Introduction

In the last few lectures, we have focused on three problems related to local multiple sequence alignments:

- Discovery
- Modeling
- Recognition

We discussed Position Specific Scoring Matrices (PSSMs) for modeling and recognition of ungapped patterns or *motifs* and the Gibbs sampler for discovering such motifs in unlabeled data. PSSMs work well for fixed length patterns in which the sites are more or less independent. However, we saw that there are other kinds of motifs for which PSSMs are not well suited. In particular, PSSMs cannot

1. model positional dependencies,
2. recognize pattern instances containing insertions or deletions,
3. model variable length patterns,
4. detect boundaries.

Modeling motifs using Markov chains: pros and cons

Perhaps another formalism would work better for positional dependencies, variable length patterns, and boundary deletion. How about Markov chains?

Recall that a finite, discrete-time Markov chain is defined \(^1\) by

- a finite set of *states*, \(S_1, S_2, \ldots S_N\);
- a *transition probability*: \(a_{ij} = P(q_{t+1} = S_j | q_t = S_i)\), the probability that the chain will be in state \(S_j\) at time \(t + 1\) given that it was in state \(S_i\) at the previous time step, \(t\); and
- an *initial state probability distribution*, \(\pi_0 = (p(q_0 = S_1), p(q_0 = S_2), \ldots p(q_0 = S_N))\).

\(^1\)Note that the notation here differs somewhat from the notation introduced earlier in the semester.
A Markov chain can also be represented graphically: Each state is represented by a circle or box. State $S_i$ is connected to $S_j$ by an arrow if $a_{i,j} > 0$.

The connectivity or topology of a Markov chain can be designed to capture dependencies and variable length motifs. Suppose, for example, that our motif has a positional dependency like this one, in which we see either RD or QH in the last two positions, but never QD or RH.

A PSSM for this motif, however, would give equally good scores to WEIRD, which conforms to the positional dependency, and WEIRH, which does not. In contrast, we can construct a Markov chain to model this pairwise dependency like this:

The transition probabilities are chosen so that sequences with RD have a high probability in states $M_4$ and $M_5$ and sequences with QH will have high probability in states $M'_4$ and $M'_5$.

To address pattern instances with gaps and variable length motifs, we can construct a Markov chain to recognize query sequences with insertions, such as $O = WECIRD$, by adding an insertion state:
We can also handle query sequences that have deletions, e.g., $O = \text{WIRD}$, by modifying the topology of the Markov chain. One approach to capturing such deletions would be to add edges allowing us to jump over any set of states:

**Boundary detection:**

What about the fourth problem, *boundary detection*? We are given a sequence and we wish to label each symbol in the sequence according to its class (e.g. introns and exons; alpha helices and beta sheets; genes and non-coding regions). It is difficult to identify sharp boundaries between classes using by scoring windows using a fixed size model, such as a PSSM.

As an example, we considered the problem of recognizing transmembrane regions in a transmembrane protein. We are given an unlabeled amino acid sequence as input. The goal is to label each residue with the cellular location for which it is found.
Initially, we considered a simpler problem: Supposing we are given a sequence fragment that is either a transmembrane (TM) region or an extracellular/cytosolic region (E/C), can we determine which it is?

To do this we constructed two Markov models, a TM model and an E/C model. In these models, each amino acid is encoded as either a hydrophobic (H) residue, or a hydrophilic (L) residue.

**Transmembrane model:**

```
H  0.7  L
\  \  /  /
\  0.7 0.3 0.3 0.5
L  0.3  H
```

**Extracellular/cytosol model:**

```
H  0.5  L
\  \  /  /
\  0.5 0.5 0.5
L  0.5  H
```

**Parameter estimation:** What distinguishes the two models is the frequency of hydrophobic residues. Transmembrane regions tend to be enriched for hydrophobic residues; the TM model assigns higher probabilities to such sequences. I made up the transition probabilities in this simple example, but in practise the transition probabilities in such a model are determined from training data. Given labeled sequences (transmembrane or not transmembrane), we determine the transition probabilities of the two models. Let $A_{ij}$ be the number of transitions from state $i$ to $j$ in the training data$^2$:

