CS 2427 - Algorithms in Molecular Biology

Lecture #6a: 27 January 2006

Lecturer: Michael Brudno

Scribe Notes by: Graham Taylor

Hidden Markov Models (continued)

Today we will continue with Hidden Markov Models (HMMs) for finding CpG- islands contained in a genomic sequence. We will not, however, go into much detail. Those who do not have experience with HMMs are pointed towards the tutorial by Larry Rabiner (found on the course webpage).

From every single state in the HMM, there is a defined output distribution (of emmissions). There is also a set of transition probabilities, which indicate the probability of transitioning from one state to another. In the HMM for CpG-islands, every state emits a unique letter. This is not always the case (i.e the emission probabilities need not be zero-one).

Formally, an HMM is a 4-tuple, consisting of an emission alphabet, Σ, set of hidden states, Q, transition matrix, A, and a set of emission probabilities, e.

A path $\pi = (\pi_1, \pi_2, \ldots, \pi_L)$ in an HMM $M = (\Sigma, Q, A, e)$ is a sequence of states $\pi_i \in Q$. Given a sequence of symbols $x = (x_1, \ldots, x_L)$ and a path π through M. Then the joint probability is

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

where L is the length of the sequence. Usually we do not know the path π through the model, but often we are content to find the most probabable path. To solve the decoding problem, we want to determine the path π^* that maximizes the probability of having generated the sequence x of symbols, that is:

$$\pi^* = \arg\max_{\pi} \Pr(\pi|x) = \max_{\pi} \Pr(\pi, x).$$

For a sequence of N symbols there are $|Q|^N$ possible paths. Therefore, we cannot solve the problem by full enumeration. However, the Viterbi algorithm, based on dynamic programming, allows us to efficiently compute the most probabable path.

This algorithm involves a recursion and requires us to save “back pointers” to the states from which we came (starting from the final state). We always know from where we came, as we choose the state with max probabability.

The forward-backward algorithm is also a recursion, and is like the Viterbi algorithm, but we take sums instead of the maximum. Here we do not find the best path, but a sort of “average” over paths. The forward-backward algorithm is called by the Baum-Welch algorithm for training HMMs when the hidden states are unknown.

The total time complexity of the Viterbi algorithm is $O(Q^2T)$ and the forward-backward algorithm is $O(2Q^2T)$ (because of the two passes). The space complexity for both algorithms is $O(KT)$ since we must store sums (or pointers) for each state at each time step.

Do we need to explicitly account for sequencing/read errors when defining our HMMs for CpG-islands? No. The probability of sequencing errors is so small that we do not have to explicitly account for them in our model.

Parameters may be tied within the model to improve its performance as well as speed up learning/inference.