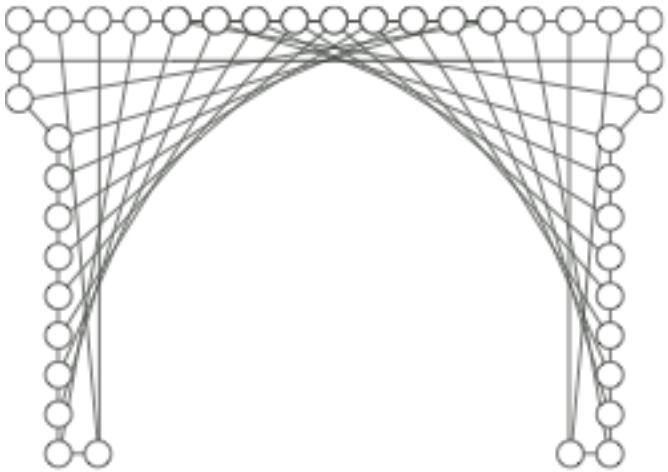
Bioinformatics

Richard Bonneau

Lecture 4: alignment via HMMs.



NEW YORK UNIVERSITY
CENTER FOR COMPARATIVE
FUNCTIONAL GENOMICS

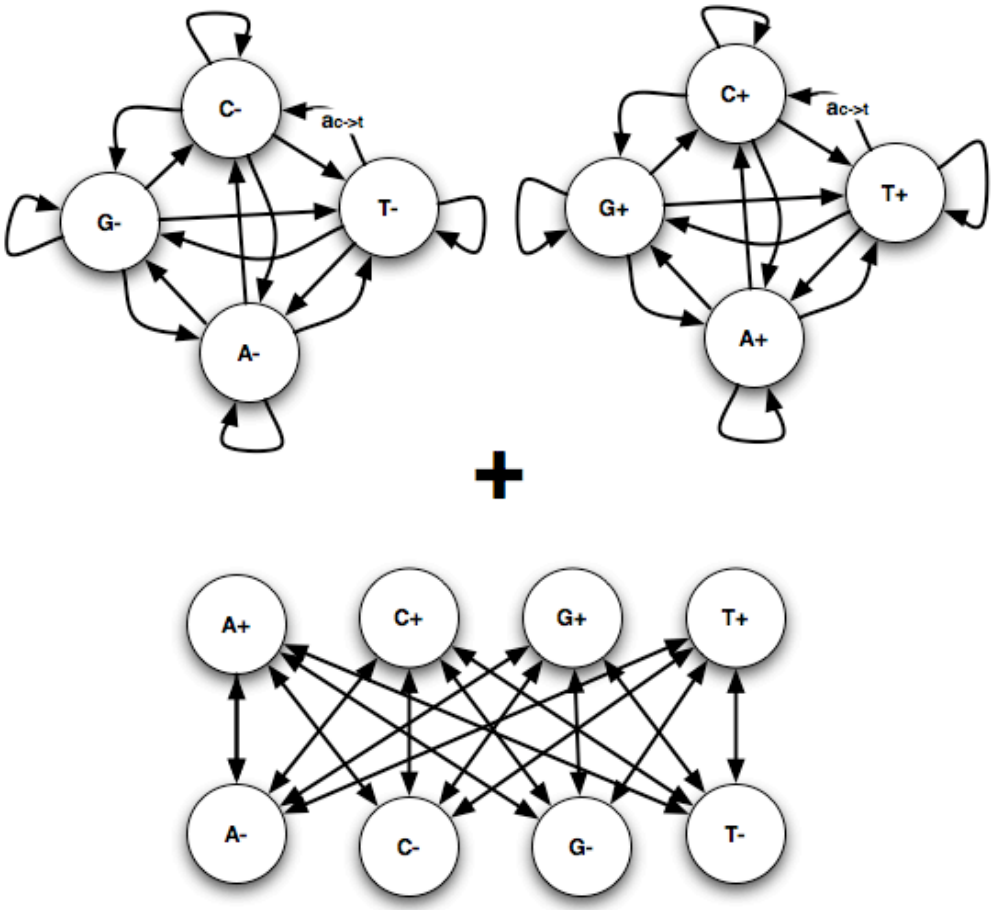
COURANT
INSTITUTE

Associated reading.

- Durbin, Eddy, et al. Ch. 4
- Papers for discussion section:
  *MAIN: Pfam: Nucleic Acids Res. 1998 Jan 1;26(1):320-2.
- AND: Proteins. 1997 Jul;28(3):405-20.

an HMM to find CpG islands

Model description: p(x,π)



$$trans: a_{kl} = P(\pi_i = l \mid \pi_{i-1} = k)$$

$$emit: e_k(b) = P(x_i = b \mid \pi_i = k)$$

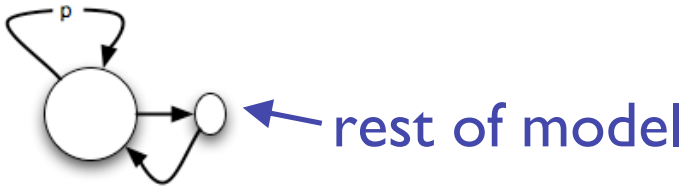
$$P(x \mid \pi) = a_{0\pi_1} \prod_{i=1}^{L} e_{\pi_i}(x_i) a_{\pi_i, \pi_{i+1}}$$

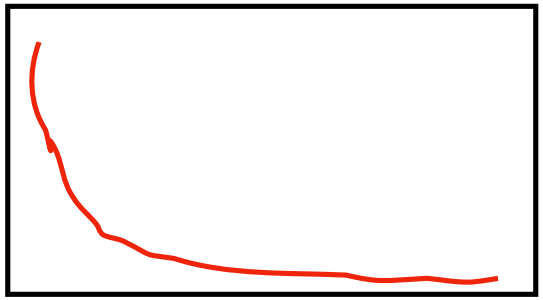
Model structure:

Duration modeling: Effect of model structure on length of any stretch of seq in one state.

