Hidden Markov Models (Part 1)

BMI/CS 576

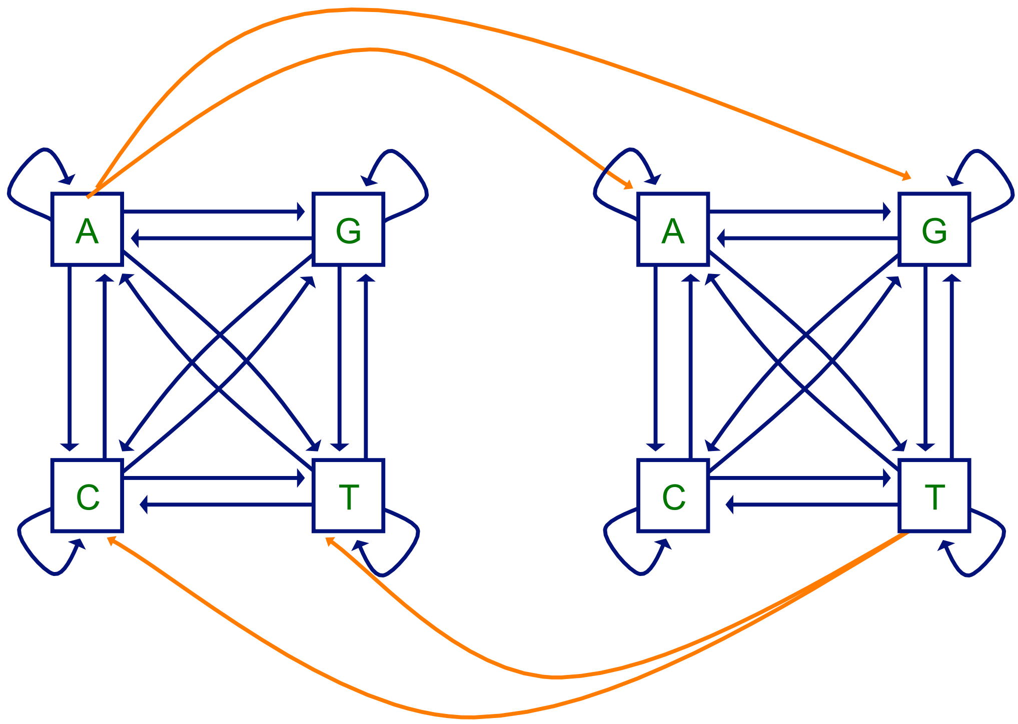
www.biostat.wisc.edu/bmi576.html

Mark Craven

craven@biostat.wisc.edu

Fall 2011

A simple HMM



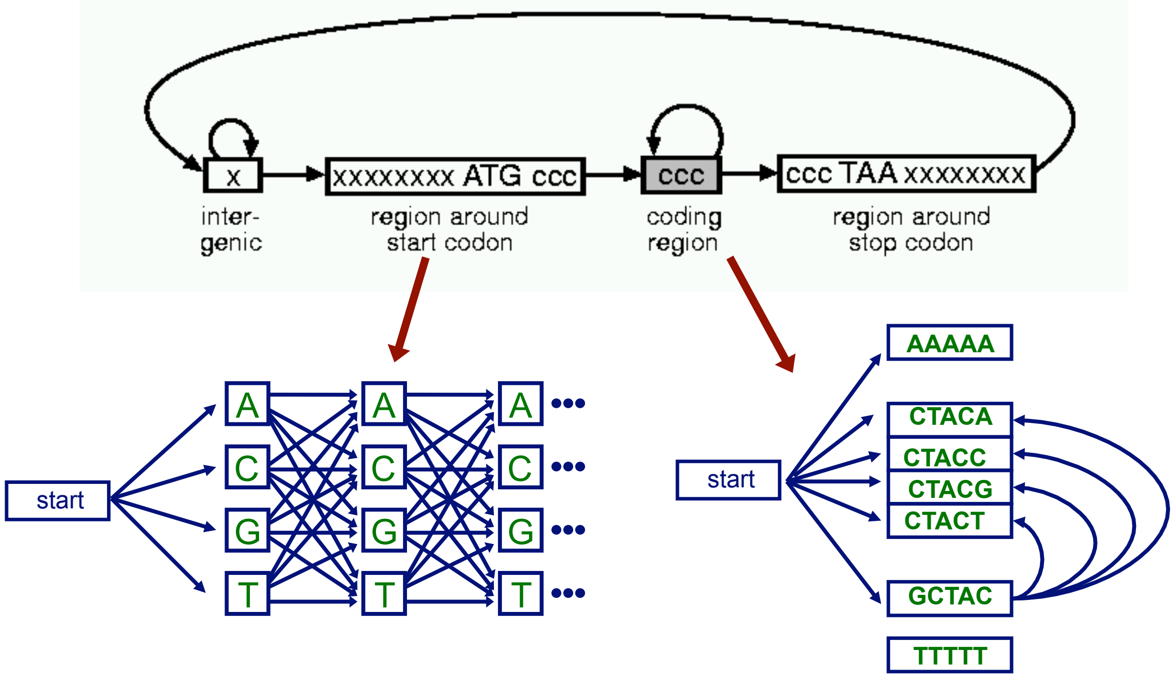
- given say a T in our input sequence, which state emitted it?

