Supplemental Material: 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*

October 9, 2017

Contents

1 Yeast Microarray Data **2**

2 Neural Network Architecture **2**

3 Motif discovery **2**

4 Ranking regulator by feature importance **5**

5 *in silico* Mutagenesis **5**

List of Figures

1 Dataset 3
2 Architecture 4
3 Example of known motifs recovered by the convolutional filters . 5
4 Distribution of change in expression stratified by MSN2/4 target or non-target genes 7
5 Difference of change in expression between by MSN2/4 target and non-target genes 7
6 Comparison of *in silico* knockout and microarray measurement . 8

Layer	Units	Size	Stride
Sequence module			
Input	1000	-	-
Conv1D	50	9	1
Maxpool1D	-	4	4
Conv1D	50	9	1
Maxpool1D	-	4	4
Dense	512	-	-
Reg module			
Input	472	-	-
Dense	512	-	-
Integration module			
Concatenation	1024 (512 + 512)	-	-
Dense	512	-	-
Dense	512	-	-

Table 1: Hyperparameters

1 Yeast Microarray Data

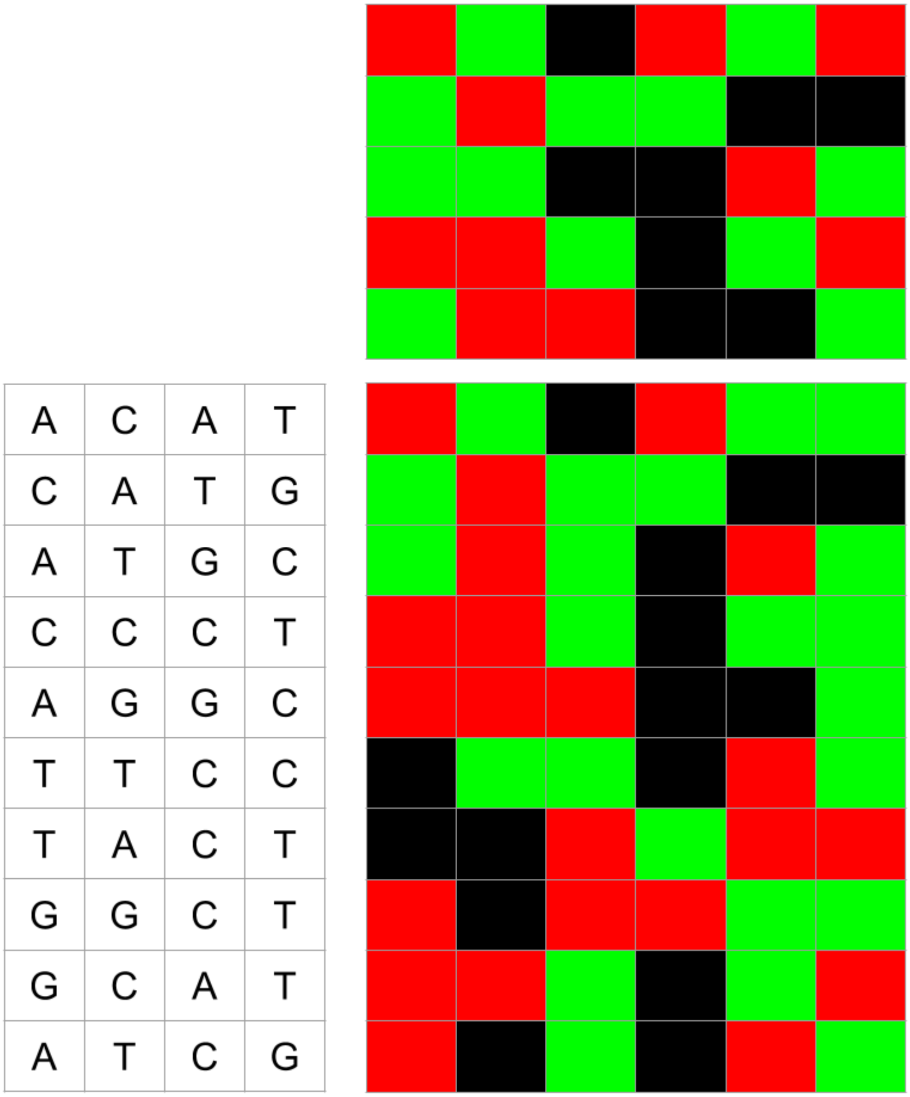
In this study we used transcriptome microarray measuring cDNA abundance by Gasch et al[1]. In total, there are 6110 genes across 173 experimental conditions. The dataset were given as as log_2 fold change w.r.t an unstimulated reference expression. We decided not to performance further normalization to preserved the interpretation of true zero, i.e. no change w.r.t the reference. We selected 472 signaling molecules and transcription factors as input the regulatory module (see Section 2). We used 1kbp promoter sequence directly upstream of the TSS as input to the sequence module. The data can be visualized in graphical form as follows. Each column represent a experimental condition, and each row is a gene.

