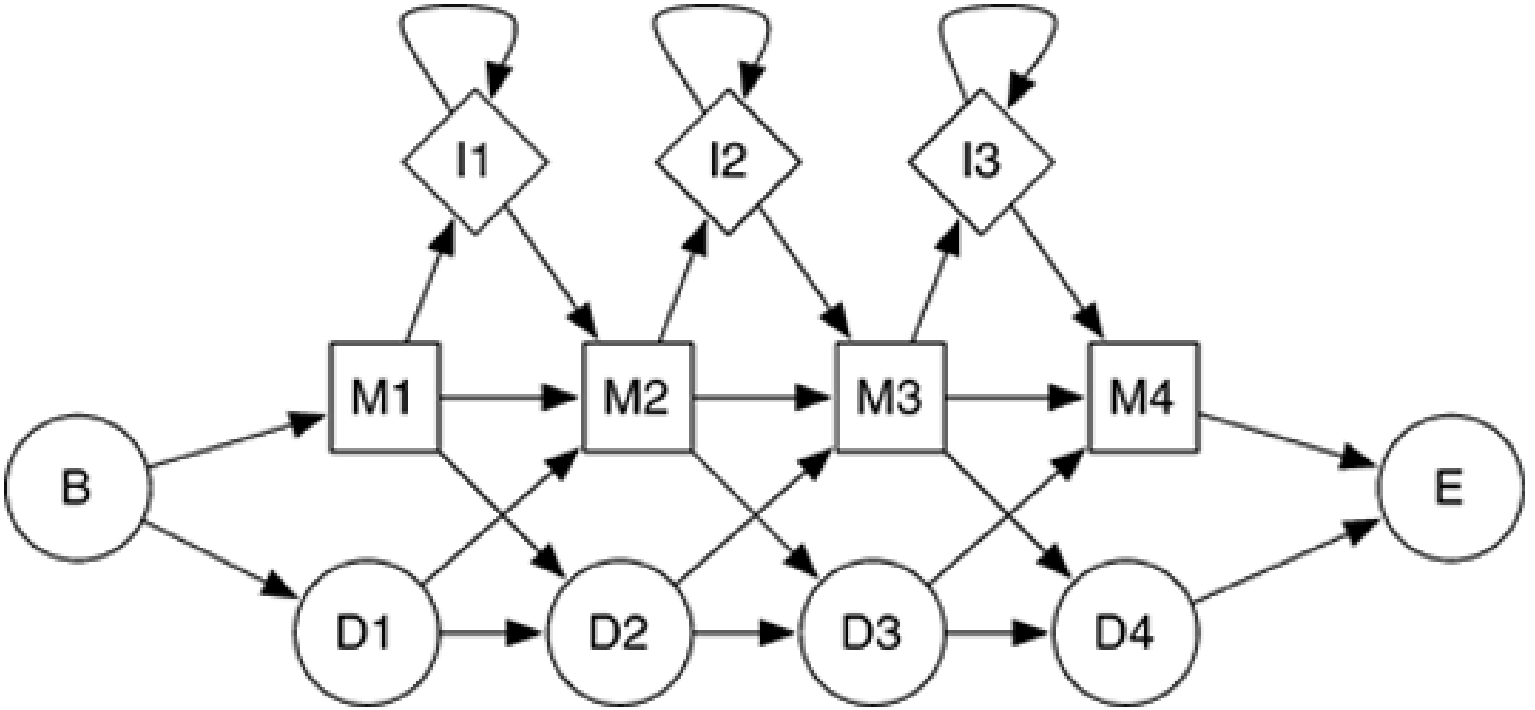
HMM Models of Alignment

Motivations

- As you may have noticed, the dynamic programming algorithm for finding optimal sequence alignments is just equivalent to the Viterbi algorithm as we learned it for HMMs.
- Implies alignment is equivalent to an HMM.
- We can gain insight into different aspects of alignment by looking at it from this different (HMM) point of view.
- Viterbi results are hard to interpret because they give no confidence values. Recasting this as an HMM would allow us to compute explicit posterior probabilities for each step of the alignment.

Profile HMM

We've already seen an HMM model of alignment, namely profile HMMs:



This is a standard, first-order HMM; we index the observation sequence $\vec{X}^n$ as usual by position ie. X_t But note how the sequence represented by the HMM is indexed only *implicitly*, i.e. each position in that sequence is represented by a different set of states in the HMM.