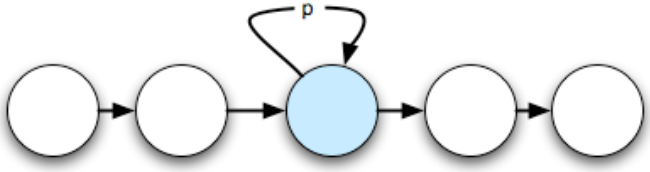
Model structure effects duration of any state in subtle ways:

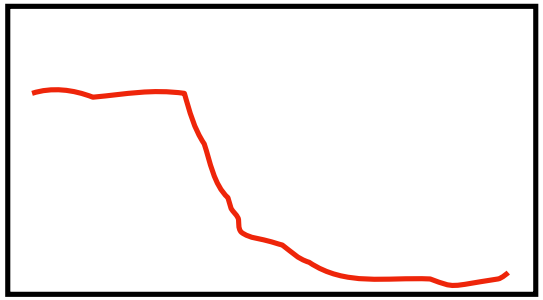
self loop p(n in a row = k) $=(n-p)p^{n-1}$



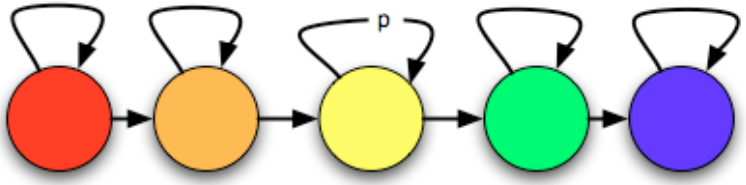


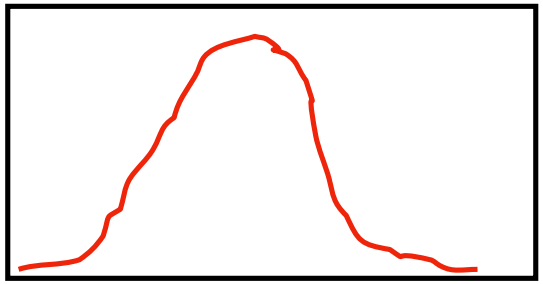
make a min # in a row by embedding geometric in linear path.





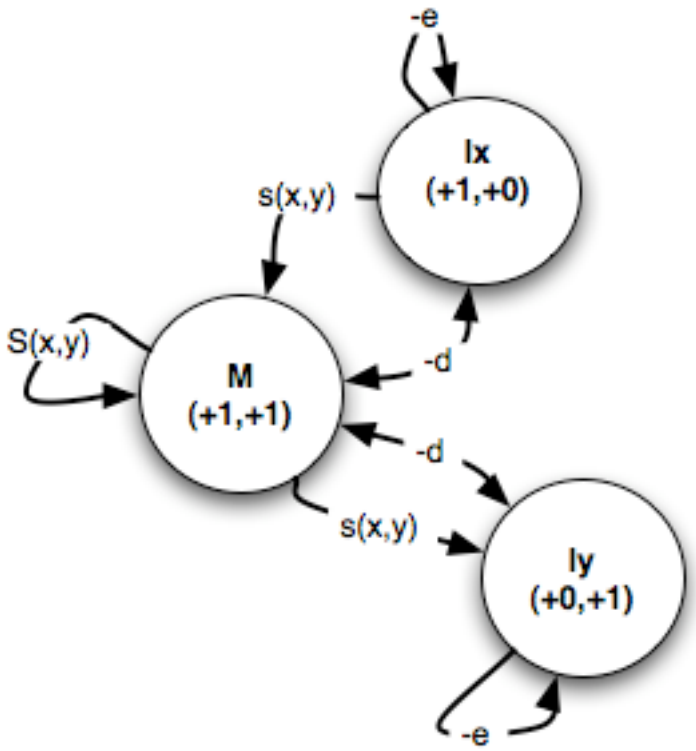
negative binomial





An FSA representation of pairwise algnment

Filling F(i,j) using a Finite State Automaton



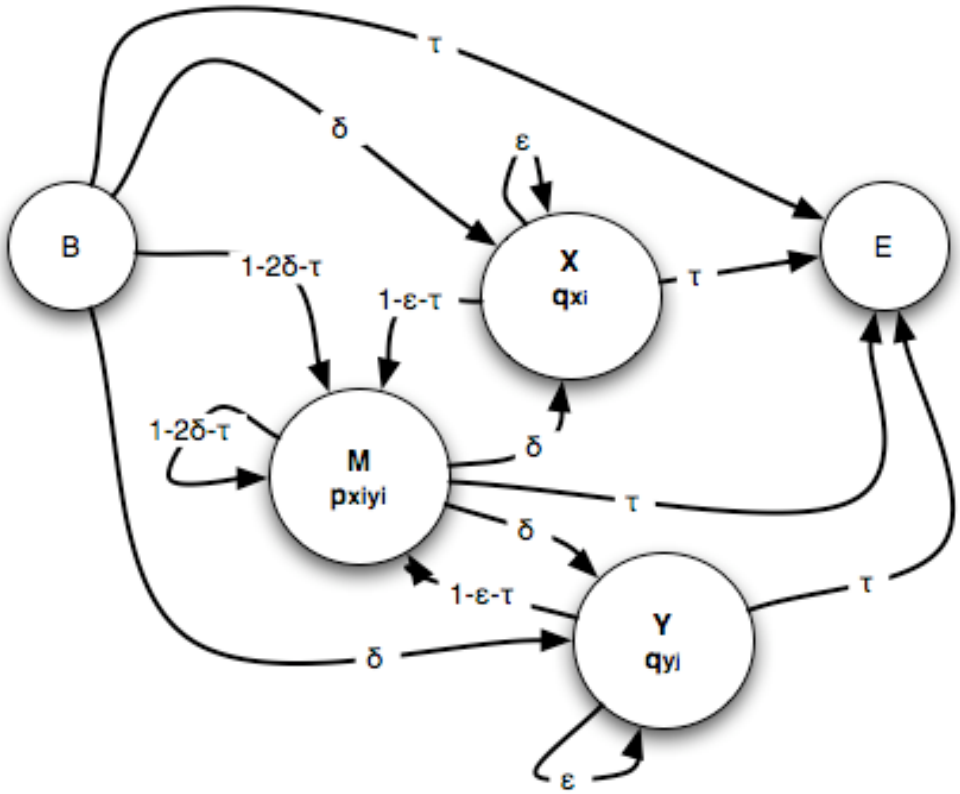
Where M means diagonal up pointer and I means adding an insertion in either seq x or seq y (vertical or horizontal pointer from F(i,j).

