Context specific sequence preference of DNA-binding proteins

Tara Friedrich with help from advisor
Katherine Pollard
University of California, San Francisco
Gladstone Institute of Cardiovascular Disease

Abstract

Deciphering why different sequences are preferred under different contexts is critical to understanding how nearby genes are regulated so precisely. Here I investigate this question using two perspectives. First, I ask if there are sequence features that are predictive of different contexts. Using a logistic regression model to identify sequence preferences when Met4 is recruited to sequences with Cbf1 and Met28 (three factor model) compared to when just Cbf1 is present (two factor model), I confirmed a previous observation that the presence of an 'AAT' motif that is predictive of the three factor model. Using this 'AAT' motif to label three factor binding sites, I tried to identify other features that were predictive of this motif within their genomic context (such as distance to other proteins binding DNA, distance to nearest gene, etc). *Here I look at different methods to predict sequence preferences between two contexts with the broad goal of understanding how cofactors affect the binding preferences of a protein. I go further by identifying other features that are predictive of this three-factor model within its genomic context.*

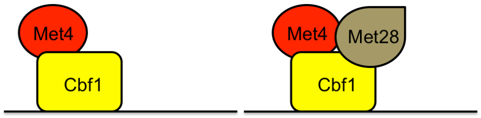
1. Introduction

Transcription factors, or DNA binding proteins, regulate genes by binding regulatory elements to turn genes on or off. Understanding why a transcription factor gets recruited to a regulatory sequence can help elucidate how genes are regulated so precisely. Delving deeper into how the genome regulates this process in a spatiotemporal fashion can shed insights into key moments in development as well as how cellular processes become dysfunctional.

To understand this problem, I have chosen a simple, well-studied example. Met4 is a protein that activates transcription of genes involved in the sulfuric metabolic network in *Saccharomyces cerevisiae*. Met4 lacks the ability to bind DNA directly. It is recruited to DNA through cofactors (Cbf1, and Met31/Met32) that target its binding to the correct locations on DNA. A third cofactor (Met28) stabilizes this complex of proteins (1).

My first goal is to investigate the sequence specificity of Met4 recruitment in the presence of two or three cofactors.

Figure 1a and 1b: Here are illustrations of two factor and three factor models of Cbf1 binding DNA. Cbf1 binds DNA and recruits Met4 and Met28. Met28 acts to stabilize the complex's interactions.



Two factor Three factor

Understanding the sequence specificity of Met4 recruitment using experimental assays can add to our knowledge of the sequence binding preferences of this factor within its genomic context. The presence of other factors binding nearby and the distance to the nearest gene could help predict the functional importance of a Cbf1 binding site. This led me to my second goal where I attempted to

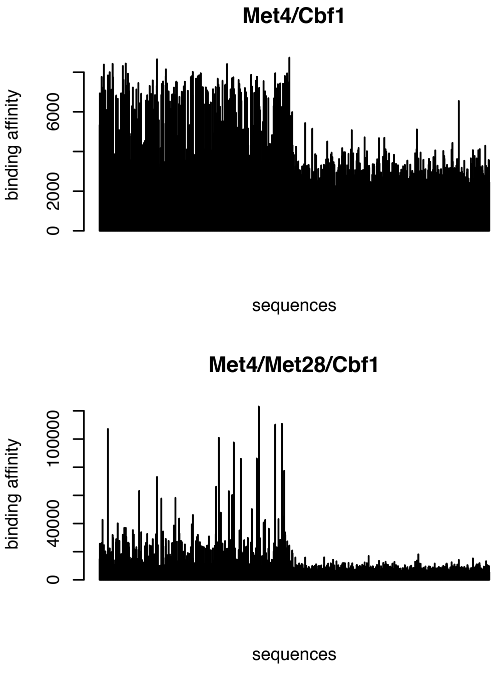
2. Data Acquisition

i. Two factor or three factor model

To answer this first question, I found published experimental data that measures the recruitment of Met4 to 1358 DNA sequences (20 bps long) that either bind Cbf1 or Met31 (2). More specifically, the data quantifies DNA binding affinity of Met4 using Protein Binding Microarray assays performed in the presence or absence of Met31, Cbf1, and Met28.

By testing the binding affinity of sequences for various combinations of cofactors, we can compare sequence preferences for different complexes. While Met4 binds non-specifically to all sites, it shows higher sequence specificity when recruited specifically by Met31/Met32. Cbf1 can recruit Met4 to specific sequences but, when stabilized by Met 28 (Fig 2a), its sequence preferences become even more selective (Fig 2b).