$$a_{ij} = \frac{A_{ij}}{\sum_{j'} A_{ij'}}$$

$^2$Note that for some models, we may want to incorporate pseudocounts in the calculation of the parameters.
where $i$ and $j$ are either H or L. In this example, we must learn four transition probabilities: $a_{HH}$, $a_{HL}$, $a_{LH}$ and $a_{LL}$. Given a sequence, $\text{HHLLLHLL...}$ we count the number of $HH$ pairs to determine $A_{HH}$ and normalize by the number of pairs of the form $H^*$. The other transition probabilities are estimated similarly. We estimate the initial probability $\pi_H$ by counting the number of sequences that begin with a hydrophobic residue, and normalizing by the total number of sequences. Once we have learned the parameters, we can use the model to recognize new transmembrane sequences.

Evaluating a new sequence: Using this model, we can classify an observed sequence, $O = O_1, O_2, \ldots O_T$, by its log odds score

$$S(O) = \log \frac{P(O|TM)}{P(O|EC)}$$

where $P(O|TM) = \pi_{O_1} \prod_{i=2}^{T} a_{O_{i-1}O_i}$ is the probability of the observed sequence given the TM model. $P(O|EC)$ is defined similarly. If $S(O) \gg 1$, we infer that $O$ is located in the membrane.

For example, $P(\text{HHLHH}|TM) = 0.7 \times 0.7 \times 0.3 \times 0.7 \times 0.7 \approx 0.072$ and $P(\text{HHLHH}|EC) = 0.5 \times 0.5 \times 0.5 \times 0.5 \times 0.5 \approx 0.031$. The likelihood ratio is $\frac{0.072}{0.031} = 2.3$; in other words, it is a little more than twice as likely that $\text{HHLHH}$ is a transmembrane sequence than an E/C sequence. The log-odds score is $\log_2(2.3) \approx 1.2$.

Note that the likelihood of a sequence is sensitive to the length of the sequence; the longer the sequence, the lower its probability. Intuitively, this makes sense since the number of sequences of length $T$ grows exponentially with $T$. As a result, $P(\text{HHHHHHHH}|TM) = 0.04 < P(\text{HHLHH}|TM) = 0.07$, even though $\text{HHHHHHHH}$ looks more like a transmembrane segment than $\text{HHLHH}$. In contrast, the likelihood ratio is not sensitive to the length of $O$ because of the normalization. The log odds score for $\text{HHHHHHHH}$ is roughly 4.4, which is greater than 1.2, consistent with our expectations.

The above models are useful for classifying a sequence fragment where all residues are of the same class (i.e., all TM or all E/C). However, for finding boundaries in a sequence that has transitions from regions of one class to regions of another class, we would still need to score successive overlapping windows along the sequence, leading to a fuzzy boundary. For this question, the Markov chain has the same problem as the PSSM.

In order to determine the location of the transmembrane regions in an unlabeled sequence, a model that assigns a probability to a sequence is not sufficient; we need a model that explicitly labels each residue with its class (TM or E/C). For this purpose, we constructed a new model by adding transitions connecting the TM and E/C models.
A four-state transmembrane HMM:

This four-state model is much better suited to the boundary detection problem: The transitions between the \(M\) states and the \(E/C\) states indicate the boundaries between membrane regions and cytosolic or extracellular regions. However, this is no longer a standard Markov chain. In a Markov chain, every sequence of symbols corresponds to a unique sequence of states. That is not true of the four-state model, above. In the four-state model, there are two states associated with hydrophobic residues and two states associated with hydrophilic residues. As a result, for any given sequence of \(H\)’s and \(L\)’s, there are multiple paths through the states of the model and each of these state paths is associated with a different probability.
Hidden Markov Models

Hidden Markov models, or HMMs, are defined formally as follows:

