Patterns of Binding Targets

John Letey David Knox

Abstract

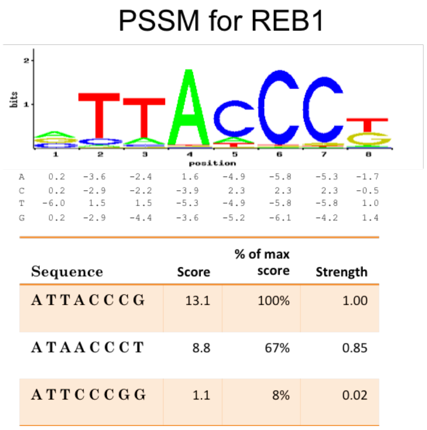
Research project investigating if weak affinity binding sites aid in unwrapping the nucleosome so that strong affinity sites can occur, and if there's a certain pattern between the different binding sites if this occurs. Using Python computer programs to simulate the binding process. We inputted pre-mapped yeast chromosomes and yeast transcription factors into our model.

1 Introduction

DNA binding factors interact differently given different sequences of DNA. If an individual factor could only recognize a single unique sequence, then we could represent the interaction with two parameters: the on-rate and off-rate. This works well for free-floating, well-mixed environments where the molecules have a limited number of interactions depending on the molecule counts and volume of the environment. However, the positively charged transcription factors (TFs) have a natural attraction to the negatively charged DNA, which leads to non-specific binding along the DNA. The non-specific binding may only be transient, but sequences that partially match the intended sequence will have stronger affinity for the factor than for random DNA sequences. High affinity binding sites will more likely be bind at low molecule counts, while low affinity sites will require high molecule counts before the site will be highly bound. The high affinity sites also have longer residency times than the transient binding of the low affinity sites.

Position Specific Scoring Matrices (PSSMs) have long been used to calculate the affinity of an individual TF for different sequences of DNA. A PSSM captures the affinity of a DNA binding factor for a sequence, position by position. The simplest method for determining a PSSM is to collect a set of sequence fragments that have been bound by a factor. These sequences are usually large fragments (20 - 100 nucleotides) of which the factor generally only recognizes a few nucleotides (4 - 20), therefore a common sub-sequence within the collection would be expected. The most common and most likely sequence would be the consensus motif that most biologists report as the binding sequence. However, the factor usually only has a few mandatory positions within the sequence and allows multiple different nucleotides to occupy the intervening positions. Aligning the pattern across multiple sequences, the number of times each nucleotide is found at each position can be calculated. Using these counts, we can calculate the log-odds of seeing a particular nucleotide at a particular position within a bound sequence of DNA. Storing the values for each nucleotide for each position results in a matrix. We can set the values of the matrix to be probabilities of each nucleotide at each position, or we can assign a score for each nucleotide at each position. The scores could be positive for a matching nucleotide or negative for an unlikely nucleotide, resulting in a PSSM. The PSSM contains the consensus motif as the highest scoring sequence, but also allows any sequence to be scored. Some alternative sequences to the consensus motif will still be able to score highly, indicating that even at low concentrations of the binding factor, sites with the alternate DNA sequence would also be bound.

PSSMs provide a powerful tool for predicting the likely binding sites for any DNA sequence. We can iteratively pass sub-sequences to the PSSM and use the high scores to predict not only the binding sites, but also the sites of high TF residency times. Consider the REB1 TF as an example (Figure 1). The TF has particular sequences to which it preferentially binds, described by its PSSM. This matrix specifies the number of positions, as well as the nucleotides that are preferentially bound at each position. For any given sequence, we can calculate a score by adding up all the values for the sequence's nucleotide at each position. The resulting score indicates how well the given sequence matches the best sequence for the factor.



