A Study of Transcription and Its Affects

John M. Letey

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

2 Methods

We started off with multiple yeast chromosomes in FASTA () format and some PSSMs (Position Specific Scoring Matrix), which correspond to the transcription factors embedded in the yeast chromosomes. Since a PSSM is a probability matrix, we made a function that takes in a string the same "length" as the PSSM and outputs the probability of that the transcription factor corresponding to that PSSM will bind to that string in the chromosome. If that probability is greater than our strong threshold (which we set to 0.7), we classify that string as a strong hit. If the probability is less than our strong threshold, but greater than our weak threshold (which we set to 0.35), it's a weak hit. Else, we don't classify it as anything at all. Since in a chromosome, one string can have a really low probability one way, but it could have a very how probability on the reverse compliment of the chromosome. We account for this by simply calculating all the hits for both ways. When we're done, we output all the hits to a GFF () formated file.

Now that we've calculated all of the hits, we can now analyze them!

3 Results

4 Code

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

[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.