d, s(x,y) are the same gap penalty and match score from before.

This is just an example of representing the same algorithm as states and transitions between states.

An HMM representation of pairwise algnment (i)

A simple pair HMM



Each state emits a pair: either an x,y; x,- or y,- pair.

Viterbi for any state is now:
$v^k(i,j)$

$$v^M(0,0) = 1; v^\bullet(i,0) = v^\bullet(0,j) = 0$$

$$v^M(i,j) = p_{x_i y_j} \max \begin{cases} (1-2\delta-\tau)v^M(i-1,j-1), \\ (1-\varepsilon-\tau)v^X(i-1,j-1), \\ (1-\varepsilon-\tau)v^Y(i-1,j-1) \end{cases}$$

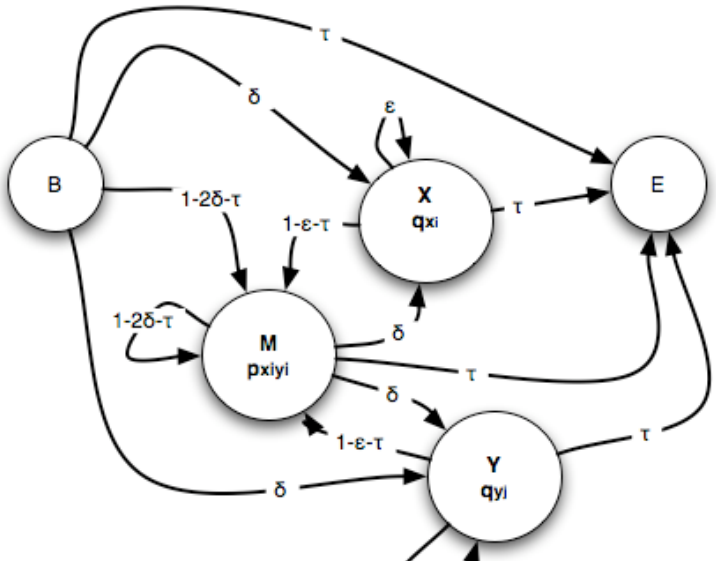
$$v^X(i,j) = q_{x_i} \max \begin{cases} \delta v^M(i-1,j), \\ \varepsilon v^X(i-1,j) \end{cases}$$

$$v^Y(i,j) = q_{y_j} \max \begin{cases} \delta v^M(i-1,j), \\ \varepsilon v^Y(i-1,j) \end{cases}$$

An HMM representation of pairwise algnment (ii)

This is in no way different from last week

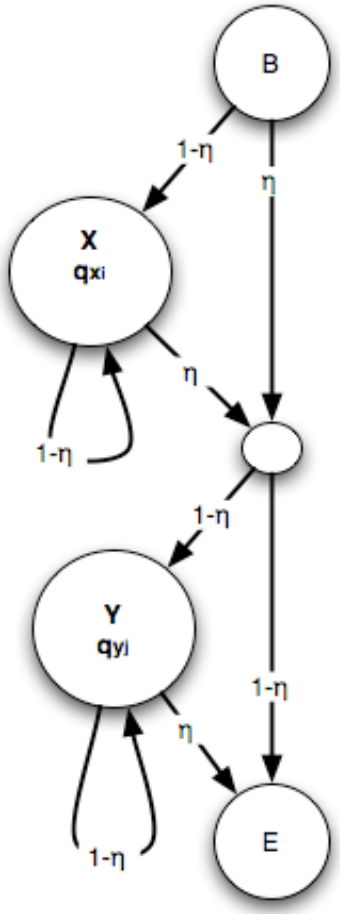
$$init: v_0(0) = 1, v_k(0) = 0, k > 0$$
$$recursion(i = 1...L): v_l(i) = e_l(x_i)\max_k(v_k(i-1)a_{kl})$$
$$ptr = v_l(i) = \arg\max_k(v_k(i-1)a_{kl})$$
$$end: P(x,\pi^*) = \max_k(v_k(L)a_{k0})$$
$$\pi_L^* = \arg\max_k(v_k(L)a_{k0})$$



$$v^M(0,0) = 1; v^\bullet(i,0) = v^\bullet(0,j) = 0$$
$$v^M(i,j) = p_{x_i y_j} \max \begin{cases} (1-2\delta-\tau)v^M(i-1,j-1), \\ (1-\varepsilon-\tau)v^X(i-1,j-1), \\ (1-\varepsilon-\tau)v^Y(i-1,j-1) \end{cases}$$
$$v^X(i,j) = q_{x_i} \max \begin{cases} \delta v^M(i-1,j), \\ \varepsilon v^X(i-1,j) \end{cases}$$
$$v^Y(i,j) = q_{y_j} \max \begin{cases} \delta v^M(i-1,j), \\ \varepsilon v^Y(i-1,j) \end{cases}$$

