HMM Lecture Notes

Thursday, November 16th

Rose Hoberman and Dannie Durand

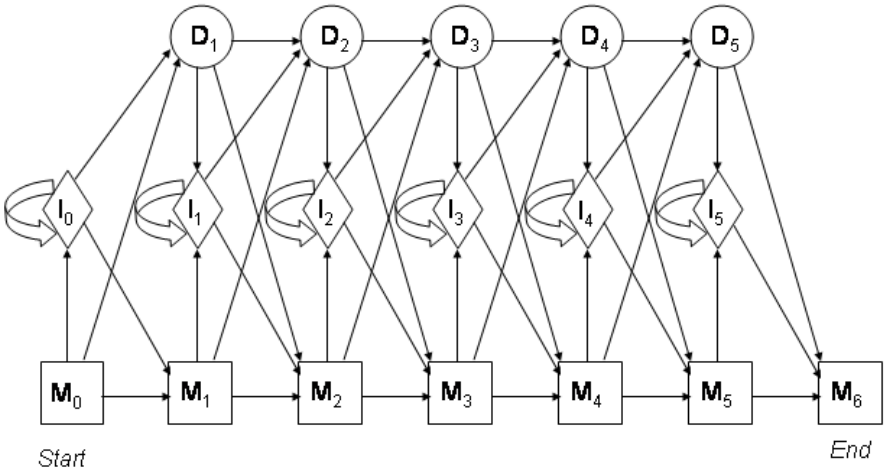
1 Notation

1. N states ($S_1..S_N$)
2. M symbols in alphabet, Σ
3. parameters, λ:
    1. initial distribution of states $\pi(i)$
    2. transition probabilities $a_{ij} = P(q_t = S_i | q_{t-1} = S_j)$. Note that $\sum_{i=1}^{N} a_{ij} = 1, \ \forall j$
    3. emission probabilities $e_i(a)$ probability state i emits a
4. Sequence of symbols: $O = O_1, O_2, ..., O_T$
5. Sequence of states: $Q = q_1, q_2, ..., q_T$

2 Profile HMMs

A *Profile HMM* is a standard topology for modeling sequence motifs. It was proposed by Krogh and Haussler in 1994.

A profile HMM of length 5



Insertion and Match states emit the 20 AA. Delete states emit "-". The emission and transition probabilities must be estimated from data.

Parameter estimation: Given labeled training data (i.e., we are given the state path), we use maximum likelihood to estimate the parameters. In general,

$$e_k(\sigma) = \frac{E_i(\sigma) + b}{\sum_\alpha E_i(\alpha) + 20b}$$

$$a(i,j) = \frac{A(i,j) + b}{\sum_l [A(i,l) + b]}$$

where $E_i(\sigma)$ is the number of instances in the training data where symbol σ is emitted in state i and $A(i,j)$ is the number of transitions from i to j in the training data, and b is a pseudocount to take transitions that are not observed into account (see page 2 for an example.)

For our Profile HMM, the estimation of the emission probabilities in state S_k might look like this:

$$e_{M_6}(a) = e_{M_0}(a) = 0$$

$$e_{I_k}(a) = p_a$$

$$e_{D_k}(a) = 0 \qquad e_{D_k}(\text{“}-\text{”}) = 1$$

$$e_{M_k}(a) = \frac{E_k(a) + b}{\sum_j E_k(j) + 20b}$$

where p_a is the background frequency of residue $a \in \Sigma$.

Constructing a Profile HMM

Labeled data: We can use the Profile HMM formalism to model a shared pattern in biomolecular sequences. If the sequences are already aligned, then we have labeled data. In other words, we can 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 determine the parameters from the labeled data.

Profile HMM's like the one above can be used to model variable length motifs, such as this one:

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

The length of the HMM should be the average of the length of the sequences. The above sequences are of lengths 3, 2, 3 and 5, respectively, yielding an average of 3.25. 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.

In order to estimate the parameters, we need to assign labels to the data using the multiple alignment. Positions in the alignment that have gaps in less than 50% of the rows correspond to match states. Those with more than 50% gaps 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	_	N
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

From these labeled sequences, we can estimate the parameters. For example, using $b = 1$ as a pseudocount, we obtain

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

and

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

The three sums in the denominator correspond to all possible transitions out of state M_2, plus pseudocounts. Specifically, in the training sequences there are two transitions from M_2 to M_3, one one transition from M_2 to I_2 and no transitions from M_2 to D_3.

Unlabeled data: Given unlabeled sequences that are known to share a pattern, we can use the Profile HMM to discover the pattern, label the data, and construct a multiple sequence alignment.

An example of this is given in Ewens and Grant, pp. 337 - 339.

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, let L, the length of HMM (i.e., the number of match states) be the average length of sequences. An example of this type of input would be sequences ≈ 50 residues long, where each sequence corresponded to a different instance of the Ig domain. If you are given sequences that contain a pattern but are much longer than the pattern, then you need to some approach to estimating the length. An example of this type of input would be a set of protein sequences, each 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 ≈ 100, since that is the length of a typical domain.

2. **The topology:** Construct a Profile HMM with $L+2$ match states. M_0 and M_{L+1} are silent states corresponding to the start state and the end state.

3. **Learn parameters:** Guess "good" initial parameters (e.g., $a_i(M_j) >> a_i(I_j)$ *or* $a_i(D_j)$). Train model using Baum Welch.

4. **Determining the motif:** Use the Viterbi algorithm ($\pi^* = argmax_j P(\pi, s_j) \forall s^j$) or posterior decoding to find path most likely to produce each sequence. The Viterbi recurrence can be greatly simplified and expressed in terms of log odds for the special case of Profile HMMs. The log odds formulation avoids underflow and to reduces length effects. This was not covered in class but you are responsible for reading the sections on the specialized forms of both the Viterbi and Forward algorithms for Profile HMMs are given in Durbin, pp 108-110. Note the similarity to the dynamic programming algorithm for pairwise alignment.

5. **Multiple Sequence Alignment:** For each sequence, the path determined using Viterbi or posterior decoding in Step 4 can be used to obtain a multiple alignment of the input sequences. If O_t^d and O_u^c were emitted by same match state, then align positions t and u. 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_i, then remove M_i, D_i, and I_i the topology. If more than half of the sequences enter the insertion state I_i, then add match, deletion and insertion states between M_i and M_{i+1} (number equal to average length of the insertion).

7. **Re-estimate the parameters:** If the states change due to model surgery, you will need to re-estimate the parameters. 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 are changed is a significant percentage of the entire HMM, then you may want to retrain with Baum Welch.

Compared with the exact dynamic programming algorithm for multiple sequence alignment, which runs in exponential time, this approach can align many sequences quickly. Note that this method doesn't say how to align indel sequences of different length. Correspond to unconserved portions, not meaningfully alignable. Often just left-justified and shaded.

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

- Calculate $\log \frac{P(O|H_A)}{P(O|H_0)}$ using the Forward algorithm. This gives a score but doesn't tell us the location.
- Find the most likely path using the Viterbi algorithm. 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. You could also use posterior decoding.

There are specialized versions of the Forward and Viterbi algorithms for profile HMM's (see Durbin, pp 109-110.)