The hidden part of the problem

- we'll distinguish between the *observed* parts of a problem and the *hidden* parts
- in the Markov models we've considered previously, it is clear which state accounts for each part of the observed sequence
- in the model above, there are multiple states that could account for each part of the observed sequence – this is the hidden part of the problem

Simple HMM for gene finding

Figure from A. Krogh, An Introduction to Hidden Markov Models for Biological Sequences



The parameters of an HMM

- as in Markov chain models, we have transition probabilities

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

probability of a transition from state *k* to *l*

π represents a path (sequence of states) through the model

- since we've decoupled states and characters, we might also have emission probabilities

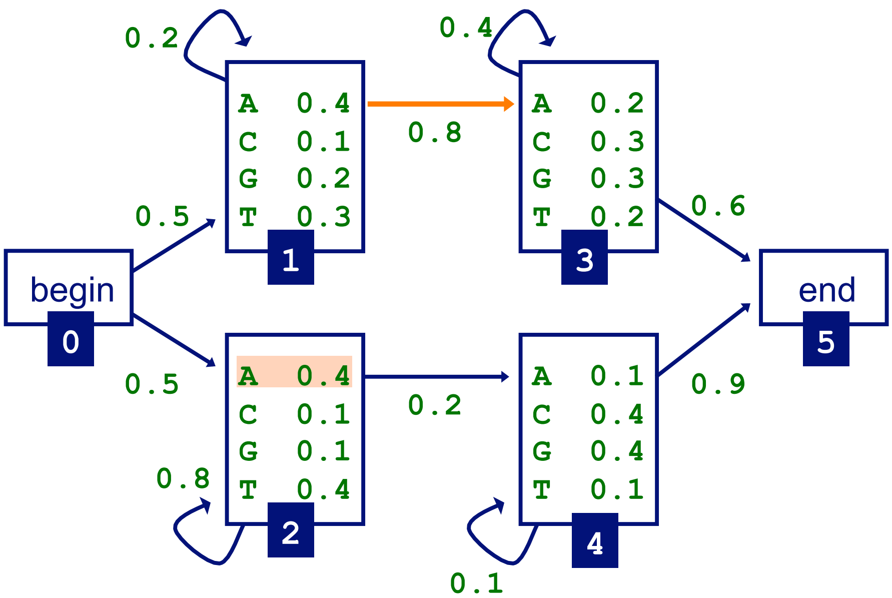
$$e_k(b) = P(x_i = b \mid \pi_i = k)$$

probability of emitting character b in state *k*

A simple HMM with emission parameters

a_{13} probability of a transition from state 1 to state 3

