Predicting Gene Expression Values from Yeast Ribosomal Protein Promoter Sequences

Teia Noel

1 Introduction

This project explores the DREAM Gene Expression Prediction Challenge. The data comes from a reporter assay, wherein 90 promoters, known to regulate the expression of ribosomal protein (RP) genes in *Saccharomyces cerevisiae*, were each inserted upstream of a reporter gene that encodes yellow fluorescence protein (YFP). Immersed in rich medium condition, *S. cerevisiae* with the plasmid constructs were allowed to grow, and for each promoter sequence, fluorescence intensity per cell per second was recorded. Although the mechanisms by which promoters regulate transcription are well studied, here I attempt to extract information from promoter sequences that is useful for predicting gene expression.

Because the nucleotide content of promoter sequences dictate binding sites for transcription-regulating proteins, the presence of these nucleotide signatures in RP promoters could be correlated with YFP expression. For instance, transcription factors (TFs) bind promoters to either activate or repress transcription. ChIP-Seq data revealed RP promoter architecture in which the TF binding sites Rap1 and Fhl1 occur sequentially, and through a reporter assay, deemed this configuration preferable to transcriptional activity [1]. ChIP data also showed that the TF binding site Sfp1 is almost exclusively associated with RP genes [2], and may recruit Fhl1 [3]. Lastly, Abf1 was a less frequent TF binding site found downstream of Rap1 [1], and mutations at this site depleted transcriptional activity [4]. Another content-based indicator of gene expression is CpG islands, or stretches of DNA with a high frequency of GC dinucleotides. CpG islands act as DNA methylation sites, and when located within a promoter, repress transcription of the downstream gene. Beyond promoter sequence content, measures of DNA structure such as free energy and denaturation have been shown to accurately detect promoter sequences [5]. It is of interest whether these same metrics are indicative of promoter sequences' transcriptional activity.

For these reasons, I hypothesize that the following metrics computed from *S. cerevisiae* RP promoter sequences are predictive of gene expression levels: (1) Log-odds of finding binding motifs for known *S. cerevisiae* and putative RP TF binding sites, (2) Measures of CpG islands, (3) GC content, and (4) free energy. By way of regression and classification, I answered whether these features can accurately predict either gene expression values, or levels of gene expression.

2 Methods

2.1 Deriving Features

In order to see if the presence of TF binding sites are predictive of gene expression, I defined a score for each motif in each promoter corresponding to known and putative TF binding sites. ScerTF is a database of position weight matrices (PWMs) composed of the nucleotide frequencies of each position in a motif, corresponding to ≈ 1200 motifs found in *S. cerevisiae*. These PWMs were those found in literature that were deemed the most predictive of motifs in in vivo data [6]. From ScerTF, I collected PWMs corresponding to the TF binding motifs for Rap1, Abf1, Fhl1, and Spf1. Additionally, I found the top five putative 15 bp motifs from the expectation maximization motif-finding algorithm MEME, and Gibbs sampling algorithm, BioProspector [7, 8]. In order to compute a motif score in a promoter, I computed the max log-odds of finding the motif in all k-mers of the motif: $argmax_s(\sum_{j=1}^{k} log_2 P(N_{ij}) - log_2(k * 0.25))$, where k is the length of the motif, s is the set of k-mers in the promoter, $P(N_{ij})$ is the frequency of nucleotide i at position j from the motif's PWM, and the subtracted term defines the background probability that assumes an equal chance of observing each nucleotide.

In terms of CpG islands, I used the criteria that an island must contain 200 basepairs with $> 50\%$ GC content (frequency of G and C nucleotides in the 200-bp stretch) and $> 60\%$ observed-to-expected CG

dinucleotide ratio. The observed-to-expected ratio was calculated by dividing the number occurrences of the CG dinucleotide by $((\#Cs) * (\#Gs))/200$. In order to count CpG islands for a promoter, I scanned every 200-mer, and counted the number of times the GC content and observed-to-expected CpG ratio crossed the 50% and 60% thresholds, respectively. Additionally, I gathered the maximum GC content and the maximum observed-to-expected CpG ratios across all 200-mers per promoter.