2 Neural Network Architecture

The detailed architecture is in Fig. 2. All hyperparameters are listed in Table 1. For simplicity we omitted batch normalization, activation and dropout layers from Table 1. If not mentioned otherwise, all activations are rectified linear units and all dropout uses a keep rate of 0.5.

3 Motif discovery

We used a method similar to Basset [2] to perform motif discovery. For each convolutional filter, we select the 100 sequence segments with the highest activation. We next calculate the PWM based on nucleotide frequency in these



🟥	🟩	⬛	🟥	🟩	🟥
🟩	🟥	🟩	🟩	⬛	⬛
🟩	🟩	⬛	⬛	🟥	🟩
🟥	🟥	🟩	⬛	🟩	🟥
🟩	🟥	🟥	⬛	⬛	🟩

A	C	A	T
C	A	T	G
A	T	G	C
C	C	C	T
A	G	G	C
T	T	C	C
T	A	C	T
G	G	C	T
G	C	A	T
A	T	C	G

🟥	🟩	⬛	🟥	🟩	🟩
🟩	🟥	🟩	🟩	⬛	⬛
🟩	🟥	🟩	⬛	🟥	🟩
🟥	🟥	🟩	⬛	🟩	🟩
🟥	🟥	🟥	⬛	⬛	🟩
⬛	🟩	🟩	⬛	🟥	🟩
⬛	⬛	🟥	🟩	🟥	🟥
🟥	⬛	🟥	🟥	🟩	🟩
🟥	🟥	🟩	⬛	🟩	🟥
🟥	⬛	🟩	⬛	🟥	🟩