$e_2(\mathrm{A})$ probability of emitting character *A* in state 2

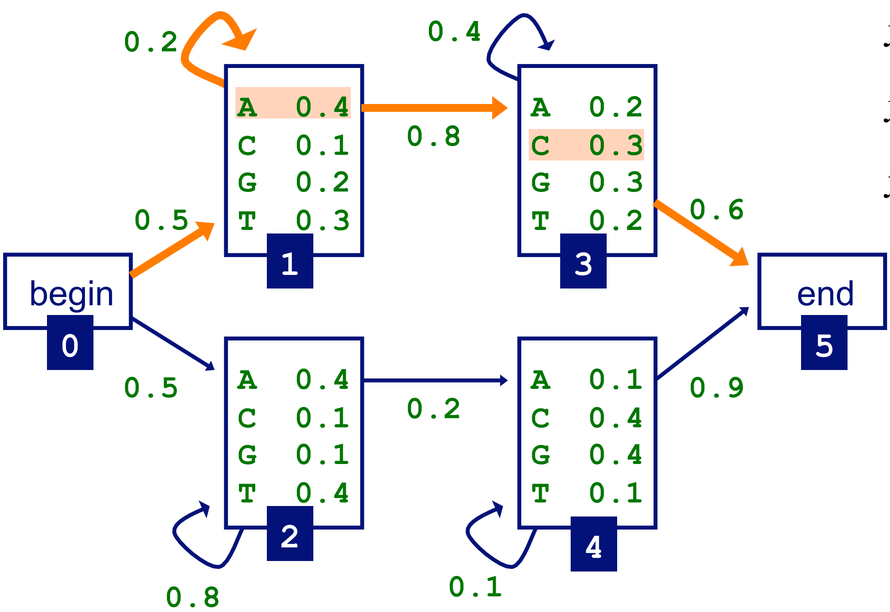


Three important questions

- How likely is a given sequence?
  the Forward algorithm
- What is the most probable "path" for generating a given sequence?
  the Viterbi algorithm
- How can we learn the HMM parameters given a set of sequences?
  the Forward-Backward (Baum-Welch) algorithm

Path notation

- let π be a vector representing a path through the HMM



$$\pi_0 = 0$$
$$x_1 = \text{A} \quad \pi_1 = 1$$
$$x_2 = \text{A} \quad \pi_2 = 1$$
$$x_3 = \text{C} \quad \pi_3 = 3$$
$$\pi_4 = 5$$

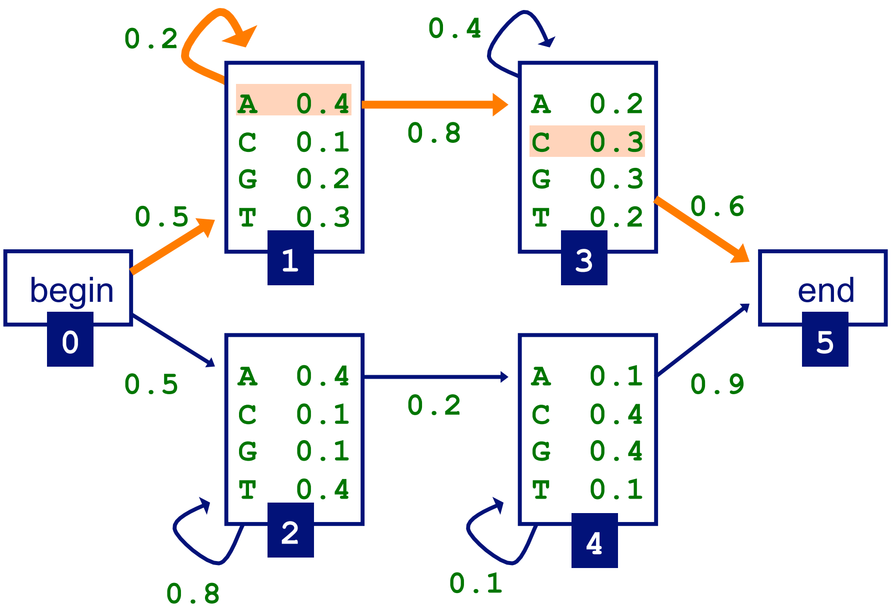
How likely is a given sequence?

- the probability that the path $\pi_0 \ldots \pi_N$ is taken and the sequence $x_1 \ldots x_L$ is generated:

$$P(x_1 \ldots x_L,\ \pi_0 \ldots \pi_N) = a_{0\pi_1} \prod_{i=1}^{L} e_{\pi_i}(x_i) a_{\pi_i \pi_{i+1}}$$

(assuming begin/end are the only silent states on path)

How likely is a given sequence?



$$P(\text{AAC},\pi\ ) = a_{01} \times e_1(\text{A}) \times a_{11} \times e_1(\text{A}) \times a_{13} \times e_3(\text{C}) \times a_{35}$$