An HMM representation of pairwise algnment (iii)

A simple pair HMM for the random model

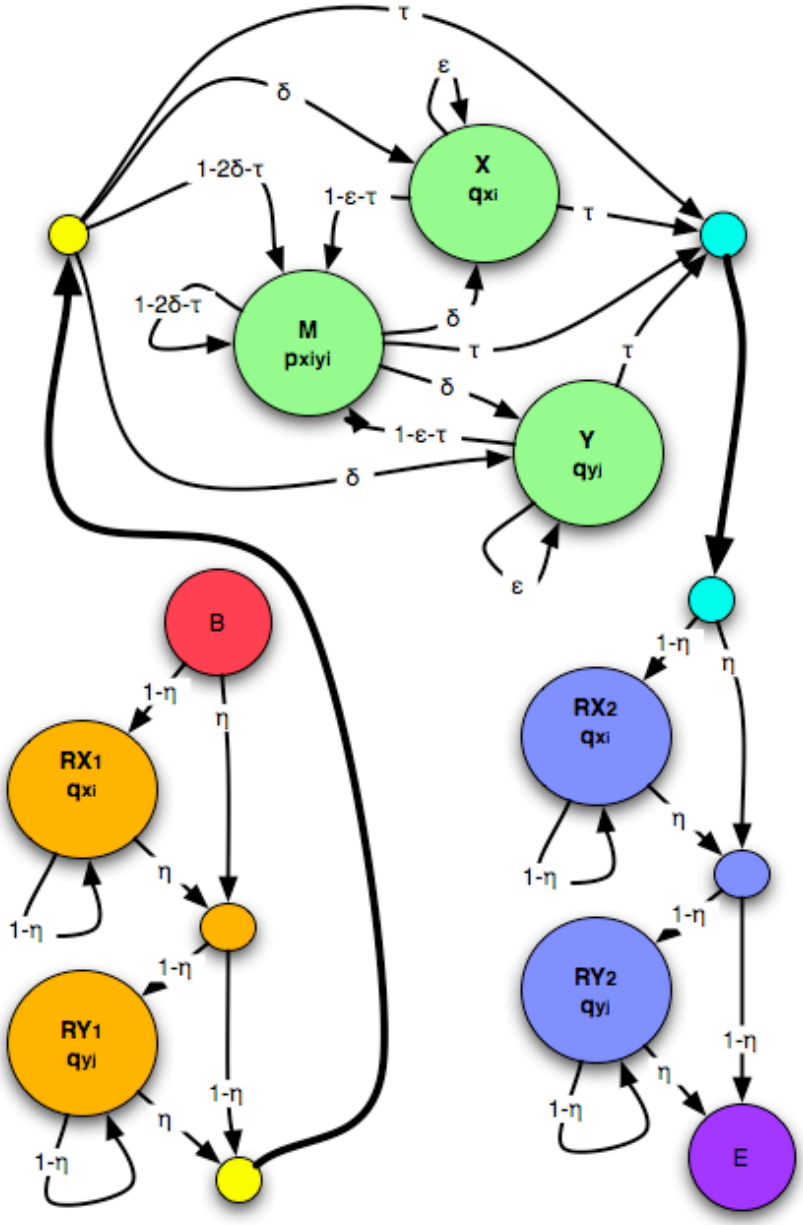


X is emitted, then Y

0 length sequences are allowed

Scores can be derived to give log-odds score (as in lecture 2) s(xi,yj) using Match model from last slide and this model.

An HMM for local pairwise algnment (iv)



We add a random sequence model to each end of our global alignment model.

Strong matches with unaligned flanking regions can be matched by this model with matched region dominating score/p of match.

Scoring alignments

$$p(x,y) = \sum_{\pi} p(x,y,\pi)$$

$$f^M(0,0) = 1; f^\bullet(i,0) = f^\bullet(0,j) = 0$$

$$f^\bullet(i,-1) = f^\bullet(-1,j) = 0$$

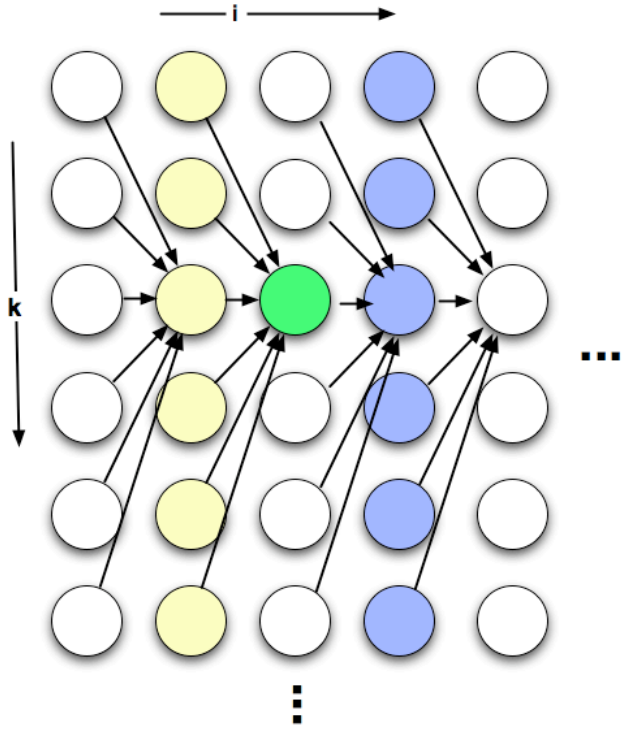
$$f^M(i,j) = p_{x_i y_j} sum \begin{cases} (1-2\delta-\tau) f^M(i-1,j-1), \\ (1-\varepsilon-\tau) f^X(i-1,j-1), \\ (1-\varepsilon-\tau) f^Y(i-1,j-1) \end{cases}$$

$$f^X(i,j) = q_{x_i} sum \begin{cases} \delta f^M(i-1,j), \\ \varepsilon f^X(i-1,j) \end{cases}$$