Figure 1: Dataset

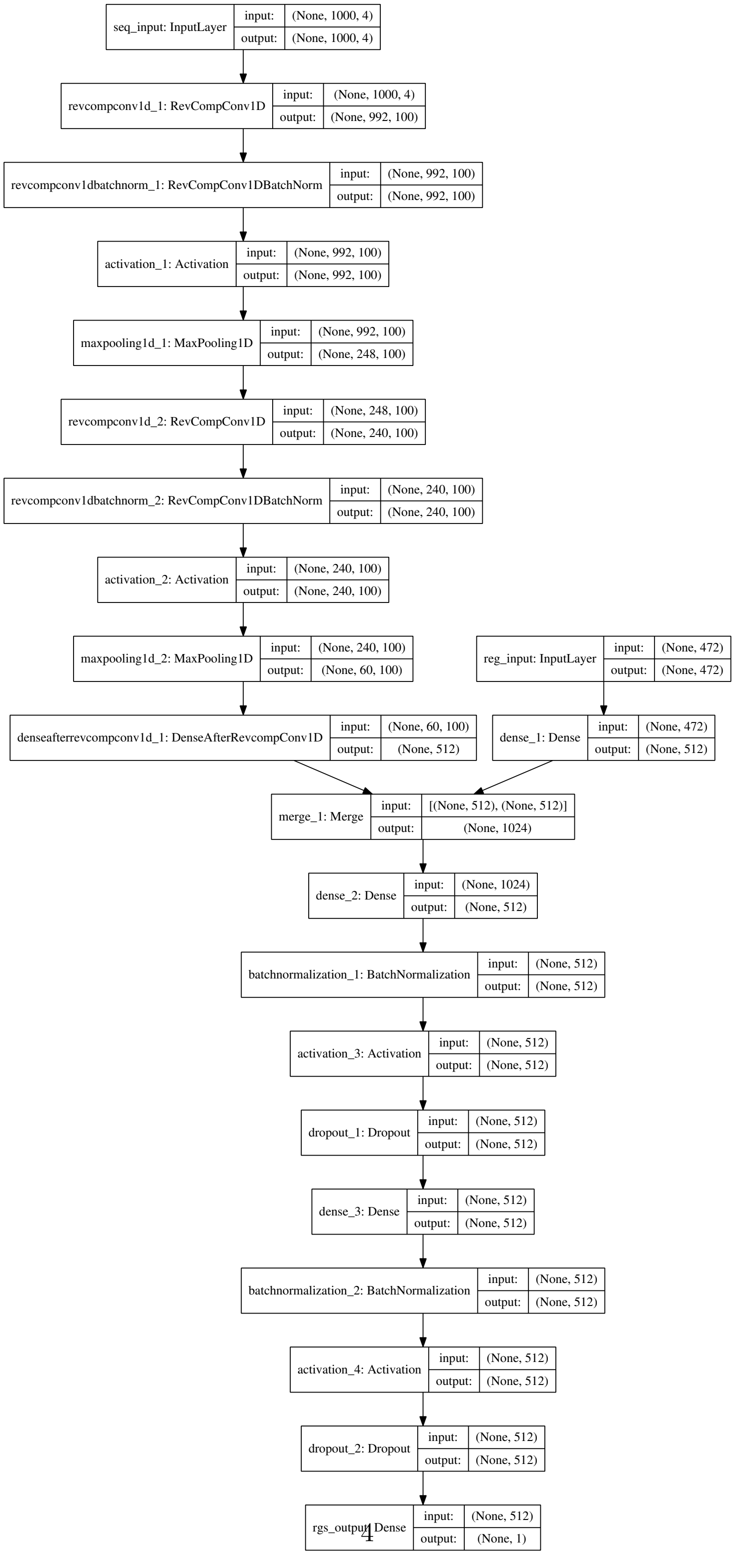


Figure 2: Architecture

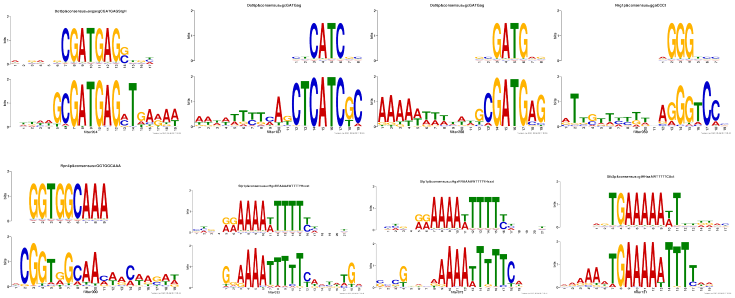


Figure 3: Example of known motifs recovered by the convolutional filters

sequence segments. To test whether any PWM correspond to known motif, we used TomTom (http://meme-suite.org/tools/tomtom) to compare against the YEASTRACT database. Below we show several more instances of motif matches.

4 Ranking regulator by feature importance

To understand which transcription factors and/or signaling molecules are most predictive, we used a backpropagation approach to rank the each input to the regulatory module [3]. We reasoned that perturbation to highly influential features would lead to large change in the output. We quantified the feature importance using gradient (w.r.t output) times the input [4]. To get the overall importance, we summed the gradient times input across all conditions and all genes.