$$= 0.5 \times 0.4 \times 0.2 \times 0.4 \times 0.8 \times 0.3 \times 0.6$$

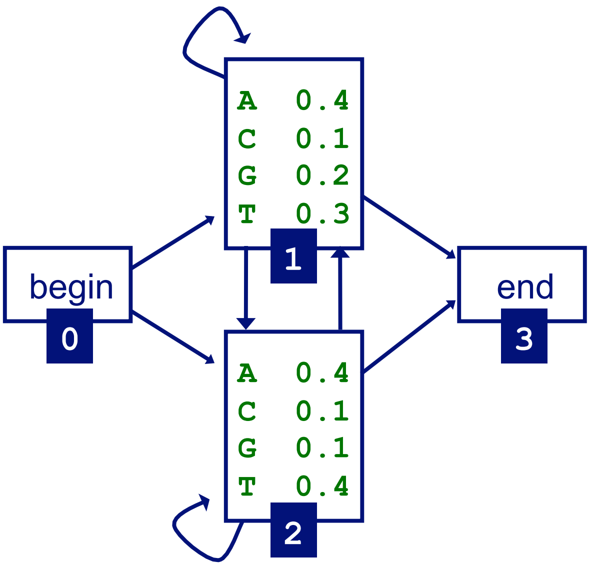
How likely is a given sequence?

- the probability over *all* paths is:

$$P(x_1 \ldots x_L) = \sum_{\pi} P(x_1 \ldots x_L, \underbrace{\pi_0 \ldots \pi_N}_{\pi})$$

Number of paths

- for a sequence of length L, how many possible paths through this HMM are there?



2^L

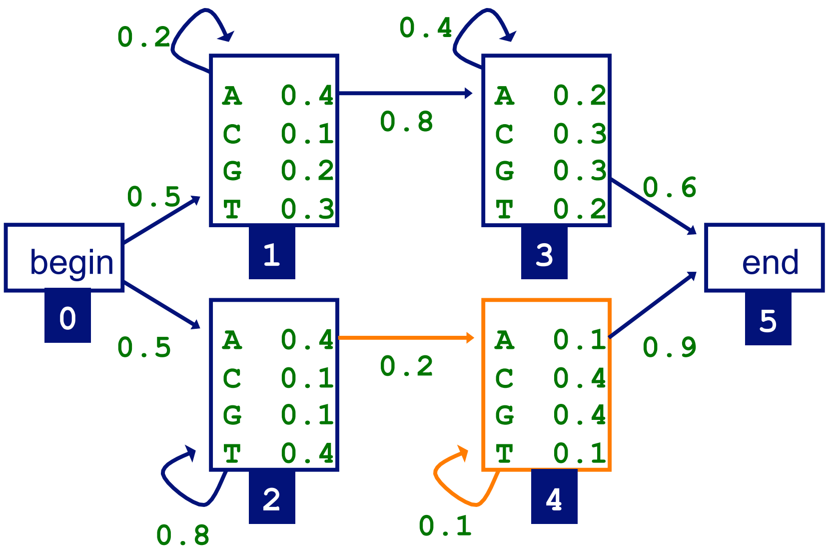
- the Forward algorithm enables us to compute $P(x_1 \ldots x_L)$ efficiently

How likely is a given sequence: the Forward algorithm

- define $f_k(i)$ to be the probability of being in state k having observed the first i characters of x
- we want to compute $f_N(L)$, the probability of being in the end state having observed all of x
- can define this recursively

The Forward algorithm

- because of the Markov property, don't have to explicitly enumerate every path – use dynamic programming instead



- e.g. compute $f_4(i)$ using $f_2(i-1)$, $f_4(i-1)$

The Forward algorithm

initialization:

$$f_0(0) = 1$$

probability that we're in start state and have observed 0 characters from the sequence

$$f_k(0) = 0, \quad \text{for } k \text{ that are not silent states}$$

The Forward algorithm

recursion for emitting states ($i = 1 \ldots L$):

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

recursion for silent states:

$$f_l(i) = \sum_k f_k(i) a_{kl}$$

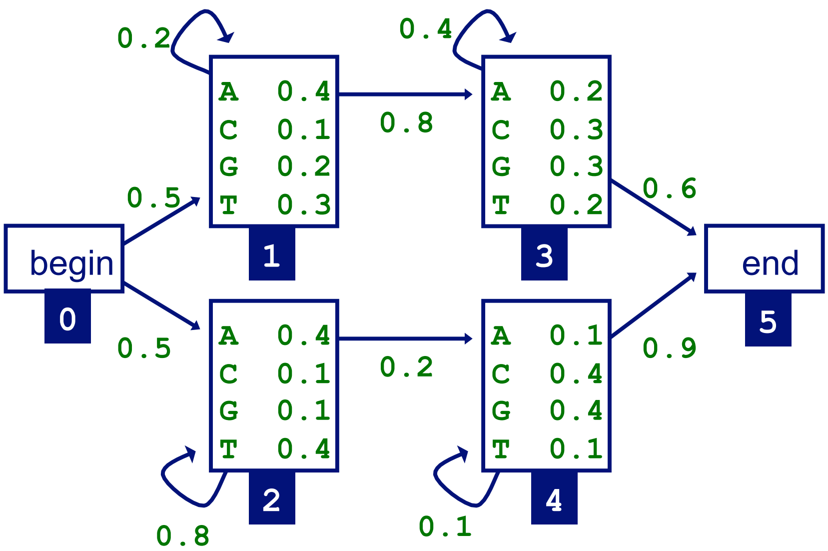
The Forward algorithm

termination:

