A multi-modal neural network for learning cis and trans regulation of stress response in S. cerevisiae

Boxiang Liu[1,2,3], Nadine Hussami[4], Avanti Shrikumar[3,5], Tyler Shimko[3], Salil Bhate[6], Scott Longwell[6], Stephen Montgomery[2,3], and Anshul Kundaje[3,6]

[1]Departments of Biology, [2]Pathology, [3]Genetics, [4]Statistics, [5]Computer Science, and [6]Bioengineering, Stanford University
{bliu2,nadinehu,avanti,tshimko,bhate,longwell,smontgom,akundaje}@stanford.edu

Abstract

Understanding gene regulatory mechanisms is a central problem in computational biology. Here, we explore the use of interpretable multi-modal neural networks to learn cis and trans regulation of stress response in the budding yeast *Saccharomyces cerevisiae*. We formulate the problem as a regression task, where the goal is learn a model that can accurately predict the real-valued gene expression of all genes across 173 different stress conditions based on two complementary regulatory inputs - a cis component represented by the raw promoter sequence of each gene and a trans component represented by the real-valued expression of a subset of regulatory genes (transcription factors and signaling molecules). The multimodal neural network includes a convolutional module to learn predictive patterns from the raw promoter sequences, a dense module to derive features from regulator expression and an integration module that learns cis-trans interactions. We use a variety of cross-validation settings to evaluate the performance of the model (held-out genes, held-out stress conditions). In all cases, the models achieve high performance and substantially outperform other state-of-the-art methods such as boosting algorithms that pre-defined cis-regulatory features (known motifs). We then use an efficient backpropagation algorithm that we recently developed to interpret the model. We interpret the model to reveal dynamic promoter sequence motif grammars, trans regulator modules and motif-regulator associations affecting individual genes and gene modules in diverse stress conditions. Our model correctly identifies known master regulators such as MSN4/MSN2, USV1, YVH1 as well as several stress-specific factors. We also use our models to perform in silico knock-out/knock-in experiments. In a MSN2/4 knockout experiment, we demonstrate that in silico predictions of target gene changes strongly correlate with the results of the corresponding knockout microarray experiment.

1 Introduction

The accurate prediction of gene expression based on the interaction between transcription regulators and DNA sequence is a critical milestone toward fully comprehending cellular regulome. Several studies have contructed computational models to tackle this problem. Some use the sequence information [1, 2], while others use the transcription regulator expression [3, 4]. In particular, GeneClass[5] and BDTree[6] model gene expression based on the interaction between sequence motif and regulator, and thus can discover previously unknown regulatory patterns. However, due to modeling assumptions, both methods require discretized expression values, which severely limits their utility. Further, they both assume prior knowledge on sequence motifs, which is unavailable in certain cases. Perhaps unsurprisingly, aforementioned studies all use yeast as the model organism.

The budding yeast *Saccharomyces cerevisiae* possesses a relatively simple transcriptional regulatory architecture governed primarily by promoter sequence directly upstream of the gene. Such architecture has very few long range interactions such as those in higher-order organisms, and is ideal for computational modeling.

We present a deep neural network architecture to address the limitations with previous methods. Our architecture models gene expression as the interaction between trans (signaling molecules and transcription factors) and cis (1kbp promoter sequence) regulatory components. Our model automatically extract sequence motifs from raw sequences, thus eliminating the need for prior knowledge and feature engineering. Our model predicts gene expression values in $\mathbb{R}$, akin to microarray and sequencing measurement. We demonstrate that our model outperforms the state-of-the-art methods, and provides sensible prediction concordant with existing knowledge of yeast biology.

2 Methodology

2.1 Sequence module

We treat one-hot encoded promoter sequences as black-and-white images of size $\{0,1\}^{4\times 1000}$. The promoter sequences contain information about transcription factor binding motifs, each representing a consensus sequence (e.g. 'TGATCA'). Like an object in natural images, a motif appears in multiple promoter sequences and a given promoter can contain several instances of the same motif with slight variations. We use a 2-layer convolutional neural networks to detect such motifs. Mathematically, the 2D convolution layers with ReLU activation can be written as:

$$x_{ij} = \max\{0, \sum_{a=1}^{4}\sum_{b=1}^{l} W_{abj}^{seq} s_{(a)(i+b)} + b^{seq}\} \tag{1}$$

Where s represents sequence input, W^{seq} represents the weight for convolutional filters, i represents the genomic location, j represents filter number, b^{seq} represents the bias, and l represents the filter length. We use $l = 9$ and $J = 50$ filters for both convolutional layers with a subsampling rate of 1. To keep model prediction invariant to small motif shifts along the genomic axis, we apply 1D maxpooling with a subsampling rate of 4 after each convolutional layer. We applied batch normalization before each ReLU activation to mitigate covariate shifts during training and to accelerate learning. Since both strands of DNA encode identical information, we incorporated the reverse complement sequences into our model, effectively doubling the number of training examples, and used filter pairs with mirroring weights to detect each motif [7]. We used a fully connected layer with 512 units to integrate information across filters.