A	0.2	-3.6	-2.4	1.6	-4.9	-5.8	-5.3	-1.7
C	0.2	-2.9	-2.2	-3.9	2.3	2.3	2.3	-0.5
T	-6.0	1.5	1.5	-5.3	-4.9	-5.8	-5.8	1.0
G	0.2	-2.9	-4.4	-3.6	-5.2	-6.1	-4.2	1.4

Sequence	Score	% of max score	Strength
ATTACCCG	13.1	100%	1.00
ATAACCCT	8.8	67%	0.85
ATTCCCGG	1.1	8%	0.02

Figure 1: **TF affinity can be scored for any sequence by using the TFs PSSM.** An example PSSM, here for REB1, describes the probability of binding at each nucleotide for any possible sequence. Positions are assumed to be independent and therefore the probability of a TF binding to any arbitrary sequence can be easily calculated.

2 Methods

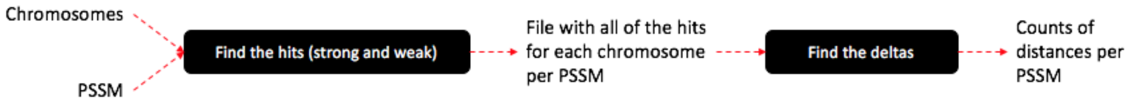

Chromosomes
PSSM
Find the hits (strong and weak)
File with all of the hits for each chromosome per PSSM
Find the deltas
Counts of distances per PSSM

Figure 2: **The Approximate Pipeline our Code Follows**

Looking at the pipeline (Figure 2): we started with a bunch of chromosomes from yeast, 18 to be exact, in FASTA form, however, we did discarded 2 of the chromosomes because they didn't give us any useful information. We then took that. Also, we had 124 transcription factors, with their corresponding PSSMs, in TAMO form. Using this data, we can then take each transcription factor's PSSM and then take each chromosome and calculate all the affinities along the chromosome for that particular transcription factor. Every transcription factor floats around on the chromosome until it finds a match, then after a while, it separates and moves on. Note that when the transcription factor gets to the end of the chromosome, it turns around and goes over the reverse strand (where there could possibly be a match), which is called the "reverse complement" chromosome. Hence, we also run over the reverse complement of each chromosome so that we get a better simulation of what the transcription factor is actually doing. We then take all of these hits and classify them as either *strong* (having a very high affinity chance of the transcription factor binding at that location in the chromosome), *weak* (having a low chance, but still a chance, of binding), or just nothing (when there's close to nothing of a chance of binding). We did output all of these hits (with the corresponding label that we gave them) to GFF (Genetic File Format) files. Please note that the *strong*, *weak*, and nothing classifications have cutoffs. We chose the cutoffs to be 0.75 or greater is *strong*, 0.35 and greater until 0.75 is *weak*, and anything below 0.35 is classified as nothing.

We wrote another Python program that takes in all of those GFF files, and sees if there are weak spots within some distance (we chose 150) around each strong spot. The program also plots a histogram of these to show us the frequency of weak binding sites around the strong binding sites.

3 Results

Unfortunately, since we are dealing with such a big data set, we only ran my code on 30 different TFs, over all 16 chromosomes (there were 18 in the data, but we discarded two of them). Here is an example of a histogram:

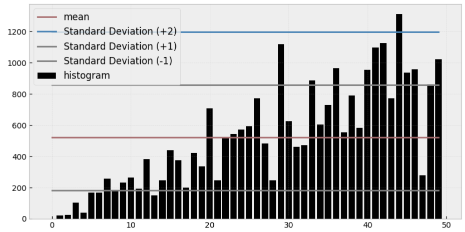


Figure 3: **Histogram for TF GCR1**

Please note that this is only a small selection of output histograms that we get. But, the good news is that this is exactly what we were looking for! As you can see from Figure 3, there is an overall peak right at 7.5.

4 Code

All the code for this project is hosted on GitHub, located here. Note that we originally started out in MATLAB, but then migrated to Python for efficiency reasons.

5 Future Work

As you may have noticed above, we didn't run our code for more than one dataset and also only for one specific settings for parameters. So, in the future we would love to write an overall program to fine tune the parameters that our program uses and test these parameters on different datasets, and see what our program outputs!

6 Mentions

[1] A special thanks to David Knox for the terrific guidance he gave me on the project!

[2] Thanks to Matt Shirley for making pyfaidx! It really helped in reading the FASTA data files quickly and efficiently.