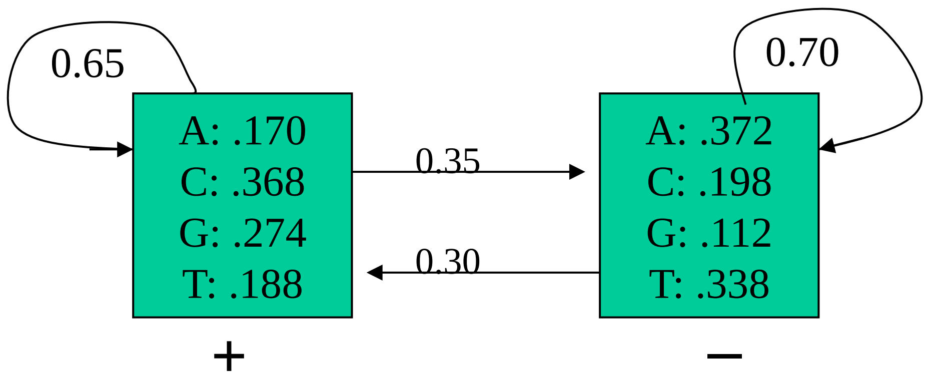
CISC 636 Computational Biology & Bioinformatics (Fall 2016)

Hidden Markov Models (II)

- The model likelihood: Forward algorithm, backward algorithm
- Posterior decoding



The probability that sequence x is emitted by a state path π is:

$$P(x, \pi) = \prod_{i=1 \text{ to } L} e_{\pi i}(x_i)\, a_{\pi i\, \pi i+1}$$

```
i:123456789
x:TGCGCGTAC
π:--++++---
```

$P(x, \pi) = 0.338 \times 0.70 \times 0.112 \times 0.30 \times 0.368 \times 0.65 \times 0.274 \times 0.65 \times 0.368 \times 0.65 \times 0.274 \times 0.35 \times 0.338 \times 0.70 \times 0.372 \times 0.70 \times 0.198.$

Then, the probability to observe sequence x in the model is

$$P(x) = \Sigma_\pi P(x, \pi),$$

which is also called the likelihood of the model.

How to calculate the probability to observe sequence x in the model?

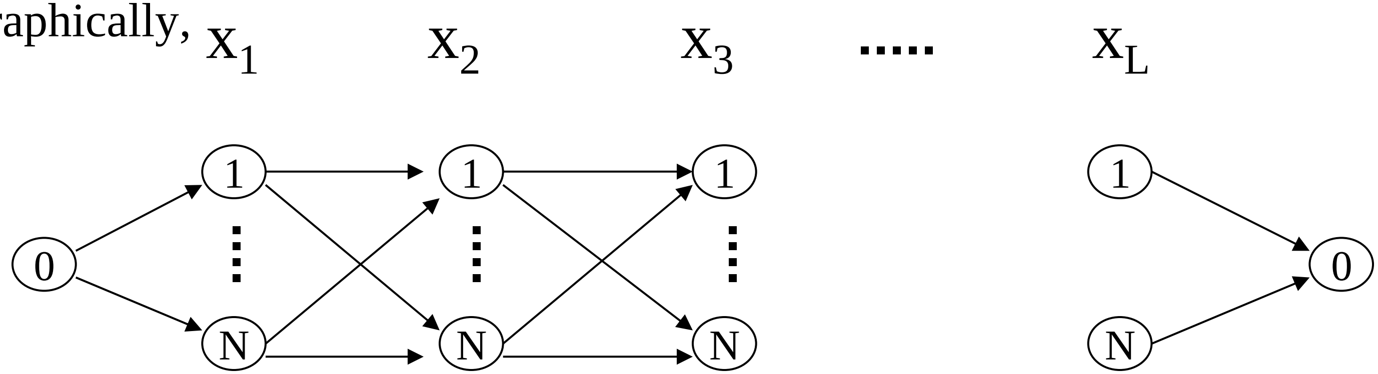
$$P(x) = \Sigma_\pi P(x, \pi)$$

Let $f_k(i)$ be the probability contributed by all paths from the beginning up to (and include) position i with the state at position i being k.

The the following recurrence is true:

$$\mathbf{f_k(i) = [\Sigma_j f_j(i\text{-}1)\, a_{jk}]\, e_k(x_i)}$$

Graphically,



Again, a silent state 0 is introduced for better presentation

Forward algorithm

Initialization: $f_0(0) = 1$, $f_k(0) = 0$ for $k > 0$.

Recursion: $f_k(i) = e_k(x_i) \Sigma_j f_j(i-1)\, a_{jk}$.

Termination: $P(x) = \Sigma_k f_k(L)\, a_{k0}$.

Time complexity: $O(N^2L)$, where N is the number of states and L is the sequence length.

Let $b_k(i)$ be the probability contributed by all paths that pass state k at position i.

$$b_k(i) = P(x_{i+1}, \ldots, x_L \mid \pi(i) = k)$$

Backward algorithm

Initialization: $b_k(L) = a_{k0}$ for all k.

Recursion (i = L-1, …, 1): $b_k(i) = \Sigma_j\, a_{kj}\, e_j(x_{i+1})\, b_j(i+1)$.

Termination: $P(x) = \Sigma_k\, a_{0k}\, e_k(x_1) b_k(1)$.

Time complexity: $O(N^2L)$, where N is the number of states and L is the sequence length.

Posterior decoding

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

Algorithm:

for i = 1 to L

do $\text{argmax}_k \, P(\pi_i = k \mid x)$

Notes: 1. Posterior decoding may be useful when there are multiple almost most probable paths, or when a function is defined on the states.

2. The state path identified by posterior decoding may not be most probable overall, or may not even be a viable path.