$$P(x) = P(x_1 \ldots x_L) = f_N(L) = \sum_k f_k(L) a_{kN}$$

probability that we're in the end state and have observed the entire sequence

Forward algorithm example



- given the sequence x = TAGA

Forward algorithm example

- given the sequence x = TAGA
- initialization

$$f_0(0)=1 \qquad f_1(0)=0 \ \ldots \ f_5(0)=0$$

- computing other values

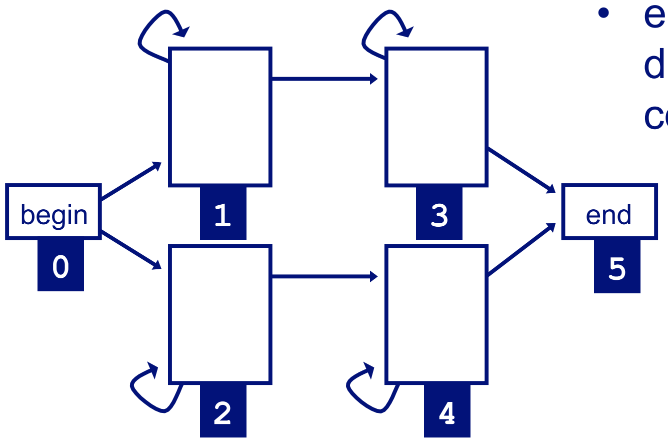
$$f_1(1)=e_1(T)\times(f_0(0)a_{01}+f_1(0)a_{11})=$$
$$0.3\times(1\times0.5+0\times0.2)=0.15$$
$$f_2(1)=0.4\times(1\times0.5+0\times0.8)$$
$$f_1(2)=e_1(A)\times(f_0(1)a_{01}+f_1(1)a_{11})=$$
$$0.4\times(0\times0.5+0.15\times0.2)$$

• • •

$$P(TAGA)=f_5(4)=(f_3(4)a_{35}+f_4(4)a_{45})$$

Forward algorithm note

- in some cases, we can make the algorithm more efficient by taking into account the minimum number of steps that must be taken to reach a state



- e.g. for this HMM, we don't need to initialize or compute the values

$$f_3(0),\ f_4(0),$$
$$f_5(0),\ f_5(1)$$

Three important questions

- How likely is a given sequence?
- What is the most probable "path" for generating a given sequence?
- How can we learn the HMM parameters given a set of sequences?

Finding the most probable path: the Viterbi algorithm

- define $v_k(i)$ to be the probability of the most probable path accounting for the first i characters of x and ending in state k
- we want to compute $v_N(L)$, the probability of the most probable path accounting for all of the sequence and ending in the end state
- can define recursively, use DP to find $v_N(L)$ efficiently

Finding the most probable path: the Viterbi algorithm

- initialization:

$$v_0(0) = 1$$

$$v_k(0) = 0, \quad \text{for } k \text{ that are not silent states}$$

The Viterbi algorithm

- recursion for emitting states (i =1…L):

$$v_l(i) = e_l(x_i) \max_k \left[v_k(i-1) a_{kl} \right]$$

$$\text{ptr}_l(i) = \arg\max_k \left[v_k(i-1) a_{kl} \right]$$

keep track of most probable path

- recursion for silent states:

$$v_l(i) = \max_k \left[v_k(i) a_{kl} \right]$$

$$\text{ptr}_l(i) = \arg\max_k \left[v_k(i) a_{kl} \right]$$

The Viterbi algorithm

- termination:

$$P(x,\pi) = \max_k \left(v_k(L)a_{kN}\right)$$

$$\pi_L = \arg\max_k \left(v_k(L)a_{kN}\right)$$

- traceback: follow pointers back starting at π_L

Three important questions

- How likely is a given sequence?
- What is the most probable "path" for generating a given sequence?
- How can we learn the HMM parameters given a set of sequences?