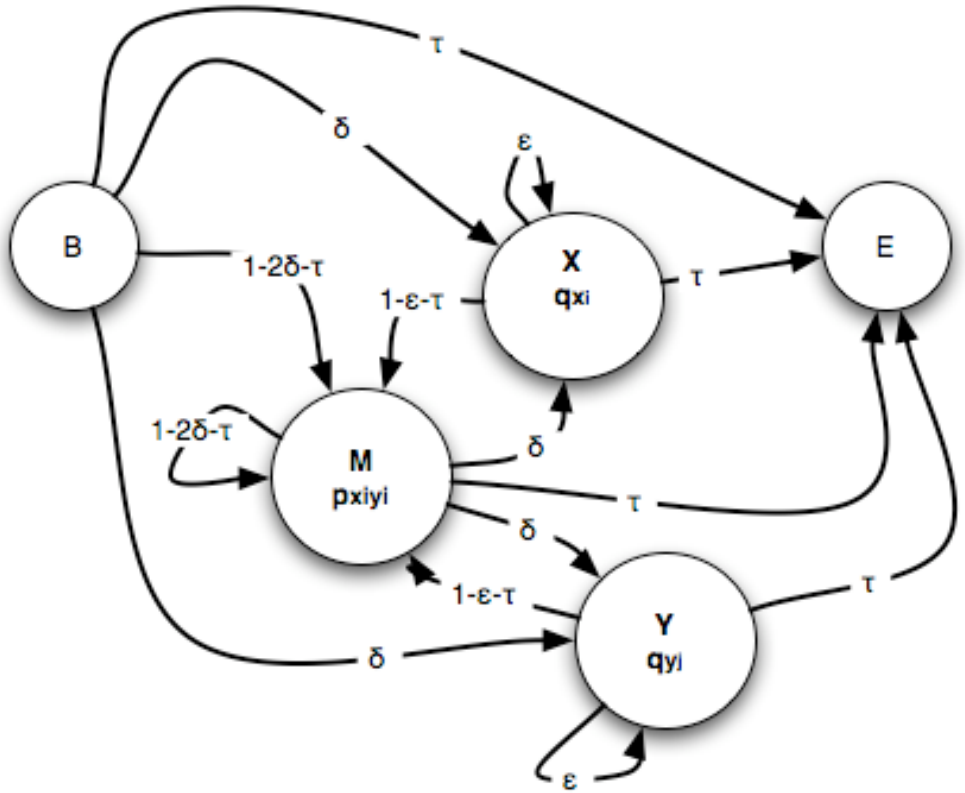
$$f^Y(i,j) = q_{y_j} sum \begin{cases} \delta f^M(i-1,j), \\ \varepsilon f^Y(i-1,j) \end{cases}$$

$$P(x,y) = f^E(n,m) = \tau[f^M(n,m) + f^X(n,m) + f^Y(n,m)]$$

We use forward algorithm to sum probability of x and y being aligned over all possible alignments.



Scoring alignments



$$p(x,y) = \sum_{\pi} p(x,y,\pi)$$

$$f^M(0,0) = 1; f^\bullet(i,0) = f^\bullet(0,j) = 0$$

$$f^\bullet(i,-1) = f^\bullet(-1,j) = 0$$

$$f^M(i,j) = p_{x_i y_j} sum \begin{cases} (1-2\delta-\tau) f^M(i-1,j-1), \\ (1-\varepsilon-\tau) f^X(i-1,j-1), \\ (1-\varepsilon-\tau) f^Y(i-1,j-1) \end{cases}$$

$$f^X(i,j) = q_{x_i} sum \begin{cases} \delta f^M(i-1,j), \\ \varepsilon f^X(i-1,j) \end{cases}$$

$$f^Y(i,j) = q_{y_j} sum \begin{cases} \delta f^M(i-1,j), \\ \varepsilon f^Y(i-1,j) \end{cases}$$