2.2 Regulator module

Transcription factors and signaling molecules interact with each other through bi-, tri-molecular and even higher order reactions with no genomic-spatial constraint. We model the regulator input with a fully connected layer of 512 hidden units. Mathematically,

$$y = \max\{0, W^{reg} r + b^{reg}\} \tag{2}$$

2.3 Integration module

In principle, the output from sequence module represents gene-specific information, and that from the regulator module represents condition-specific information. Treated separately, neither module will be able to predict the gene-condition interactions critical for accurate final prediction. We integrate two separate sources of information through a simple concatenation layer, and use a 2-layer fully connected network to model the interaction between motifs and regulators.

2.4 Dataset and Pre-processing

We utilized the Gasch [8] dataset which used microarray to measure a total of 6100 genes under 173 experimental conditions. The measurements were given as log_2 expression values representing the fold change w.r.t. the untreated reference condition. We used 1kbp sequences upstream of the transcription start site as the sequence predictors and a set of 472 transcription factors and signaling molecules as the regulator predictors. We used a 80-10-10 split among training, validation, and test datasets.

3 Results

3.1 Regression and Classification Performances

We compared our model against two state-of-the-art models, GeneClass [5] and BDTree [6], using the same dataset by Gasch et al [8]. The GeneClass model is a boosted alternating decision tree, and the BDTree model is a bidirectional regression tree. We trained our model to predict real-valued expression levels as well as discrete labels in $\{-1, 0, 1\}$ to represent upregulation, baseline, and downregulation. In the classification task, the dataset is discretized such that expression values in $[-\infty, -0.5]$ are converted to -1, $[-0.5, 0.5]$ to 0, and $[0.5, \infty]$ to +1. Our best model outperformed the previous state-of-the-art by 16.6% (Table 1). In addition, our model achieved a pearson correlation of 0.82 for the regression task on the test set (Figure **??**). For regression, we could not compare our models because GeneClass or BDTree only reported classification results.

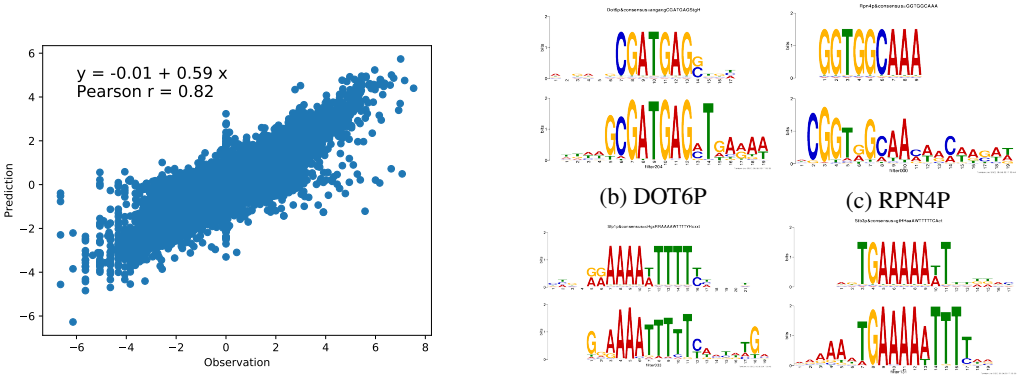
Table 1: Classification performance

Method	Accuracy
GeneClass	60.9%
BDTree	62.9%
DNN	**79.5%**

	Predicted by DNN		
	Down	Baseline	Up
Down	10.14	7.47	0.13
Baseline	3.29	59.77	3.02
Up	0.18	6.42	9.59

3.2 Recovering Known Motifs

Unlike GeneClass and BDTree which rely on existing motif annotations, our model learn filter weights and extracts motif information automatically. We found that our model learns both known and *de novo* motifs. Figure 1 shows four known motifs recovered by our model.

