Seth Cooper
CSE 527 notes
October 24, 2007

# Approaches to Finding Sequence Motifs

## - DNA Binding Site Summary

Binding sites are not perfect, they can tolerate variability. One helix turn is 10 base pairs, so about 6 to 8 base pairs are bound to by the protein.

There are two meanings to "motif". One is a "structural" motif, which is the structure of a protein, such as helix-turn-helix. A few examples are illustrated in the slides. The other is a "sequence" motif, which is the sequence which is bound to, such as "GGCTA". The motifs discussed here are sequence motifs.

There are several ways to determine a binding motif. One is gel electrophoresis. In this method, DNA is placed at one end of a gel, and an electric field is applied. Because DNA is negatively charged, it moved toward the positive side, but it is slowed by the gel. Small molecules will move further along in the gel, and this information can be used to determine what has bound. E.g., if there is a shift in position of a band on a gel between one lane containing just DNA vs another containing DNA + protein, a likely explaination is that the protein binds to that DNA sequence, therefore slowing its migration in the gel. Another method is immuno-pulldown such as ChIP-chip. In this method, formaldehyde causes covalent linkage between protein and DNA. Antibodies will then attach to the proteins to capture them. It is then possible to look at the DNA that has been removed to determine the binding sequence.

## - TATA Box Frequencies

The "TATA Box" is one studied binding site motif of length 6. To determine the frequencies you must first know the correct alignments of many motifs. This gives information about the variability of the binding sites.

## - TATA Scores

In order to score a given sequence, we need to go from frequency to score. To do this, we use the log likelihood over background. This gives us a Weight Matrix Model we can use to score.

- Scanning for TATA

To scan a long sequence for a TATA Box, we slide along the sequence scoring each sequence of 6 base pairs with the WMM. Higher scores are more likely to be TATA.  Scores have been shown to correlate with binding affinity.

- Score Distribution

There will be some distribution of scores for TATA as well as for the background.  Some sequences will clearly be one or the other but some cases will still be ambiguous.

- Pseudocounts

If we didn't see any of a particular base at a particular position in our data, then it will have a score of negative infinity in our WMM.  This means it is not possible to occur in the motif. However, it may be possible, but we didn't see any in the data. We may not want one unseen data point to rule out an otherwise strong match.  We can add a small pseudocount to each observed count to get around this problem.

- WMM Summary

WMMs can be used to locate motifs in sequences.  Higher order models may capture effects of neighboring base pairs, but won't be as good at capturing effects of distant base pairs, and need more training data.

- How-to Questions / Motif Discovery

Discovering a motif given a set of sequences which contain the motif is difficult, but there are several approaches.  We want to find the subsequences in each sequence with the highest relative entropy.  We know the motif length we are looking for, usually around 10.  Finding an exact solution is not tractable.

- Brute Force

This approach examines all possible alignments of subsequences in the given sequences.  The possibilities are explored in a tree structure and all nodes are expanded.  This will exhaustively try all possible candidates for motif.  This method is guaranteed to find the best motif; however, it is very slow.

- Greedy Best-First Approach

In this method, we examine the possibilities by building a tree structure as in the greed approach. However, now we will only expand the best few nodes in each step. This improves the speed of the algorithm. However, it is not guaranteed to find the best motif. The danger is throwing out possible later better matches early on.

- Expectation Maximization

There is another method based on Expectation Maximization called MEME. This method uses EM and considers the motif start locations as the hidden variables and the WMM as the parameters. If we know where the motif starts in each sequence, we can compute the WMM. Similarly, if we know the WMM, we can find where the motif starts.

- Expectation Step

In this step we find the motif start probabilities. To do this, we scan the WMM across the sequences and get scores. Higher scores are more likely starting positions.

- Maximization Step

In this step we find the WMM. To do this, we do a count of letter frequencies in the sequences, weighted by the probability that the motif starts at that position.