Conditional Random Fields, DNA Sequencing

Armin Pourshafeie

February 10, 2015

CRF Continued

HMMs represent a distribution for an observed sequence x and a parse π $P(x, \pi)$. However, usually we are interested in the parse that the sequence we have implies ($P(\pi \mid x)$). We can model this probability using CRF's

Definition

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

where

F : (state, state, observations, index)$\in \mathbb{R}^n$ is called the local feature mapping. Note that this form is more general than HMM's. In particular, since we are looking at entirety of x we can model events that we couldn't using HMM's.

$w \in \mathbb{R}^n$ is the parameter vector (a set of weights)

Relationship to HMM's

From the definition, we can notice that CRF's can model a greater set of systems. they can look at the entire set of emissions x. In particular any HMM can mapped to a CRF.

For a HMM we have

$$\begin{aligned} \log P(x, \pi) &= \log P(\pi_0) + \sum_{i=1}^{|x|} \log P(\pi_i|\pi_{i-1}) + \log P(x_i \mid \pi_i) \\ &= \log a_{0\pi_0} + \sum_{i=1}^{|x|} \log a(\pi_{i-1}, \pi_i) + \log e_{\pi i}(x_i) \end{aligned}$$

For a CRF, we can define weights

$$\mathbf{w} = \begin{pmatrix} \log a_0(1) \\ \vdots \\ \log a_0(K) \\ \log a_1(1) \\ \vdots \\ \log a_K(K) \\ \log e_1(b_1) \\ \vdots \\ \log e_K(b_M) \end{pmatrix} \in \mathbb{R}^n$$

and feature vector

$$F(\pi_i, \pi_{i-1}, x, i) = \begin{pmatrix} 1\{i = 1, \pi_{i-1} = 1\} \\ \vdots \\ 1\{i = 1, \pi_{i-1} = K\} \\ 1\{\pi_{i-1} = 1, \pi_{i-1} = 1\} \\ \vdots \\ 1\{\pi_{i-1} = K, \pi_i = K\} \\ 1\{x_i = b_1, \pi_i = 1\} \\ \vdots \\ 1\{x_i = b_M, \pi_i = K\} \end{pmatrix} \in \mathbb{R}^n$$

Where $1\{x\}$ is the indicator function for x.
This gives log probability

$$\log P(x, \pi) = \sum_{i=1}^{|x|} w^T F(\pi_i, \pi_{i-1}, x, i)$$

From which we can get the conditional probability

$$P(\pi \mid x) = \frac{P(x, \pi)}{\sum_\pi P(x, \pi)} = \frac{\exp\left(\sum_{i=1}^{|x|} w^T F(\pi_i, \pi_{i-1}, x, i)\right)}{\sum_\pi \exp\left(\sum_{i=1}^{|x|} w^T F(\pi_i, \pi_{i-1}, x, i)\right)}$$

This shows:

- $HMM \subsetneq CRF's$
  - HMM parameter vectors have restrictions that are not present in CRF. for instance in an HMM $\sum_i a_{0i} = 1$ (since this refers to probability of transition.) but in CRF parameter vectors only need to be positive.
  - CRFs can have features that HMM's cannot have.

    - For an HMM the feature vector is $F(\pi_i, \pi_{i-1}, x_i, i)$
    - HMM's can only generate each observation once whereas CRF's can have features dependent on x_{i-1}, x_{i+1}
    - We could in principle define a very large set of features but this will require a very large training data set.
  - The down side is that CRF's don't model the observations.
    - i.e. They are not generative models.
  - CRF needs categorized data for learning.

Basic questions for CRF's

Just like HMM's, there are 3 important questions we would like to answer.

- Evaluation: Given a sequence of observations (x), and a sequence of states (π), what is $P(\pi \mid x)$?
- Decoding: Given a sequence of observations (x) what is the maximum probability sequence of states $\pi = \arg\max_\pi P(\pi \mid x)$
- Learning: Given a CRF, compute the weights that maximize the likelihood of π given x. $w_{ML} = \arg\max_w P(\pi \mid x, w)$

Decoding

Decoding can be done with the Viterbi for CRF's

Using $P(\pi \mid x)$ given earlier, we can write

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

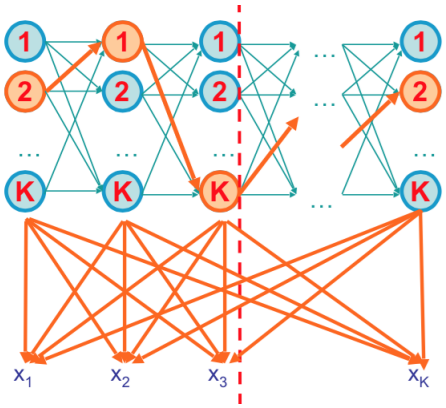
Where we have used monotonicity of exp to get from the second to the third line.

Now, we can easily extend the Viterbi algorithm using

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

This is very similar to the Viterbi for HMM's however, there is no emission and transmission probabilities, and now we have a dot product of the weights with the feature vector.