$$P(x,y) = f^E(n,m) = \tau[f^M(n,m) + f^X(n,m) + f^Y(n,m)]$$

=

$$init : f_0(0) = 1, f_k(0) = 0, k > 0$$

$$recursion(i = 1...L) : f_l(i) =$$

$$e_l(x_i) \sum_k (f_k(i-1) a_{kl})$$

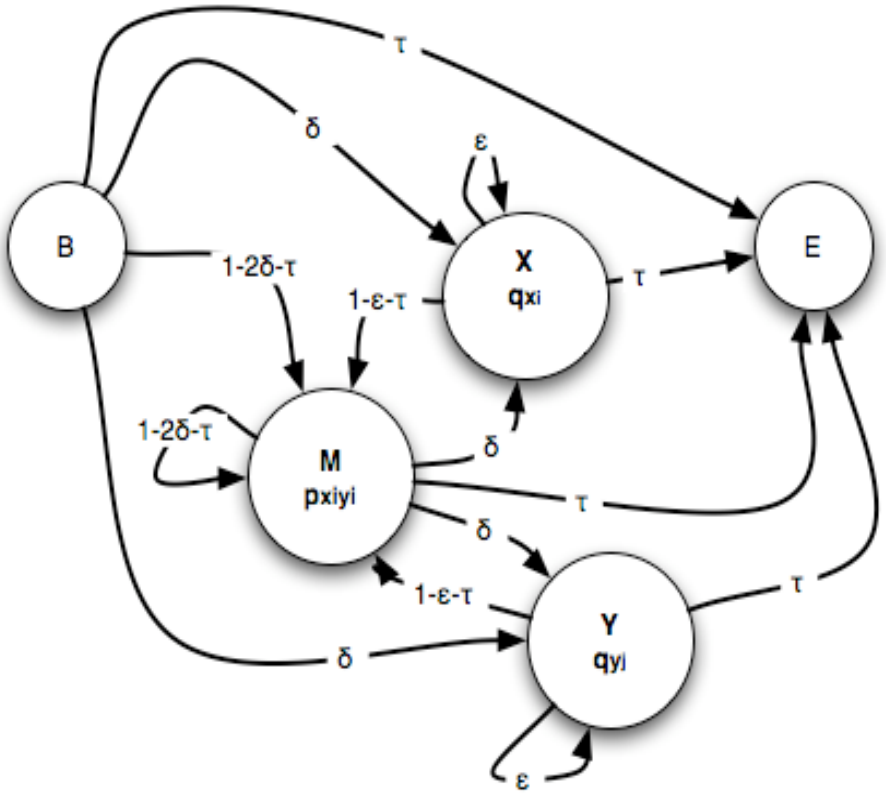
$$term : P(x) = \sum_k f_k(L) a_{k0}$$

Choosing suboptimal alignments

$$M(i,j) \to \begin{cases} M : p_{xi,yj}(1-2\delta-\tau) f^M(i-1,j-1) / f^M(i,j), \\ X : p_{xi,yj}(1-\varepsilon-\tau) f^X(i-1,j-1) / f^M(i,j), \\ Y : p_{xi,yj}(1-\varepsilon-\tau) f^Y(i-1,j-1) / f^M(i,j) \end{cases}$$

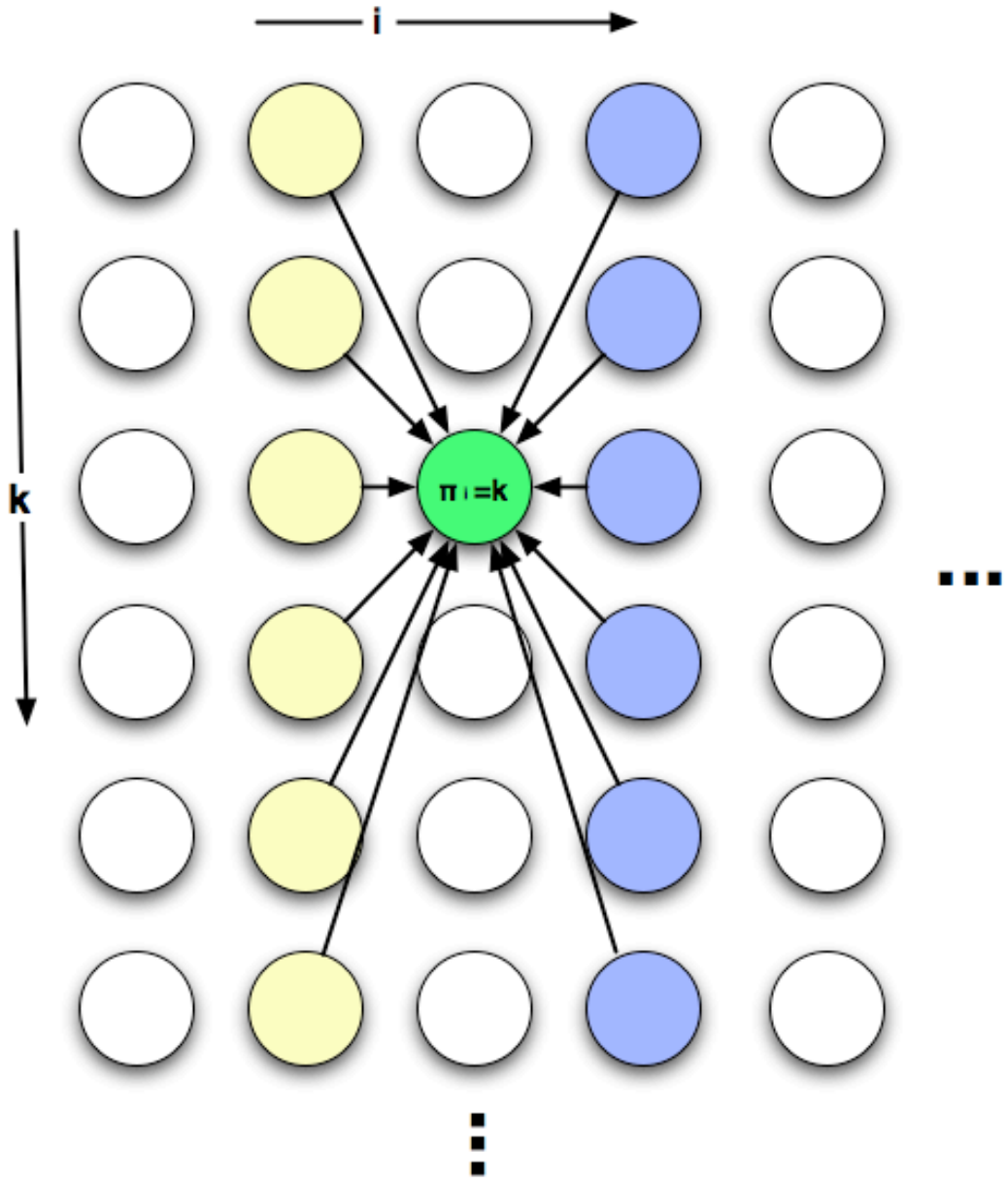
$$X(i,j) \to \begin{cases} M : q_{xi}\delta f^M(i-1,j) / f^X(i,j), \\ Y : q_{xi}\varepsilon f^X(i-1,j) / f^X(i,j) \end{cases}$$

$$Y(i,j) \to \ldots$$



We use the results from the forward algorithm to calc the probability of a given step in the path, and sample from that probability. We thereby generate ensembles of paths.

Using f and b to get $p(\pi_i=k|x)$



At any position in our sequence of observations (x) we want to know the probability that the path goes through state k. We want this to predict the hidden state, given the model and a sequence of observations.