5 *in silico* Mutagenesis

To test whether our model makes biologically relevant predictions, we decided to compare *in silico* MSN2/4 knockout against actual microarray measurement. MSN2/4 are two master transcription regulators with about 300 target genes. They are only activated under stress conditions such as heat shock. We removed the MSN2/4 binding sequence ('AGGGG' and its reverse complement sequence) and reduced the expression by 32 fold. Since MSN2/4 are only activated under stress conditions, we expect that the expression of their target genes will have larger change under stress than normal growth conditions. As expected, in heat shock experiments (the first five groups), MSN2/4 target genes show more pronounced decreases in expression values compared with non-target genes. This indicates that MSN2/4 has a greater influence in known target genes (Fig. **??**). As a negative control, we looked at steady-state growth conditions. In these conditions, since MSN2/4 are not activated, knocking them down should not influence the target genes. Indeed, we observe smaller differences in the next five groups, which are steady-state growth experiments. In addition, the smaller errorbars indicate smaller overall effect by MSN2/4 knockdown. The same pattern can be observed for the last five groups corresponding to exponential growth conditions at different temperature. Notably, the last group, exponential growth at 37 C, shows a pattern close to a heat shock experiment. We reason that heat-shock experiments are generally conducted by raising culturing temperature temporarily to 37 C, and exponential growth at this temperature disrupts the normal function of the yeast regulatory machinery, thus activating MSN2/4. Fig. 5 shows the difference in mean between MSN2/4 target and non-target genes. The heat shock experiments clearly show greater differences than steady-state and exponential growth conditions. Interestingly, as the heat shock becomes milder, the difference becomes smaller. Again, we observe that 37 C growth is similar to heat shock.

As an orthogonal validation, we performed *in silico* MSN2/4 knockout experiments for yeast exposed to 37C heat shock. The resultant predictions were compared with actual MSN2/4 strains exposed to 37 C heat shock. We ranked the genes according to the changes in their expression - downregulation is ranked higher than upregulation. Their rank correlation is shown below as a smoothed scatter plot. The rank is consistent for extreme upregulated/downregulated genes. However, for genes whose expression underwent less dramatic change, the prediction can be noisy. We reason that this is due to regulatory buffering. Genes with large expression change are often directly regulated by MSN2/4, where those with smaller change are indirectly regulated through intermediate genes. Since we only knock out MSN2/4, directly regulated genes will be strongly affected, and therefore show consistency with microarray measurement. On the other hand, the effect of MSN2/4 knockout may not penetrate deep enough to indirectly regulated genes due to regulatory buffering.

References

[1] 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.

[2] David R Kelley, Jasper Snoek, and John L Rinn. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome research*, 26(7):990–999, July 2016.

[3] Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. Not Just a Black Box: Learning Important Features Through Propagating Activation Differences. *arXiv.org*, May 2016.

[4] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the Visualization of What a

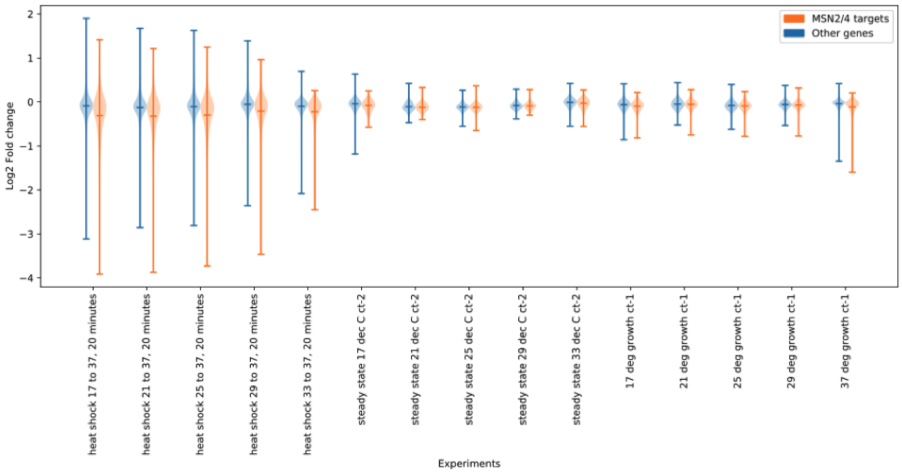


Figure 4: Distribution of change in expression stratified by MSN2/4 target or non-target genes

