+1} are silent states corresponding to the Start state and the End state.

3. **Parameter estimation:** Guess "good" initial parameters (e.g., $a_{M_i\, M_j} \gg a_{M_i\, I_j}$ and $a_{M_i\, M_j} \gg \; a_{M_i D_j}$) and train the model using the Baum Welch algorithm.

4. **Determining the motif:** Use the Viterbi algorithm or posterior decoding to infer the state path that emitted each sequence. The Viterbi recurrence can be greatly simplified and expressed in terms of log odds for the special case of Profile HMMs (Durbin, pp 108-110). The log odds formulation avoids underflow and reduces length effects. This was not covered in class and you will not be tested on it. Note the similarity to the dynamic programming algorithm for pairwise alignment.

5. **Multiple Sequence Alignment:** The most likely paths for each sequence obtained from decoding can be used to obtain a multiple alignment of the input sequences. If symbols O^c_t and O^d_u were emitted by the same Match state, then align position t in sequence O^c with position u in sequence O^d. See Ewens and Grant, p 337 - 339 for a discussion and example of multiple sequence alignment using Profile HMMs.

6. **Model surgery:** The topology of the model can be iteratively refined. If more than half of the sequences enter the Delete state, D_j, then remove M_j, D_j, and I_j from the topology. If more than half of the sequences enter the Insertion state, I_j, then add

Match, Insertion and Deletion states between positions j and $j+1$.

7. **Re-estimate the parameters:** If the states change due to model surgery, the parameters must be re-estimated. Label the multiple alignment with the new states and calculate the transition and emission probabilities as described above for labeled data. If the number of states that changed is a substantial fraction of the entire HMM, then you may obtain better results by retraining with the Baum Welch algorithm.

Compared with the exact dynamic programming algorithm for multiple sequence alignment, which runs in exponential time, this approach can align many sequences quickly.

Pattern recognition with profile HMMs: Once you have constructed your Profile HMM, how do you determine whether a new, unlabeled sequence, O, contains the motif?

If you have a model for a suitable null hypothesis, H_0, you can obtain a log odds ratio,

$$\log \frac{P(O|H_A)}{P(O|H_0)} \,,$$

using the Forward algorithm to determine the probability of the sequence for each model. Typically, H_A would be represented by a profile HMM and H_0 by a background model such as the one shown in Fig. 5.11. This gives a score, but does not infer the location of the motif.

Alternatively, you can find the most likely path using the Viterbi algorithm or posterior decoding. The location of the motif corresponds to the symbols emitted by the Match states. If no symbols were emitted by Match states, then the motif is not present in O.