Lastly, for structural features, I measured free energy and DNA denaturation for each promoter. For free energy, I used the nearest neighbor method, defined by $\Delta G^{\circ}_{37} = \Delta G^{\circ}_{37}(initiations) + \sum_{i=1}^{n-2} \Delta G^{\circ}_{37}(i)$, where n is the length of the promoter, $\Delta G^{\circ}_{37}(initiations)$ are the free energies of the leading and terminating nucleotide pairs, and $\Delta G^{\circ}_{37}(i)$ is the free energy of the i^{th} dinucleotide pair [9]. The free energies used in this calculation were those derived experimentally from 108 oligonucleotide duplexes [9]. DNA denaturation was measured by the GC content of each promoter. The reasoning here is that Gs form triple bonds with Cs, whereas As form double bonds with Ts, and therefore GC content correlates positively with melting temperature and negatively with denaturation.

2.2 Feature Selection

In order to select for features relevant in predicting gene expression values, I evaluated the Pearson coefficient of correlation by correlating each feature to the gene expression values. Furthermore, for each feature vs. expression levels comparison, I computed the p-value, the probability of observing the correlation coefficient given the null hypothesis that there is 0 correlation between the two variables. I selected features most highly correlated with the expression value response values and with p-values$<$0.05.

2.3 Predictive Models

I chose a number of regression models to predict gene expression with the selected features, including multiple linear regression and quadratic support vector regression. For classification, I redefined the output variables by assigning 0s to low expression values, or those falling below the median, and 1s to high expression values, or those falling above or at the median. The classification models I chose to predict either high or low expression values with the selected features were logistic regression and quadratic support vector classification. I scored the regression models by conducting 5-fold cross validation, and finding the maximum R^2 value across all folds. I scored the classification models by finding the maximum mean accuracy across a 5-fold cross validation.

3 Results

3.1 GC content, observed to expected GC dinucleotide content, and Fhl1 and Sfp1 TF binding motif scores were most correlated with gene expression values

Feature selection analysis deemed total GC content and max GC content out of all 200-mers across all promoters most negatively correlated with gene expression values (r=-0.239;p= 0.031 and r=-0.227;p=0.023, respectively; Figure 1A) , while observed to expected GC dinucleotide content, and Fhl1 and Sfp1 ScerTF binding motif scores for all promoters were found to be most positively correlated with gene expression values (r=0.232;p= 0.027, r=0.292;p=0.005, and r0.256=;p=0.014, respectively; Figure 1A). I selected all of these features but max GC content for all 200-mers for my predictive models, as this was strongly correlated with total GC content across all promoters (Figure 1B).

3.2 Neither regression nor classification acted as outstanding predictors of gene expression

While quadratic support vector regression produced a higher R^2 score than multiple linear regression, neither score was far above the baseline score of 0 (Figure 2A). Furthermore, while quadratic support vector classification yielded a higher mean accuracy score than logistic regression, neither scores were far beyond 0.5, a score that you would expect if the classifier randomly guessed the correct classifications (Figure 2B). Thus, neither regression nor classification performed exceptionally well in predicting

expression values and expression levels, respectively.

Figure 1A

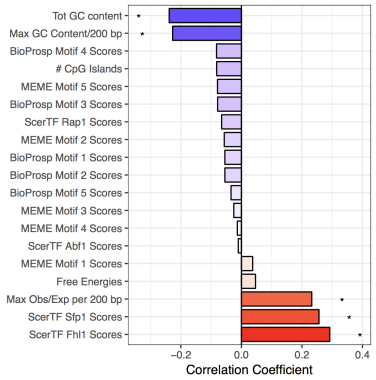


Figure 1B

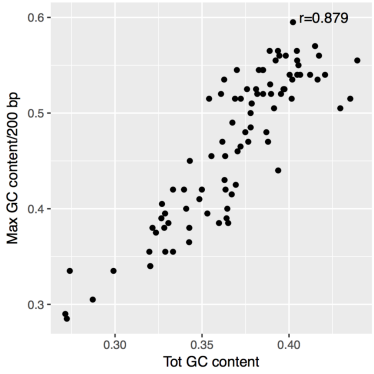


Figure 2A

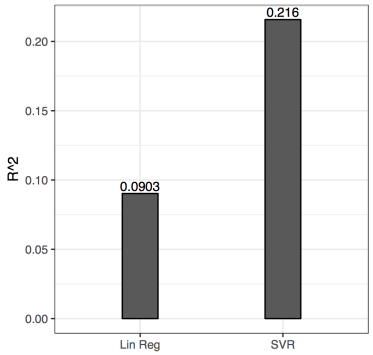
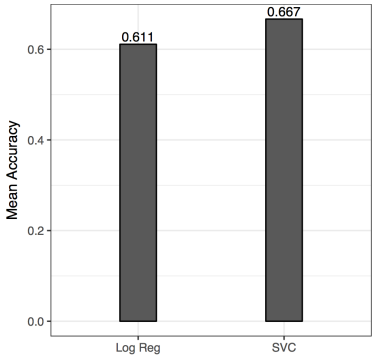


