CS262: Computational Genomics **Winter 2016**

Lecture 8: January 28

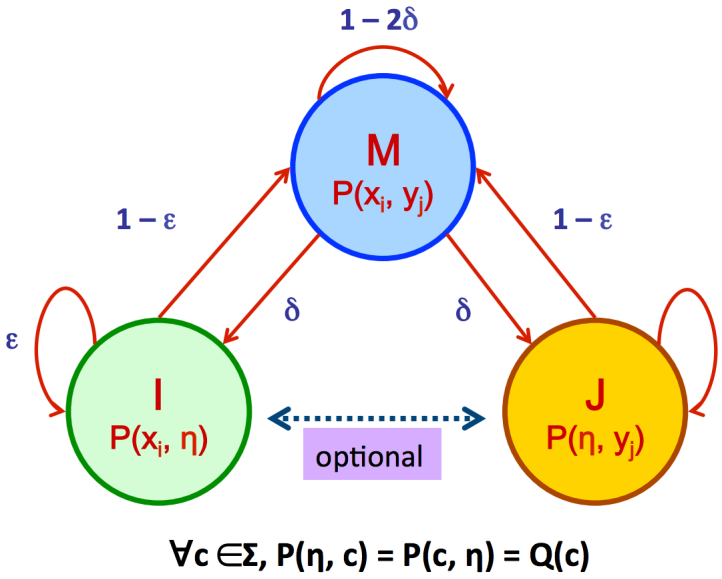
Lecturer: Serafim Batzoglou *Scribes: Sam Kim*

This lecture can be divided into three parts:

- We reviewed pair HMM's
- We discussed Conditional Random Fields (CRF's)
- We started discussing DNA sequencing

8.1 Review pair HMM's

Let us briefly review the pair HMM's that we discussed from last lecture. We wanted to apply some of the ideas and algorithms that we have learned from discussing the hidden Markov model to sequence alignments. In a regular HMM, each state emits a single sequence. However, in a sequence alignment, we observe two sequences and want to make sense of how they relate to each other. Therefore, instead of letting a single state emit a single sequence, we can allow it to emit a pairwise alignment such as (x_i, y_j), (x_i, η), and (η, y_j). We refer to such model as a pair HMM. We gave the following example and studied it last time.



We also discussed that we can adapt the dynamic programming methods of Viterbi, forward, and backward algorithms for single sequence HMM's to pair HMM's with small modifications. For instance, we can run

Viterbi with the following relations

$$V_M(i,j) = P(x_i, y_j) \max \begin{cases} (1-2\delta) \cdot V_M(i-1, j-1) \\ (1-\epsilon) \cdot V_I(i-1, j-1) \\ (1-\epsilon) \cdot V_J(i-1, j-1) \end{cases}$$

$$V_I(i,j) = Q(x_i) \max \begin{cases} \delta \cdot V_M(i-1, j) \\ \epsilon \cdot V_I(i-1, j) \end{cases}$$

$$V_J(i,j) = Q(y_j) \max \begin{cases} \delta \cdot V_M(i, j-1) \\ \epsilon \cdot V_J(i, j-1) \end{cases}$$

By taking the log of the likelihood ratios of getting the pair of sequence x_i, y_i individually with gaps as opposed to obtaining them as a match, and removing terms that are very close to zero, we get the Needleman-Wunsch algorithm.

$$\log V_M(i,j) = \log \frac{P(x_i, y_j)}{Q(x_i)Q(y_j)} + \max \begin{cases} \log V_M(i-1, j-1) \\ \log V_I(i-1, j-1) \\ \log V_J(i-1, j-1) \end{cases}$$

$$\log V_I(i,j) = \max \begin{cases} \log \delta + \log V_M(i-1, j) \\ \log \epsilon + \log V_I(i-1, j) \end{cases}$$

$$\log V_J(i,j) = \max \begin{cases} \log \delta + \log V_M(i, j-1) \\ \log \epsilon + \log V_J(i, j-1) \end{cases}$$

Pair HMM's can be designed for a variety of different specific situations. For instance, we can classify gaps into different groups such as long gaps or short gaps. We can also classify matches into different groups such as high or low conservation. In all these situations, no matter how complex the model becomes, the algorithms to compute the optimal alignment follows automatically from the three dynamic programming algorithms we discussed and therefore, HMM's can be quite useful.

8.2 Conditional Random Fields

The in-depth details of CRF's are outside the scope of this class, but it is good to have a general idea of what they are and how they are used since they are used quite extensively in practice. Conditional random fields (CRF) address some of the shortcomings of HMM's where the features that we have at our disposal are limited and can be quite clumsy to define.

For instance, let's say that we have a casino player with a fair and loaded die. Unlike before, where the player switches from the two dice with some probability, say that the player can observe the last few rolls and if he observes too many 6's, he changes the die in fear of being caught. Also, if the player observes that he is losing too much money, he switches to a loaded die. These type of situations are difficult to define with HMM's because the player sees not only the immediately previous state, but a number of previous states.

