A Study of Transcription and Its Affects

John Letey **David Knox**

Abstract

Research project investigating hits calculated from yeast chromosomes and PSSMs and their correlations to each other. Transcription clusters and PSSMs have information about the transcription process, so modeling it can help us understand more about transcription. We hypothesized that the histogram of hits contains a pattern of having a peak around 30. We were able to prove this [INSERT ACCURACY HERE]% of the time.

1 Introduction

The method of transcription, as depicted in Figure **??**,

DNA
RNA
Protein

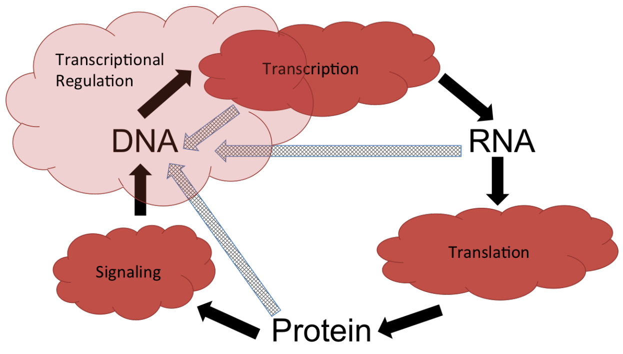


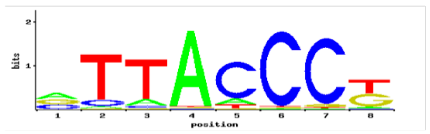
Figure 1: **The Central Dogma of Biology as seen from viewpoint of Biologist and Computational Biologist.**
a) The central dogma of Biology embodies the notion that a DNA template is used to create the RNA copies that are used to create the proteins used in biological processes. b) In Computational Biology, the focus is changed to the processes, where DNA, RNA, and proteins are the inputs and outputs of the processes. Transcription is the process that takes the current DNA state as input and produces an RNA as the output. The translation process converts the RNA input and outputs a set of amino acids that form the protein represented by the RNA. Many of the proteins do not immediately interact with the processes, but are sequestered away from the processing environment, waiting until the cell detects an internal or external signal that causes the proteins to return to the processing environment and change its behavior. Current biology textbooks usually show the dogma as a linear process, but we show it as a circular set of processes that feedback upon each other. The processes are shown as clouds and the solid black arrows show the inputs and outputs of the processes. The gray hatched arrows depict the influence on the DNA state by the proteins, the RNA products, and even the transcription process itself.

2 Methods

PSSMs (Position Specific Scoring Matrices), have been used to calculate the affinity of an individual transcription factor for different sequences of DNA for a long time. 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 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 bound 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.

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) and the factor generally only recognizes a few nucleotides (4 - 20), therefore a common sub-sequence within the collection would be expected. The most common or 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 the 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 transcription factor residency times. Consider the REB1 transcription factor as an example (Figure 1). The transcription factor 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.

PSSM for REB1



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
A T T A C C C G	13.1	100%	1.00
A T A A C C C T	8.8	67%	0.85
A T T C C C G G	1.1	8%	0.02

Figure 2: **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. (Knox, Page 24, Figure 1.5)

Assuming the above, we can then calculate each one of the scores for the entire DNA sequence given each transcription factor. When we do this, we can classify each score as being a strong hit, if it's greater than our set strong threshold, a weak hit, if it's greater than our set weak threshold but less than our set strong threshold, and else, we don't classify it at all.

3 Results

4 Code

All the code for this project is open sourced on GitHub here.

5 References

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

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