Figure 1:



Note that the features in the CRF can depend on any position in x but DP depends only on the previous state.

We could extend the forward/backward algorithm to compute the partition function.

We can intuitively justify this algorithm. Even though the features could depend on any x, given that at time i we end up in state k, we can separately maximize the scores to the left and to the right. Since x is fixed, the parse to the left and to the right are do not effect each other except for the fact that they have to end/start at k.

Learning

Note that $-\log P(\pi \mid x, w)$ is both differentiable and convex. Therefore we can use gradient descent and be sure to find the global solution. (Contrast this with Baum-Welch which gives the local solution)

In order to run gradient descent (ascent) we first compute the partial derivatives w.r.t w_j:

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

Where $E_{\pi' \sim p(\pi' \mid x, w)}[F_j(x, \pi')]$ is the expected value of jth feature given the current parameters and

$F_j(x, \pi)$ is the value of the jth feature given the correct parameters.

Notice that the if the true value of the jth feature is higher than the mean calculated by our parameters this derivative is positive, which will increase w_j and if the case is reversed then the derivative is negative which will decrease w_j upon iteration.

Figure 2:

DNA Sequencing

Background

DNA sequencing allows us to determine the order of nucleotides in an individuals DNA. The DNA is consisted of about 3 billion base pairs.

The Human Genome project started in 1990 and took more than 10 years (2001 for the first draft) and over $3 billion dollars to complete.

During the 1990's two separate programs attempted to sequence the human genome. The public effort (Human Genome Project) collected DNA from 12 anonymous regional individuals. The private effort by Celera Genomics attempted to sequence the DNA of the CEO Dr. Craig Venter. At the end of the day, any two individual's DNA sequence is more than 99.9 % similar so either sequence could be fairly representative of the Human population.

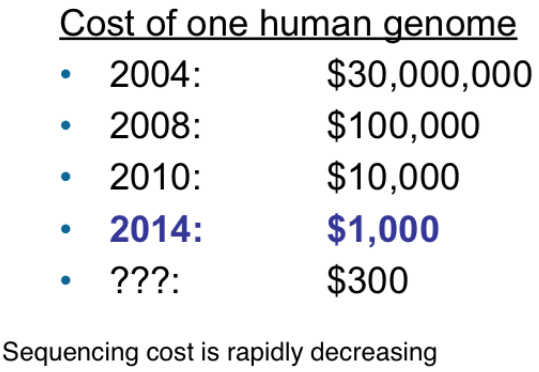
Why is the Human genome so similar? On average, only 1 out of 1000 letters of two individuals is different. In contrast, in Ciona savignyi, two individuals may have up to 10% difference in their DNA. This low polymorphism rate is related to the population size. For most of human history, human population has been very small (˜10,000 individuals living in Africa). At a certain choke point, the human population reduced to ˜1000. This small population interbred with individuals with very similar genomes which reduced the genetic variation. With the expansion of human population in recent history, we are acquiring some rare mutations. In particular, we can compute the heterozygosity of a population using $h = \frac{4N\mu}{1+4N\mu}$ where N is the population size ($\sim 10^4$ for most of human history) and μ mutation rate (10^{-8}).

Polymorphism comes in many forms. The most common is Single Nucleotide Polymorphism (**SNP**) which is when a single base pair is different among two individuals. Another kind of polymorphism is copy number variation (**CNV**) which is referred to the case where a large region of DNA is duplicated more or

less than usual.

Cost

While the Human Genome Project cost 3 Billion dollars for a single sequence, the cost has been dropping sharply. With about $1000 per sequence in 2014. This is good news as it opens up the possibility of sequencing different species, different people or even different tissues of the same person.



Sequencing

Old technology, Sanger vectors

This method was developed by the biochemist Frederick Sanger, who won two Nobel prizes for his work in DNA sequencing. This method was the dominant technology for ~25 years after it's invention in 1977 and has more recently been replaced by "Next-Generation" sequencing.

In this method:

1. We break the DNA into fragments of different sizes by sonication, (each fragment is a few thousands base pairs).

2. Each small fragment then is copied many times. In order to copy the fragments, each piece is inserted into a known region of a circular, self-replicating piece of genome known as **cloning vector** (bacterium, plasmid)**.** After insertion, the cloning vector is replicated many times using cellular DNA-replication machinery.

Since the location of the DNA fragment is known, we can cot it from the cloning vector.

3. Next we determine the ordering of nucleotides within the copied pieces. In order to do this the DNA is placed in a solution so it can replicate however, the solution will have small concentration of base pairs similar to ACGT which are

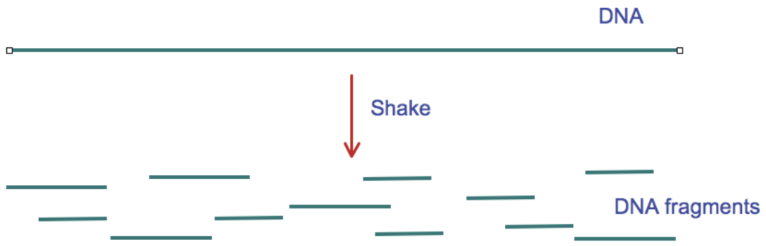


Figure 3:

fluorescent and terminate the copying process. This produces many substrings of the original DNA each starting from the beginning but ending at different points with a fluorescent base-pair. These subcopies are then organized based on length using electrophoreses and the base pairs can be read using the fluorescent base-pairs at the end.