Like before, given an observed sequence x, we want to find the distribution of the sequence of states π. The CRF model gives this distribution as follows:

$$P(\pi|x) = \frac{\exp(\sum_{i=1,\ldots,|x|} w^T F(\pi_i, \pi_{i-1}, x, i))}{\sum_{\pi} \exp(\sum_{i=1,\ldots,|x|} w^T F(\pi'_i, \pi'_{i-1}, x, i))}$$

where F :(states, states, observations, index) $\mapsto \mathbb{R}^n$ is the *local feature mapping* and $w \in \mathbb{R}^n$ is the *parameter vector*. At an intuitive level, we can view the function F as indicating the set of events and the vector w as the set of weights for each of the possible events.

We can view CRF's as a generalization of HMM's. In fact, any HMM can be converted to a CRF. Note that in an HMM we can view the probability of observing a sequence x with states π is given as follows:

$$P(x, \pi) = a_{0,\pi_1} e_{\pi_1}(x_1) a_{\pi_1,\pi_2} e_{\pi_2}(x_2) \cdots a_{\pi_{n-1},\pi_n} e_{\pi_n}(x_n)$$

Taking the log on both sides, we present the product as a sum:

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

Let $\{1, ..., K\}$ be the set of states in an HMM and let $\{b_1, ..., b_M\}$ be the alphabet set. Then define the parameter vector $w \in \mathbb{R}^n$ as follows:

$$w = \begin{pmatrix} \log a_{0,1} \\ \vdots \\ \log a_{0,K} \\ \log a_{1,1} \\ \vdots \\ \log a_{n,n} \\ \log e_1(b_1) \\ \vdots \\ \log e_K(b_M) \end{pmatrix}$$

and the feature vector

$$F(\pi_i, \pi_{i-1}, x, i) = \begin{pmatrix} \mathbf{1}\{i = 1, \pi_{i-1} = 1\} \\ \vdots \\ \mathbf{1}\{i = 1, \pi_{i-1} = K\} \\ \mathbf{1}\{\pi_{i-1} = 1, \pi_{i-1} = 1\} \\ \vdots \\ \mathbf{1}\{\pi_{i-1} = K, \pi_{i-1} = K\} \\ \mathbf{1}\{x_i = 1, x_i = 1\} \\ \vdots \\ \mathbf{1}\{x_i = b_M, \pi_i = K\} \end{pmatrix}$$

for the indicator function $\mathbf{1} \mapsto \{0, 1\}$. Each row represented by the local feature mapping represents a particular event and each row of w gives the corresponding log probability of the event. Taking the sum of the inner products,

$$\exp\left(\sum_{i=1,...,|x|} w^T F(\pi_i, \pi_{i-1}, x, i)\right) = a_{0,\pi_1} e_{\pi_1}(x_1) a_{\pi_1,\pi_2} e_{\pi_2}(x_2) \cdots a_{\pi_{n-1},\pi_n} e_{\pi_n}(x_n) = P(x, \pi)$$

Therefore, we can check that the conditional probability gives the correct distribution for HMM's as before

$$\frac{\exp(\sum_{i=1,...,|x|} w^T F(\pi_i, \pi_{i-1}, x, i))}{\sum_{\pi} \exp(\sum_{i=1,...,|x|} w^T F(\pi'_i, \pi'_{i-1}, x, i))} = \frac{P(x, \pi)}{\sum_{\pi} P(x, \pi)} = P(\pi|x)$$

We note that in the above feature mapping, all the events require only the observed sequence x_i. However, the feature mappings of general CRF's can rely on the whole observed sequence $x_1 \cdots x_n$. In this way, CRF's can be seen as a generalization of HMM's, which can be quite powerful. When x represents the sequence of throws of a die and π represents the sequence of fair or loaded die, then a valid event in the feature mapping can be the event that the last fifty throws resulted in at least thirty 6's, which is not possible with HMM"s.

The basic questions that we are interested in for CRF's are the evaluation, decoding, and learning like in the case of HMM's:

- **Evaluation**: Given a sequence of observations x and a sequence of states π, compute $P(\pi|x)$