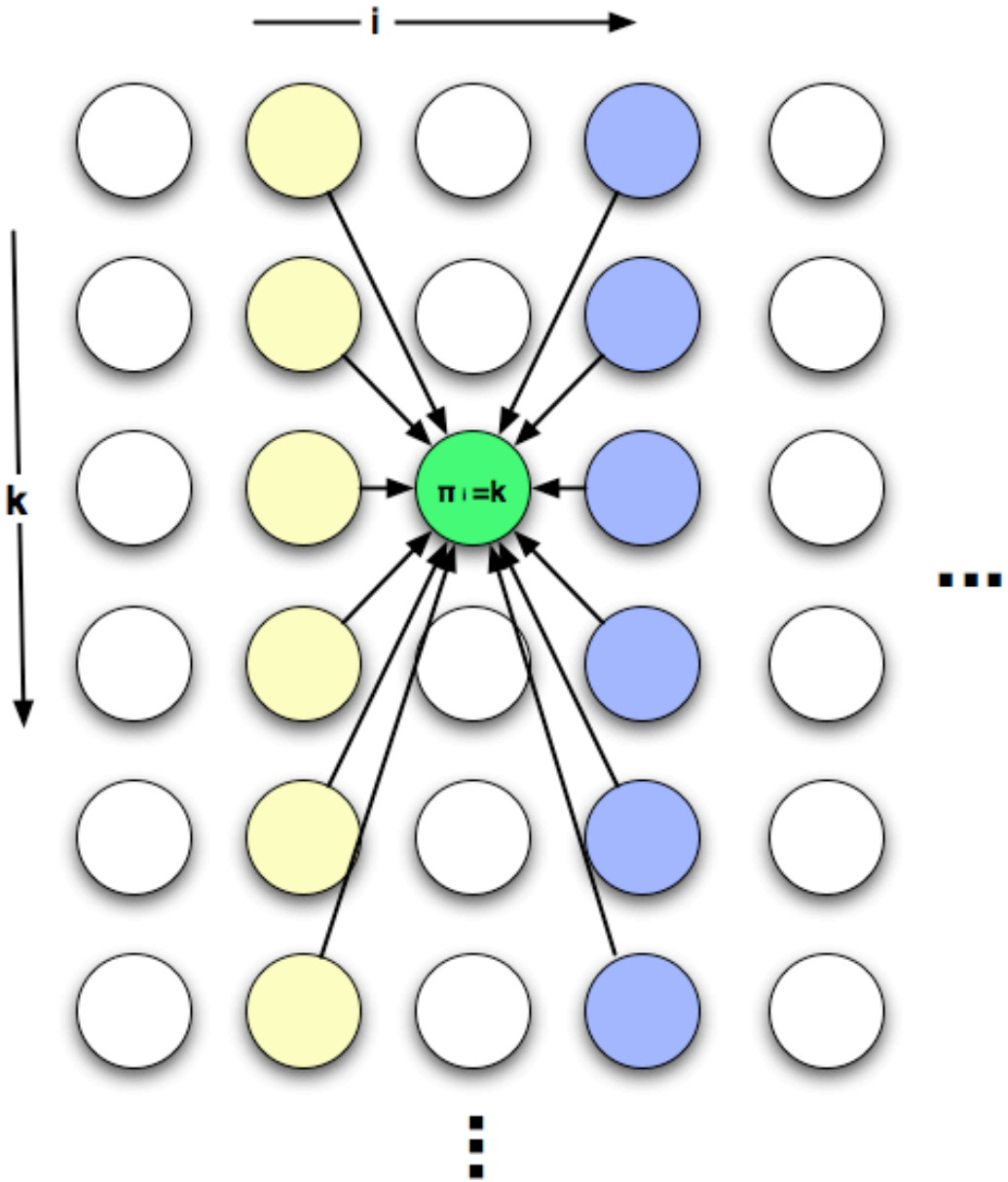
$$P(x,\pi_i = k) = f_k(i)b_k(i)$$

$$P(\pi_i = k \mid x) = \frac{f_k(i)b_k(i)}{P(x)}$$

$$P(X,Y) = P(X \mid Y)P(Y)$$

$$\frac{P(X,Y)}{P(Y)} = P(X \mid Y)$$

Using f and b to get $p(x_i \rightarrow y_j | x,y)$

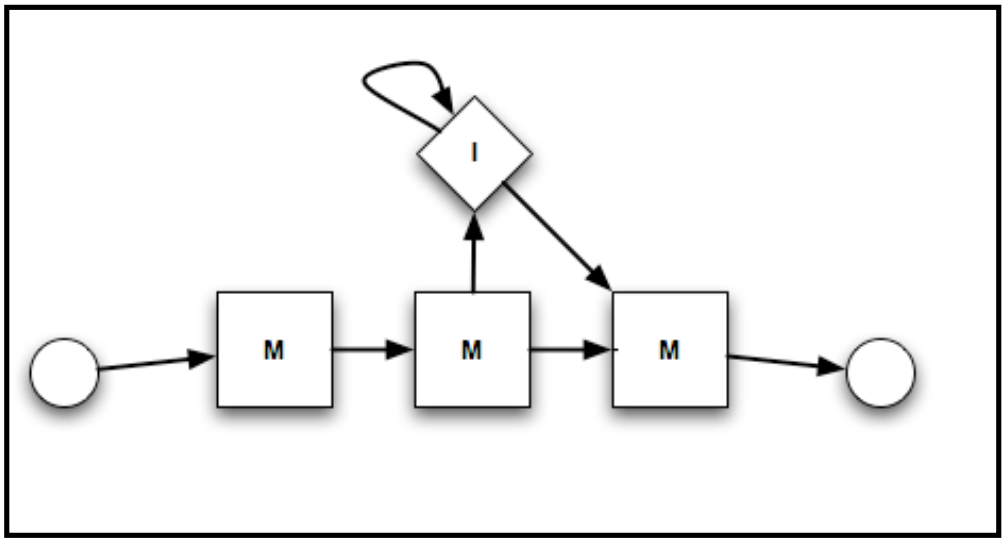


We sum over all paths that pass through the state with xi and yj in the match state M(i,j)