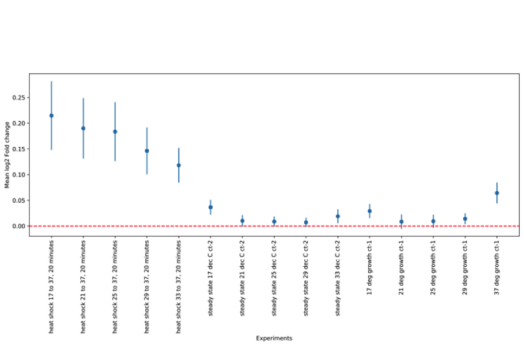


(a) Correlation (b) DOT6P (c) RPN4P (d) SFP1P (e) STB3P

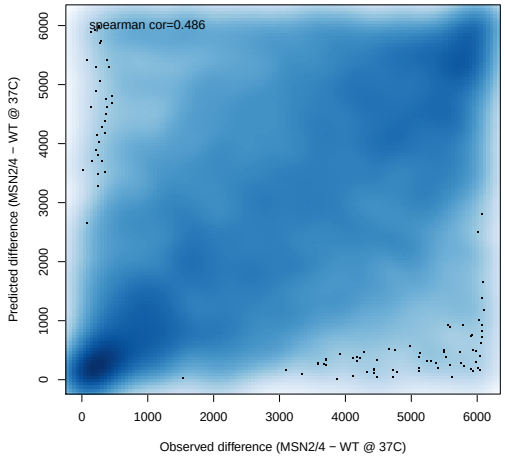
Figure 1: (a) Predicted vs grounth truth (b-e) examples of recovered motifs

3.3 Recovering Master Regulators

A few studies have shown that yeast master regulators often manifest as key feature in prediction models [9]. To understand whether our model reflect the tran-regulatory component, we decided to



(a) Target gene expression change



(b) Predicted vs ground truth

Figure 2: (a) The expression of MSN2/4 known target gene experience larger change under stress conditions. (b) Predicted expression change vs actual microarray experiment.

rank the relative importance of the regulator module inputs, and test if such rank captures the master regulators in *S. cerevisiae*. We reasoned that perturbations in regulators with high feature importance would lead to large change in the output. Therefore, we estimated the relative importance as the gradient (w.r.t to output) times the magnitude of input (G-by-I). We summed the G-by-I values across all genes and conditions to obtain the global estimate of feature importance. Notably, several master regulators such as USV1, YVH1, MSN4 appear as top features.

Table 2: Rank of regulator module inputs

Rank	Regulator
1	USV1 / YPL230W
2	DAL80 / YKR034W
3	XBP1 / YIL101C
4	PPT1 / YGR123C
5	LSG1 / YGL099W
6	CIN5 / YOR028C
7	YVH1 / YIR026C
8	TPK1 / YJL164C
9	GAC1 / YOR178C
10	MSN4 / YKL062W

3.4 *in silico* Mutagenesis Reflects *in vivo* Measurements

Ultimately, the model should make predictions similar to actual experimental measurement, even for previously unseen conditions. Therefore, we performed *in silico* knockout experiment on MSN2/4, where we replaced all instances of 'AGGGG' with neutral 'NNNNN', and reduced the expression level of MSN2/4 by 32 fold. MSN2/4 are activated under heat shock but are inactive under steady-state growth [10]. As expected, we observed that MSN2/4 target genes experience greater change under heat shock conditions than steady-state growth condition 2a. We also compared our prediction against microarray measurement in a actual MSN2/4 knockout strain, and observed general agreement between the two (spearman correlation = 0.486, 2b)

References

[1] Harmen J Bussemaker, Hao Li, and Eric D Siggia. Regulatory element detection using correlation with expression. *Nature Genetics*, 27(2):167–174, February 2001.

[2] Tu Minh Phuong, Doheon Lee, and Kwang Hyung Lee. Regression trees for regulatory element identification. *Bioinformatics*, 20(5):750–757, March 2004.

[3] Lev A Soinov, Maria A Krestyaninova, and Alvis Brazma. Towards reconstruction of gene networks from expression data by supervised learning. *Genome biology*, 4(1):R6, January 2003.

[4] Eran Segal, Michael Shapira, Aviv Regev, Dana Pe'er, David Botstein, Daphne Koller, and Nir Friedman. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics*, 34(2):166–176, June 2003.

[5] Manuel Middendorf, Anshul Kundaje, Chris Wiggins, Yoav Freund, and Christina Leslie. Predicting genetic regulatory response using classification. *Bioinformatics*, 20(suppl 1):i232–i240, August 2004.

[6] Jianhua Ruan and Weixiong Zhang. A bi-dimensional regression tree approach to the modeling of gene expression regulation. *Bioinformatics*, 22(3):332–340, February 2006.

[7] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Reverse-complement parameter sharing improves deep learning models for genomics | bioRxiv.

[8] A P Gasch, P T Spellman, C M Kao, O Carmel-Harel, M B Eisen, G Storz, D Botstein, and P O Brown. Genomic expression programs in the response of yeast cells to environmental changes. *Molecular Biology of the Cell*, 11(12):4241–4257, December 2000.

[9] Anshul Kundaje, Manuel Middendorf, Mihir Shah, CHRIS H WIGGINS, Yoav Freund, and Christina Leslie. A classification-based framework for predicting and analyzing gene regulatory response. *BMC bioinformatics*, 7(1):S5, March 2006.

[10] A Sadeh, N Movshovich, M Volokh, L Gheber, and A Aharoni. Fine-tuning of the Msn2/4-mediated yeast stress responses as revealed by systematic deletion of Msn2/4 partners. - PubMed - NCBI. *Molecular Biology of the Cell*, 22(17):3127–3138, August 2011.