Profile HMM Computation

If we run a Viterbi (or Forward-Backward calculation) on this profile HMM, what will the actual calculation look like?

- draw the basic data structure that stores the state of the Viterbi algorithm during the calculation;
- indicate precisely how the different aspects of the problem (i.e. the observation sequence, the HMM states) are represented on this data structure.

Profile HMM Computation Answer

It forms a two-dimensional matrix:

- on the vertical dimension we have the set of states of the HMM $D_1 M_1 I_1 D_2 M_2 I_2 \ldots D_m M_m I_m$.
- on the horizontal dimension we have the observation sequence letters $X_1 X_2 \ldots X_n$.
- We fill in the two dimensional matrix by Viterbi (or forward-backward) in the usual way.
- Any part of $\vec{X}^n$ could align to (be emitted by) any part of the HMM, because the deletion and insertion can shift the observation sequence relative to the HMM sequence.

Profile HMM “Moves”

- Whether we’re doing Viterbi or forward-backward, we follow our standard process.
- I.e. consider all possible moves to (from) a given cell, and write the *max* (or *sum*) of those moves in this cell.

Draw the “moves” incoming to each of the possible states in terms of how they are represented on the matrix.

Profile HMM "Moves" Answer

- moves to the M state are "diagonal", i.e. they advance both the observation index t and our position in the HMM profile.
- moves to the I state are horizontal, i.e. the observation index advances but the profile index does not.
- moves to the D state are vertical, i.e. the profile index advances but the observation index does not.
- Since every profile position has exactly one D state, one M state, and one I state, it is convenient to keep one separate matrix just for the $D_1, D_2, \ldots D_m$ states, another matrix for just the $M_1, M_2, \ldots M_m$ states, and a third matrix for the $I_1, I_2, \ldots I_m$ states. Then the vertical coordinate u in each matrix is just the position in the profile that this coordinate represents.
- All the moves in the M matrix are diagonal; all the moves in the D matrix are vertical; all the moves in the I matrix are horizontal.
- There are also moves between the matrices, e.g. $D \rightarrow M$ and vice versa.

Where is the Hidden Variable?

- For posterior probability calculations, we want to sum $p(\theta_t, \vec{X})$ over all possible states of the hidden variable, to obtain $p(\vec{X})$.
- So far, we're used to thinking of the set of "all possible states" of one hidden variable θ_t as corresponding to a single column of this two-dimensional matrix.
- Will that work here?

Probability Summation with Silent States

- Note that in previous HMM matrices we never had “vertical moves” (i.e. “silent states” that emit nothing) before.

Does this affect how we would calculate the probability of the observation sequence via the forward-backward algorithm?

- if not, why not?
- if so, how?

Probability Summation with Silent States Answer

- When calculating forward and backward probabilities, be very careful to include *emission* components *only* for the M and I states, not the D states!
- The fact that we have silent states (i.e. the D states don't emit anything), means that one position in the observation sequence can traverse multiple D states.
- The "moves" on the matrix no longer all come from the previous column: moves to **D** states are "vertical" (i.e. they stay in the same column).
- We can no longer sum $p(\vec{X}) = \sum_{\theta_t} p(\vec{X}, \theta_t)$ over the whole column, because the states are no longer mutually exclusive. E.g. at $t = 1$ we could traverse states $D_1 \rightarrow D_2 \rightarrow D_3 \rightarrow \ldots$.

What Are the Hidden Variables?

- because of these complications, less obvious exactly what we should take as our hidden variables.
- traditionally, each successive state transition equals one more hidden variable. Concretely, there is uncertainty about what the next state will be at each step, so each step in the path is a hidden variable forming a chain $\theta_1 \rightarrow \theta_2 \rightarrow \theta_3 \rightarrow \ldots$
- We could still do that, but note that the "path step index" w has a non-trivial mapping to the observation and profile indexes t, u; specifically $w = t + u - N_M$, where N_M is the number of match states traversed so far on this path.
- More convenient: note that on any given path, one position X_t can still be emitted by only one M or I state. We can refer to one such specific emission as $\theta_t = M_u$, which means state M_u in observation column t in the matrix.

Total Likelihood Computation

Propose what you think would be the simplest way to compute $p(\vec{X})$ over this matrix structure.