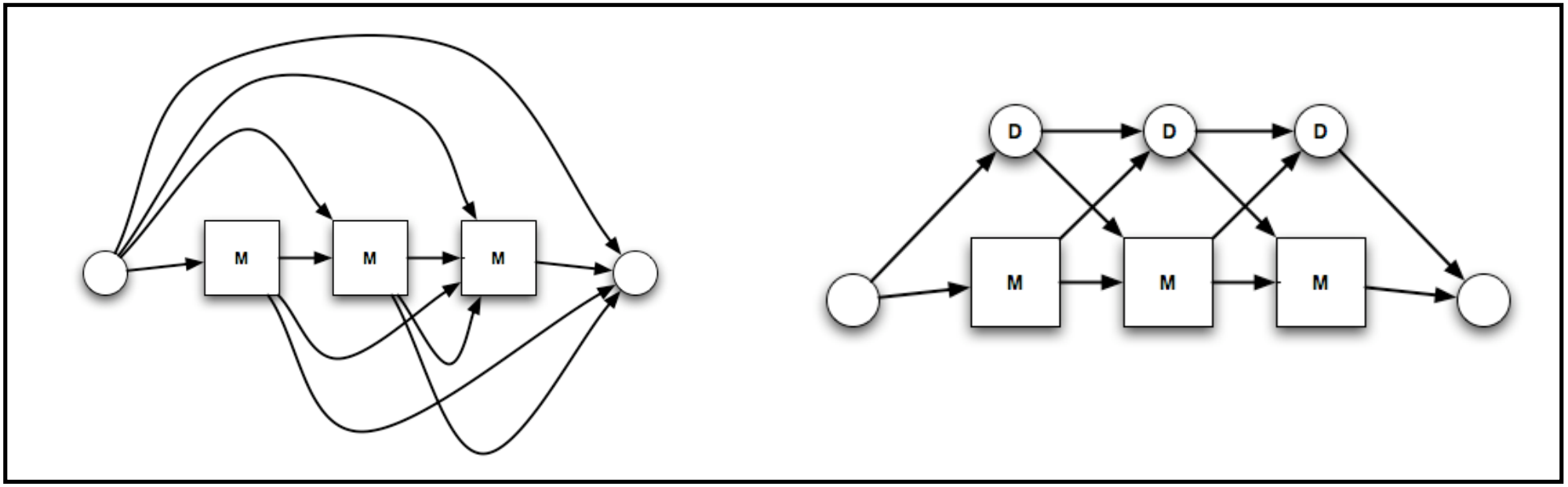
We can also calc p of gapped parts of alignments.

$$P(x_i \Diamond y_j) = \frac{f^M(i,j) b^M(i,j)}{P(x,y)}$$

Profile HMMs (i)

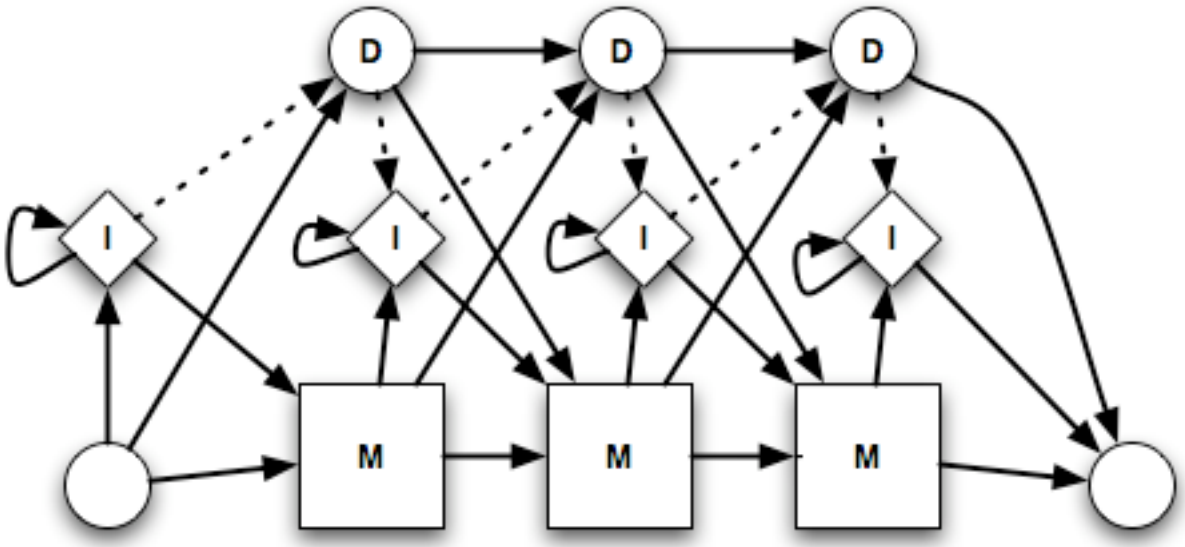


At each M node emission probabilities are given by profile (PSSM), now we model insertions (above) with I states and deletions with silent states (D).



Profile HMMs (i)

We can combine these into a model with all three states (M, D and I) into what is known as a profile HMM. Each position in a PSSM will have a corresponding M, I and D state.



Next week's reading

- Slight change of plans: Phylogeny, Tree Building is next (and we'll skip protein 3D structure)

- BSA Chapter 7 (next next week will be 8)

- Drummond AJ,Ho SYW,Phillips MJ,Rambaut A (2006) Relaxed Phylogenetics and Dating with Confidence. PLoS Biol 4(5): e88