Figure 2B



4 Discussion

While the analysis presented here failed to produce a strong predictor of gene expression given RP promoter sequences, feature selection produced some promising and biologically relevant results. Out of all features, GC content per promoter produced the largest negative correlation coefficient when correlated with gene expression levels. This makes sense because high GC content is associated with high DNA methylation levels, which in turn, blocks TF binding and silences gene expression. Furthermore, Sfp1 and Fhl1 motif scores correlated positively with gene expression values. Studies have shown that Sfp1 recruites the Fhl1 TF, which is found in a TF binding site in favor of transcriptional activity.

For future analyses, it would be advantageous to gather additional reporter assay data, since the data associated with 90 promoters provided by this challenge were sparse. With additional data, it would be more feasible to use motif-finding algorithms like BioProspector and MEME, used in this study, along with DeepBind, which utilizes deep learning. Additionally, large amounts of data would make finding a predictive model more robust. Perhaps a larger dataset of expression values associated with many types of promoters can yield general features predictive of expression in an organism. Furthermore, additional features can be explored, including positions of motifs, or aggregate scores of multiple high-scoring motifs in a promoter. If stronger features are found, additional models can be tried and tested. Assuming a reliable predictive model is generated, in vitro mutational analyses, as well as varying experimental conditions can be induced on promoters to analyze how these conditions affect gene expression levels.

References

[1] B. Knight, S. Kubik, B. Ghosh, M. J. Bruzzone, M. Geertz, V. Martin, N. Dénervaud, P. Jacquet, B. Ozkan, J. Rougemont, *et al.*, "Two distinct promoter architectures centered on dynamic nucleosomes control ribosomal protein gene transcription," *Genes & development*, vol. 28, no. 15, pp. 1695–1709, 2014.

[2] R. M. Marion, A. Regev, E. Segal, Y. Barash, D. Koller, N. Friedman, and E. K. O'Shea, "Sfp1 is a stress-and nutrient-sensitive regulator of ribosomal protein gene expression," *Proceedings of the National Academy of Sciences*, vol. 101, no. 40, pp. 14315–14322, 2004.

[3] P. Jorgensen, I. Rupeš, J. R. Sharom, L. Schneper, J. R. Broach, and M. Tyers, "A dynamic transcriptional network communicates growth potential to ribosome synthesis and critical cell size," *Genes & development*, vol. 18, no. 20, pp. 2491–2505, 2004.

[4] R. F. Lascaris, E. d. Groot, P.-B. Hoen, W. H. Mager, and R. J. Planta, "Different roles for abf1p and a t-rich promoter element in nucleosome organization of the yeast rps28a gene," *Nucleic acids research*, vol. 28, no. 6, pp. 1390–1396, 2000.

[5] Y. Gan, J. Guan, and S. Zhou, "A comparison study on feature selection of dna structural properties for promoter prediction," *BMC bioinformatics*, vol. 13, no. 1, p. 4, 2012.

[6] A. T. Spivak and G. D. Stormo, "Scertf: a comprehensive database of benchmarked position weight matrices for saccharomyces species," *Nucleic acids research*, vol. 40, no. D1, pp. D162–D168, 2011.

[7] T. L. Bailey, N. Williams, C. Misleh, and W. W. Li, "Meme: discovering and analyzing dna and protein sequence motifs," *Nucleic acids research*, vol. 34, no. suppl_2, pp. W369–W373, 2006.

[8] X. Liu, D. L. Brutlag, and J. S. Liu, "Bioprospector: discovering conserved dna motifs in upstream regulatory regions of co-expressed genes," in *Biocomputing 2001*, pp. 127–138, World Scientific, 2000.

[9] J. SantaLucia, "A unified view of polymer, dumbbell, and oligonucleotide dna nearest-neighbor thermodynamics," *Proceedings of the National Academy of Sciences*, vol. 95, no. 4, pp. 1460–1465, 1998.