Total Likelihood Computation Answer

- Best to think of $p(\vec{X})$ as summing over *all possible paths* across the matrix.
- Easiest place to sum it is at the places where all the paths converge to a single state, i.e. START or END. i.e.

$$p(\vec{X}) = f_{n+1,END} = b_{0,START}$$

- Also possible to sum over all possible M_u and I_u states that could emit a specific observation letter X_t:

$$p(\vec{X}) = \sum_{u=1}^{m} f_{t,M_u} b_{t,M_u} + \sum_{u=1}^{m} f_{t,I_u} b_{t,I_u}$$

Posterior Computation

- Computing the posterior that a specific state M_u emits a specific letter X_t is just

$$p(\theta_t = M_u | \vec{X}^n) = \frac{p(\theta_t = M_u, \vec{X}^t) p(\vec{X}_{t+1}^n | \theta_t = M_u)}{p(\vec{X}^n)}$$

$$= \frac{f_{t,M_u} b_{t,M_u}}{p(\vec{X}^n)}$$

Forward Computation

- Draw the move(s) that must be taken into account for computing the forward probability f_{t,M_u} (i.e. for aligning letter X_t to profile letter M_u).
- Write the corresponding recursion equation for f_{t,M_u}.

Forward Computation Answer

- The possible moves are all diagonal (because the M state advances both t, u), originating from $\theta_{t-1} = D_{u-1}, I_{u-1}, M_{u-1}$.
- The recursion has the usual forward, transition and emission components:

$$f_{t,M_u} = \sum_{s=D,I,M} f_{t-1,s_{u-1}} p(M_u | s_{u-1}) p(X_t | M_u)$$

Backward Computation

- Draw the move(s) that must be taken into account for computing the backward probability b_{t,M_u} (i.e. for aligning letter X_t to profile letter M_u).
- Write the corresponding recursion equation for b_{t,M_u}.

Backward Computation Answer

The possible moves are

- $M_u \to M_{u+1}$ (diagonal);
- $M_u \to D_{u+1}$ (vertical);
- $M_u \to I_u$ (horizontal);

The recursion has to be written carefully, because the details of the transition and the emission depend on exactly which state we are transitioning to:

$$b_{t,M_u} = b_{t+1,M_{u+1}} p(M_{u+1}|M_u) p(X_{t+1}|M_{u+1})$$

$$+ b_{t+1,I_u} p(I_u|M_u) p(X_{t+1}|I_u)$$

$$+ b_{t,D_{u+1}} p(D_{u+1}|M_u)$$

Deletion Posteriors?

- What if we want to know the posterior $p(D_u|\vec{X}^n)$ that profile position u is deleted (doesn't emit anything)?
- Note that this isn't tied to a specific observation position t, because any part of $\vec{X}^n$ could align to any part of the HMM.
- We need to know $p(D_u, \vec{X}^n)$. As usual we get this by summation

$$p(D_u, \vec{X}^n) = \sum_{t=1}^{n} p(\theta_t = D_u, \vec{X}^t) p(\vec{X}_{t+1}^n | \theta_t = D_u)$$

- Note that $p(\vec{X}_{t+1}^n | \theta_t = D_u)$ will include a contribution from $p(\vec{X}_{t+1}^n | \theta_t = D_{u+1})$. Transitioning from $D_u \rightarrow D_{u+1}$ does not advance t!

What If We Score Every Position the Same?

- We could still use our profile HMM, but it seems unnecessarily complicated.
- We only need a single M state, a single D state, and a single I state.
- But how are we going to track what position in the profile sequence we're at?
- Treat the profile sequence as another observation sequence $\vec{Y}^m$ with position index u.
- Now we have to track not just index t but also u as we move between states.
- **M** increments both t, u; **I** increments only t; **D** increments only u.

HMM with Two Observation Sequences

- Again, any part of $\vec{X}$ can align against any part of $\vec{Y}$.
- A single state (e.g. M) is represented not by a one dimensional row (all possible θ_t) but instead by a two dimensional matrix (all possible $\theta_{t,u}$).
- Again, best to think of $p(\vec{X}, \vec{Y})$ as summed over all possible paths across the matrices. Again, this summation is easiest to do at the START state (i.e. $p(\vec{X}, \vec{Y}) = b_{0,START}$).
- To get the posterior that X_t, Y_u align, just compute it from $p(\theta_{t,u} = M, \vec{X}, \vec{Y})$.
- To get the posterior that X_t is *not* aligned to $\vec{Y}$, need to sum over all possible positions in Y, i.e. $1 - \sum_{u=1}^{m} p(\theta_{t,u} = M, \vec{X}, \vec{Y})$