1. N states $S_1..S_N$
2. An alphabet, $\Sigma = \{\sigma_1, \sigma_2 ... \sigma_M\}$
3. Transition probability matrix $a_{ij}$
4. Emission probabilities $e_i(\sigma)$ probability state $i$ emits character $\sigma \in \Sigma$
5. Initial distribution vector $\pi = (\pi_i)$

We refer to the transition probabilities, the emission probabilities and the initial distribution, collectively, as the parameters of the model, designated $\lambda = (a_{ij}, e_i(a), \pi_i)$. Following the notation used in Durbin, we will refer to the sequence of observed symbols as $O = O_1, O_2, O_3, ... O_T$ and the sequence of states visited as $Q = q_1, q_2, q_3, ... Q_T$ (the “state path”). To avoid confusing “sequences of symbols” with “sequences of states,” from now on we will use the term state path to designate a sequence of states. When considering more than one sequence or state path, we will use superscripts to distinguish them: $O^d = O^d_1, O^d_2, O^d_3, ...$ and $Q^b = q^b_1, q^b_2, q^b_3, ...$

HMMs differ from Markov chains in a number of ways. First, in an HMM, each state emits symbols from a fixed alphabet each time a state is visited. Emission probabilities are state-dependent, but not time-dependent. Note that an HMM is a generative model. It gives the probability of generating a particular sequence (hence, the emission probabilities.) This allows us to ask: Given an observed sequence, $O$, and an HMM model, what is the probability that $O$ was generated by this HMM?

Second, unlike Markov chains, in an HMM there is no longer a one-to-one correspondence between states and symbols in a Markov chain. For a given sequence, $O$, there is a unique state path corresponding to $O$, which determines the probability of $O$ with respect to the model. In contrast, in an HMM, there may be more than one, and often very many, state paths associated with $O$. Therefore, the “true” sequence of states that generated the observed sequence is unknown, or hidden, hence the term, “Hidden” Markov model. The sequences are hidden because it is not possible to tell the state merely by the output symbol. This hidden sequence of states corresponds to what we want to know, namely the classification of each symbol.

In the four-state model above, we used different states to distinguish between hydrophobic residues in the TM region and hydrophilic residues in the TM region. Another possibility is to use only one state for the transmembrane region and use the emission probabilities to distinguish between hydrophobic and hydrophilic residues. This approach is used in the following three-state HMM. Notice that this model is also a bit more sophisticated than previous models because it distinguishes between extracellular residues and cytosolic residues using separate $E$ and $C$ states.
A three-state transmembrane HMM:

The parameters for this model are

\[
\begin{array}{c|ccc}
  i & E & M & C \\
  \pi_i & 0.5 & 0 & 0.5 \\
  e_i(H) & 0.2 & 0.9 & 0.3 \\
  e_i(L) & 0.8 & 0.1 & 0.7 \\
\end{array}
\]

In this example, we assume that all transmembrane sequences are equally likely to start in the cytosol or in the extracellular matrix (ECM); i.e., \( \pi(C) = 0.5 \) and \( \pi(E) = 0.5 \).

Now that we have constructed an HMM model of transmembrane protein sequences, how do we use it to determine which amino acids are located in the membrane? In this HMM model, the subcellular location of each residue is represented by the state that emitted the symbol associated with that residue. There are many state paths that can generate a given sequence of amino acids. If we are given both the observed sequence and the state path, then calculating the probability is straightforward. Given a sequence \( O = O_1, O_2, \ldots, O_T \) and a state path \( Q = Q_1, Q_2, \ldots, Q_T \), the probability of visiting the states in \( Q \) and emitting \( O \) is

\[
P(O, Q|\lambda) = \pi_{q_1} \cdot e_{q_1}(O_1) \cdot a_{q_1q_2} e_{q_2}(O_2) \cdot a_{q_1q_2} \cdot e_{q_3}(O_3) \ldots a_{q_{T-1}q_T} e_{q_T}(O_T).
\]

For example, suppose \( O = \text{LHHHL} \) and \( Q = \text{CMMME} \), then

\[
P(\text{LHHHL}, \text{CMMME}|\lambda) = \pi_C \cdot e_C(L) \cdot a_{CM} \cdot e_M(H) \cdot a_{MM} \cdot e_M(H) \cdot a_{ME} \cdot e_E(L).
\]

However, the state path that actually generated a given protein sequence is unknown. In order to infer the location of the transmembrane regions, we must infer the “true” state path that generated the protein sequence. The boundaries between the transmembrane, extracellular and cytosolic regions are precisely defined by the transitions between \( C, M, \) and \( E \) states along this state path and the predicted subcellular location of each residue in the sequence is the state \( (C, M, \) or \( E) \) that emitted it.