Figures 2a and 2b: Raw data of binding affinities (median probe fluorescence) for each of the 1358 sequences of length 20 under the **a)** two factor model and the **b)** three factor model context.



ii. Genomic Context

To identify Cbf1 binding sites within their genomic context, I identified the promoters, or 1000bp regulatory regions, of 45 genes that have been shown to be highly regulated by Met4 (1). I scanned these promoters for Cbf1 binding sites and looked for the 'AAT' motif upstream of these Cbf1 sites. I then scanned the promoters for a second factor, Met31.

3. Methods

The primary goal of this project was to distinguish sequence preferences for a transcription factor between two contexts. Once sequence preferences were found, certain motifs could be used to label Cbf1 binding sites in the genome and other genomic features could be used to predict these labels.

3.1 Features

i. Two factor or three factor model

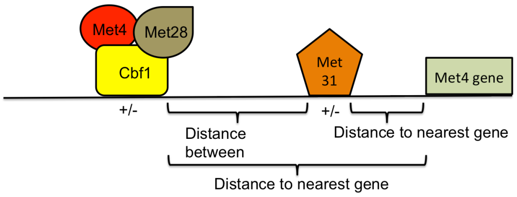
The 20 bp sequences (n=1358) are converted into a binary feature vector. If the ratio of the binding affinities between the two contexts exceeds a threshold of four, it is labeled as "three factor". Each nucleotide (A/C/G/T) for each position (1-20) found within a sequence will be considered and each feature will essentially mark the presence of absence of a nucleotide at that position for each sequence.

ii. Genomic context

Using the data mentioned above, I labeled each Cbf1 binding site within a promoter as having (positives) or not having (negatives) the 'AAT' motif upstream. From the data I processed, I curated four features (Fig 6)

- Distance between Cbf1 and Met31
- Distance between Cbf1 and gene
- Distance between Met31 and gene
- Orientation of Met31 (forward or reverse strand)

Figure 6: Cbf1 sites are labeled as either containing an 'AAT' motif 3 bp upstream (positives) or not (negatives). Here I illustrate features used to predict presence or absence of 'AAT' motif near the Cbf1 binding site.



3.2 Models and metrics

To answer my first question discerning Cbf1's sequence preferences between the three or two factor model, I asked which nucleotides in which positions found within a sequence are most predictive of sequences favoring the three factor model using a logistic regression model with an L2 penalty. For both questions, I compared the performance between logistic regression and SVM linear kernel. I used 10X CV to test the robustness of my model's accuracy and AUC. I implemented these models using python's scikit-learn package.

4. Results

Using my features, I tested the performance of different models by measuring the AUC (Area Under the Curve). I tested the robustness of my model using 10X CV. Holding the sample size constant, I compared the performance of each model and noted that logistic regression performed better than SVM (Table 1).

Table1: Benchmarking the performance of different models

	Sample size	Training AUC	Testing /CV AUC
Logistic Regression (LR)	504	0.93	0.90
Linear kernel SVM	504	0.88	0.88

Interestingly, training and cross-validation mean scores appear to converge more quickly in the SVM than in the logistic regression model (Fig 3a, 3b). The SVM model mean CV score is unaffected by increasing training example siz (Fig 3b).

Figure 3a: Learning curves for the logistic regression model. Here we are plotting the average training and cross-validation score with increasing training sample size. The colors indicate indicate the standard deviation from the mean.

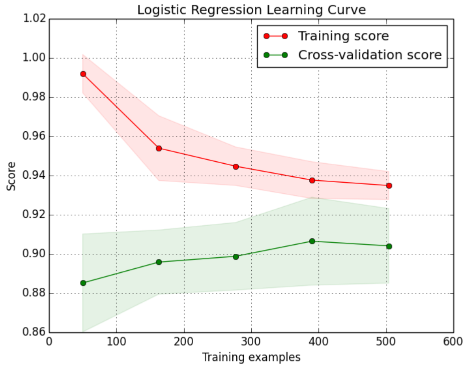


Figure 3b: Learning curves for the linear SVM.

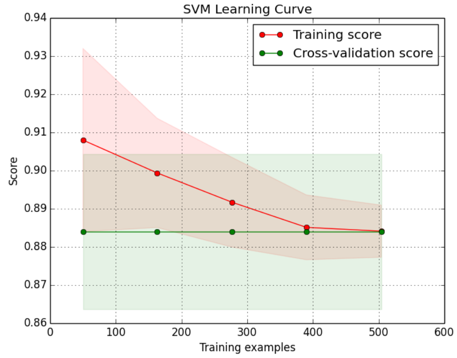
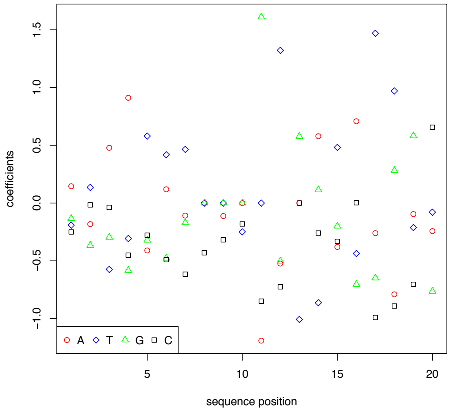


Figure 4: Coefficients for logistic regression model.