Illumina

In this method,

1. The DNA is fragmented and tagged with a different adaptor at each end.
2. Each fragment is then amplified in a flow cell. The flow cell contains receptors that are complementary to the tags on the DNA. The DNA fragment will attach to these tags. Then a complement of the fragment is produced and the original fragment is washed away. The complementary fragment then clones itself through bridge amplification. In this method, the fragment bridges so the other end tag is connected to the other type of receptor. Then the fragment is cloned and the two clones are separated so that one version of the clone is on each receptor. This process is then repeated many times. At the end, the reverse strands are washed off.
3. Finally, the DNA is sequenced using **Sequencing by Synthesis.** In this method, the complement of the DNA is made one base at a time. Each base has a fluorescent tag and can be read after it is attached.

This entire sequencing process is done in a massively parallel way.

Assembly

After sequencing, the reads need to be assembled into a DNA sequence. There are two types of assembly problems

1. De Novo Assembly
2. Resequencing

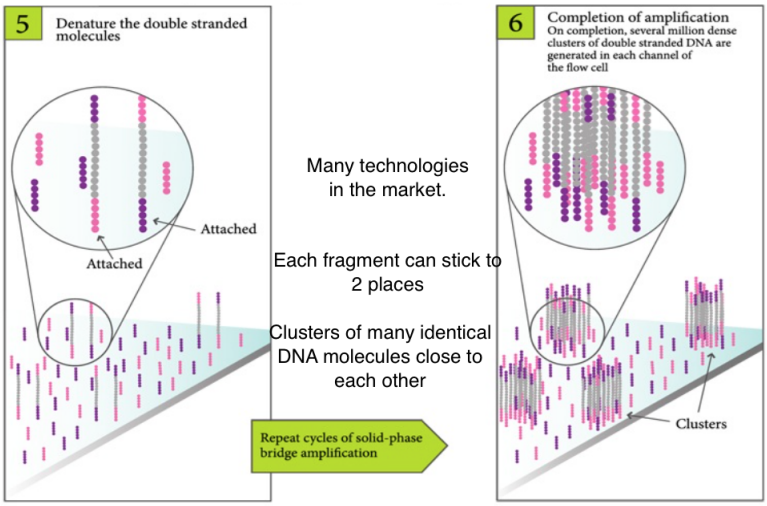


Figure 4:

Slide Credit: Arend Sidow

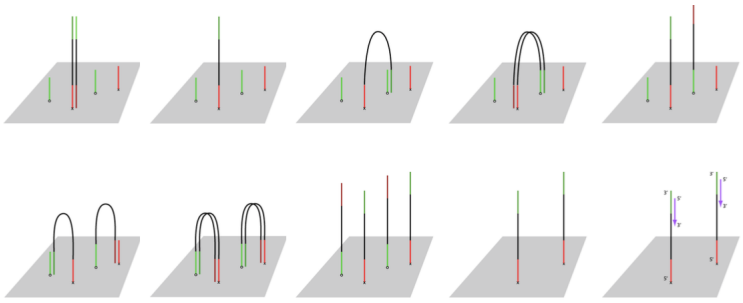


Figure 5:

Slide Credit: Arend Sidow

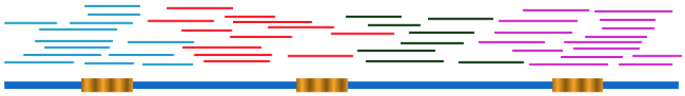


Figure 6: Clustering

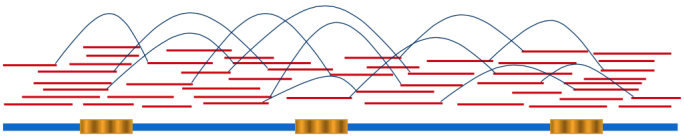


Figure 7: Linking

De Novo Assembly

De Novo Assembly refers to the problem of assembling from scratch. For example, assembling the first human DNA.

In this method, we try to cover each region with high redundancy, then we will overlap and extend each read. Here, **coverage** is defined as the expected number of reads per region and is given by: $C = NL/G$

Where N: is the number of reads, L is the length of reads and G is the length of the genome.

Repeat problem

One major problem with the sequence reconstruction is the problem of repeats. While repeats are not common in lower organisms (~5% for bacteria) in mammals they are very common (~50%) repeats are dangerous because they could cause two faraway parts of the DNA to concatenate. Of course, repeats that are shorter than the length of a read, are not a problem.

There are two usual methods to deal with the repeat problem:

1. Clustering the reads: Clustering the reads to different regions will reduce the chance of having long repeats in each cluster. (Figure 6)

2. Linking the reads (Figure 7)