Fig. 1 shows a cartoon of the probability distribution of sequence, state pairs for a hypothetical HMM. Every point on the horizontal plane corresponds to a particular sequence, \( O^d \), and a par-
Figure 1: The joint probability $P(O^d, Q^b)$ for every sequence $O^d$ and state path $Q^b$. The volume under this curve is one.
ticular state path, $Q^b$. The value on the vertical axis is the joint probability, $P(O^d, Q^b)$, that the HMM will visit the states on path $Q^b$ and emit sequence $O^d$. In the three-state TM model example, the set of all possible sequences, $O^1, O^2, O^3, \ldots$ corresponds to $H$, $L$, $HH$, $HL$, $LH$, $LL$, $HHH$, $\ldots$ and the set of all possible state paths, $Q^1, Q^2, Q^3, \ldots$ corresponds to $C$, $M$, $E$, $CC$, $CM$, $CE$, $MC$, $\ldots$. Note that $P(O^d, Q^b) = 0$ for many $(O^d, Q^b)$ pairs. For example, $P(O^d, Q^b) = 0$ when $O^d$ and $Q^b$ have different lengths. In our three-state model, $P(O^d, Q^b) = 0$ for any state path that contains $C$ adjacent to $E$, because $a_{CE} = 0$.

An HMM emits each sequence $O^d \in \Sigma^*$ with probability $P(O^d) \geq 0$. Since a sequence can, potentially, be emitted from more than one state path, to obtain the total probability of a sequence we must sum over the all possible paths:

$$P(O) = \sum_b P(O|Q^b) \cdot P(Q^b) = \sum_b P(O, Q^b).$$

Fig. 2 shows a cartoon representation of $P(O, Q)$ for a single sequence, $05$, for the set of all possible state paths, $Q$. The area under the curve is equal to $P(O)$ the total probability of sequence $O$. When all possible sequences and all possible paths are considered, the probability distribution shown in Fig. 1 sums to one:

$$\sum_d \sum_b P(O^d, Q^b) = 1.$$

Given a sequence, $O$, there are several questions we may wish to ask:

1. What is the state path that emitted sequence $O$? Otherwise stated, we wish to assign labels to an unlabeled sequence.
   
   Example: Identify the cytosolic, transmembrane, and extracellular regions in the sequence.
   
   In this case, we wish to assign the labels $E$, $M$, or $C$ to the unlabeled data.

2. What is the probability that a given sequence $O$, was generated by this HMM?
   
   Example: Is the sequence a transmembrane protein?

The process of inferring the correspondence between symbols and states is called “decoding”; we decode the sequence of symbols to determine the hidden sequence of states. HMMs were developed in the field of speech recognition, where recorded speech is “decoded” into words or phonemes to determine the meaning of the utterance. There are two common approaches to decoding. In Viterbi decoding, the most likely sequence of states is taken as an estimate of the state path that actually emitted the sequence. In posterior decoding, the sequence of most likely states is used. We will discuss these in the next lecture.
Figure 2: The probability of sequence $O = O5$ for every state path $Q^1, Q^2, Q^3 \ldots$. This curve corresponds to the intersection of the probability distribution in Fig. 1 and the vertical plane at $O = O5$ (shown as a blue line in Fig. 1). The area under this curve is $P(O5)$, the probability of $O5$. The maximum point on the curve is the most probable path, $Q^* = \arg\max_Q P(Q|O)$ is the highest point on this curve.