- **Decoding**: Given a sequence of observations x, compute the maximum probability sequence of states

$$\pi^\star = \arg\max_\pi P(\pi|x)$$

- **Learning**: Given a CRF with unspecified parameters w, compute the parameters that maximizes the likelihood of π given x,

$$w^\star = \arg\max_w P(\pi|x, w)$$

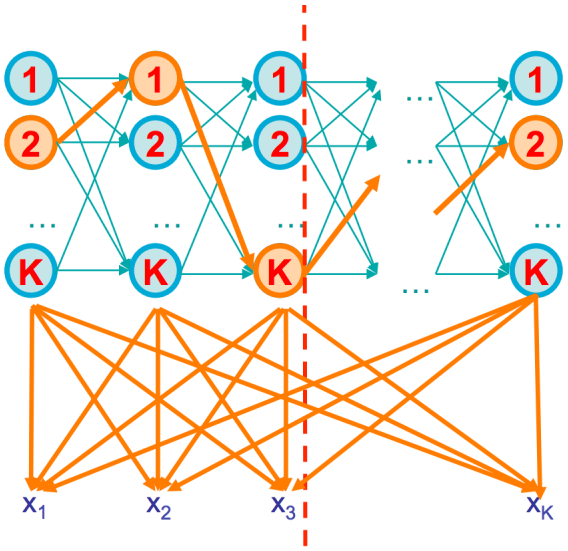
All of these problems can be computed in a similar manner as in the case of HMM's by formalizing the Viterbi, forward, and backward algorithms for CRF's. Take decoding as an example. We want to compute

$$\begin{aligned}
\arg\max_\pi P(\pi|x) &= \arg\max_\pi \frac{\exp(\sum_{i=1,\ldots,|x|} w^T F(\pi_i, \pi_{i-1}, x, i))}{\sum_\pi \exp(\sum_{i=1,\ldots,|x|} w^T F(\pi'_i, \pi'_{i-1}, x, i))} \\
&= \arg\max_\pi \exp\left(\sum_{i=1,\ldots,|x|} w^T F(\pi_i, \pi_{i-1}, x, i)\right) \\
&= \arg\max_\pi \sum_{i=1,\ldots,|x|} w^T F(\pi_i, \pi_{i-1}, x, i)
\end{aligned}$$

The second equality follows from the fact that the normalizing factor is fixed. The last equality follows from the fact that exp is monotonic.

In this case, we can use Viterbi with the relation as follows

$$V_k(i) = \max_j [w^T F(k, j, x, i) + V_j(i-1)]$$



Here, note that there are no probabilities and we are maximizing potential values. However, the intuition is the same as in the original Viterbi algorithm in that at each state k, we maximize the score to the left and to the right. We note that even though the features depend on x, x is fixed and parse to the left of step i, given we end in state k, does not affect parse to the right of step i.

For learning, we observe that the function $-\log P(\pi, x, w)$ is a convex function of w. Therefore, there is a global solution unlike the case of HMM's where there are local solutions. Since the function is also differentiable, we can use the gradient descent/ascent method and other standard techniques in convex optimization.

To run gradient descent, for instance, we can first compute the partial derivatives with respect to each parameter w_j

$$\frac{\partial}{\partial w_j} \log P(\pi|x, w) = F_j(x, \pi) - E_{\pi' \sim P(\pi'|x,w)}[F_j(x, \pi')]$$

Here $F_j(x, \pi)$ is the correct value for the jth feature given the correct parameters and $E_{\pi' \sim P(\pi'|x,w)}$ is the expected value of the jth feature given the current parameters.

If the correct feature values are greater than the predicted values calculated by the parameters, this derivative is positive and we can increase w_j. Otherwise, if the correct feature values are less than the predicted values, then we can decrease w_j. This process will move the probability mass from the incorrect parses to the correct parses.

8.3 DNA sequencing

We finished the lecture with a brief introduction to DNA sequencing.

DNA sequencing is the process to determine the order of the nucleotides in an individual's DNA. Knowledge of DNA sequences has become an important tool for basic biological research and in numerous applied fields such as medical diagnosis, biotechnology, forensic biology, virology, and biological systematics. There is no one simple machine(yet) that takes a DNA in one end and gives perfect sequence of nucleotides in the other end. It is often a series of complex steps with different method variations.

In the 80's, it was not uncommon for a PhD student to spend 5 years in sequencing parts of a DNA sequence. Starting from the mid 80's there was a movement for a public effort to sequence the human DNA. Back then, this was considered a crazy idea by many with the debate as to whether the entire genome should be sequence or just the gene regions.

In 1990's there was a race between the public and private sector to sequence the entire human DNA. In the public sector, the Human Genome Project(HGP) got underway as an international collaborative research project in 1990. On the other hand, Celera Genomics launched its private research project to sequence the human genome in 1998. Both projects were declared complete by 2001. HGP used the DNA of 12 anonymous regional individuals, while Celera Genomics sequenced the DNA of the CEO Dr. Craig Venter.

Does it matter which human genome we sequence? On average, about 1 out of 1000 letters of two individuals' genomes will be different. This is due to the fact that for most of human history, the human population has been very small (about 10,000 individuals). This small population interbred with individuals with very similar genomes which reduced the genetic variation. Therefore, any two individual's DNA sequence is more than 99.9% similar, which means that any sequence is fairly representative of the human genome and therefore, we can talk about *the* human genome.

We will start next lecture discussing different sequencing technologies.