I analyzed the logistic regression model's coefficients to identify nucleotide preferences found within the sequence that might indicate context specific sequence preferences (Fig 4). Positions 8-12 mark the binding site of Cbf1. Despite the fact that Cbf1 binds both model contexts, it shows a preference for a 'GT' in positions 11-12 in the three factor model. In addition, an 'AAT' motif in positions 3-6 is more prevalent in the three factor model and has been experimentally shown to affect Cbf1's recruitment of Met4 to DNA (2).

Using the fact that AAT appears predictive of models requiring more specificity, we can label Cbf1 sites found in the yeast genome as either being three factor binding sites or not. Here I assume that sites containing the 'AAT' motif are indicative of sites that are more functional than sites that lack that motif. I assume this because sites that show more sequence specificity might be more functionally important in regulating particular sets of genes at precise timepoints.

However, I was unable to select for features that were particularly predictive of this label ('AAT' presence/absence) Cross-validation showed me that my models predicted little better than random (Table 2). This might be due to the fact that my labels are unbalanced.

Table2: Model performance for determining features relevant to functional genomic regulatory sites

	10x CV AUC
LR	0.54
SVM	0.59

5. Discussion

Identifying rules as to why proteins prefer to bind certain sites under varying conditions can help identify how genes are regulated. This valuable information can then be used to understand how cells regulate genes during development and how regulation of these genes can be disrupted during illness or disease.

Trying to distinguish sequence preferences between two contexts is not a novel idea. There are a variety of tools that attempt to accomplish this goal (3,4). However, many of these tools struggle to to distinguish subtle differences between the same protein binding under slightly different conditions.

Here I attempted to implement my own models to distinguish between the binding preferences of Cbf1 with differing cofactors present. This dataset provided an opportunity to detect subtle binding site preferences between the two contexts. Identifying the 'AAT' motif that has been shown to be present in the three factor model in previous literature (2) indicated that straightforward logistic regression could pick up these differences. One interesting result was that the SVM model mean CV score seemed unaffected by increasing training example size. This could be due to the fact that the data is easily separable even with smaller training sizes. One thing to note is that the models treat these nucleotide positions independently. It might be worthwhile to throw in terms that incorporate dependency between adjacent nucleotides. However, doing so might lead to over-fitting.

Attempting to understand how other genomic features could predict the presence of this 'AAT' motif proved to be less fruitful. The fact that the labels were unbalanced (more negatives than positives) likely made this problem harder. In addition, Cbf1 sites within their genomic contexts are more likely to be false positives (not real Cbf1 binding sites) and this increases the noisiness of the dataset.

6. Conclusion

Using logistic regression and SVM, I was able to identify sequence preferences when Met4 is recruited to sequences with Cbf1 and Met28 (three factor model) compared to when just Cbf1 is present (two factor model). More specifically, I identified the presence of an 'AAT' motif that was predictive of the three factor model that has been shown experimentally imporatant(2). Using this 'AAT' motif to label three-factor binding sites within the genome, I hoped to identify features that were predictive of this motif but was unsuccessful in this pursuit.

7. Future Directions

Future models attempting to distinguish Cbf1 binding preferences may consider the dependency between adjacent nucleotides. In attempting to identify functional three factor Cbf1 binding sites within their genomic context, I could spend more time refining the features before including them in the model. For example, I could alter the false positive rate when calling Cbf1 or Met31 binding sites. In addition, I could include the expression magnitude of the nearest gene as a feature. Most importantly, I will switch to a metric that handles classes of differing sizes (f-statistic, Mathews correlation coefficient). This is especially important because there are very few positives in this particular analysis.

8. References:

1. Lee et al. Dissection of Combinatorial Control by the Met4 Transcriptional Complex. Molecular Biology Cell. (2009)
2. Siggers T. et al. Non-DNA-binding cofactors enhance DNA-binding specificity of a transcriptional regulatory complex. Molecular Systems Biology (2011)
3. Yao Z et al. Discriminative motif analysis of high-throughput dataset. Bioinformatics (2013)
4. Patel R et al. Discriminative motif optimization based on perceptron training. Bioinformatics (2013)
5. Shultzaberger et al. Determining Physical Constraints in Transcriptional Initiation Complexes Using DNA Sequence Analysis. PLOSone. (2007)