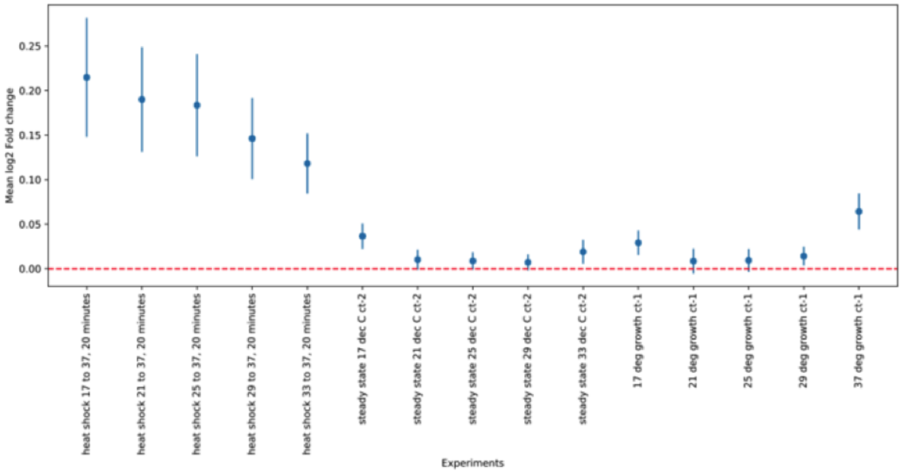


Figure 5: Difference of change in expression between by MSN2/4 target and non-target genes

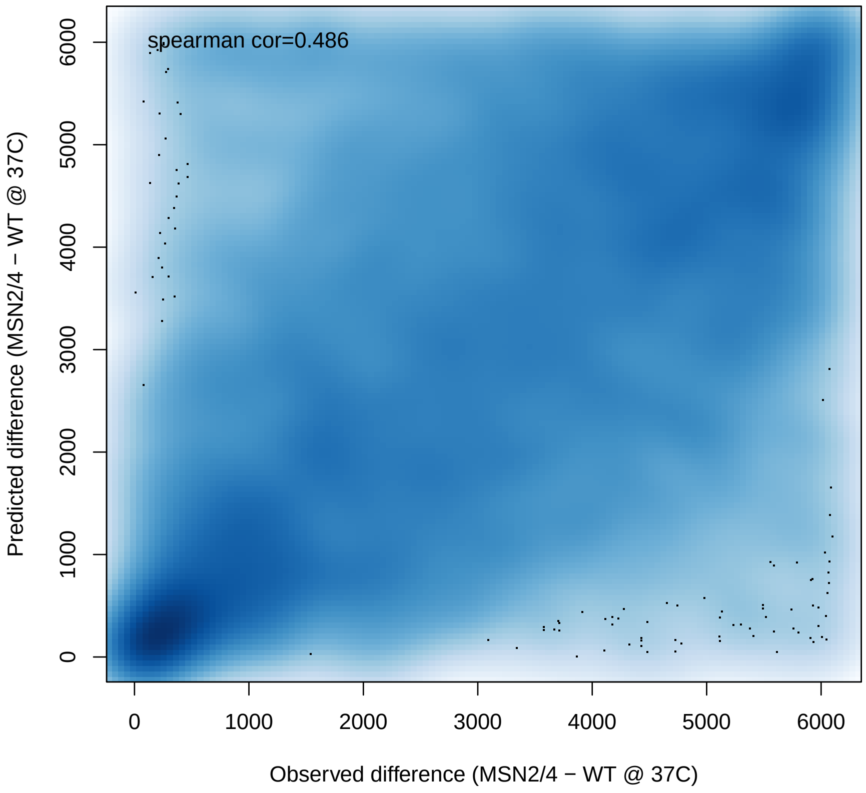


Figure 6: Comparison of *in silico* knockout and microarray measurement

Deep Neural Network Has Learned. *IEEE transactions on neural networks and learning systems*, pages 1–14, August 2016.