Pathfinder: A Tool to Predict Regulatory Networks from Time-Course RNA-Seq in *Saccharomyces cerevisiae*

Oriel Savir, Varen Talwar

Abstract

Motivation: Utilize RNA-seq time course data to predict possible gene expression pathways in yeast. We seek to develop a tool that can construct possible genetic regulatory pathways. In this paper, our model is yeast under exposure to hypoxia. The long-term goal is to create a tool that can generalize our techniques for hypoxia to be able to construct possible pathways from any RNA-seq time course dataset.

Results: We obtained a basic model for predicting regulators and regulated genes from a time-course dataset. However, more sophisticated techniques need to be employed to overcome significant caveats of false positives and inferring connections between genes, instead of merely generating broad lists of potential targets.

1 Introduction

RNA-seq is a technique that uses deep-sequencing technologies to provide a measure of levels of transcription of genes in cells. The technique relies on reverse transcription of RNA transcripts in cells, followed by sequencing of the cDNA and computational analysis to find expression levels, such as transcripts per million (TPM) (Wang et al.). Thus, using RNA-seq data taken over time intervals, it is possible to analyze the change in expression of genes.

In our data, we observed the change in gene expression in the yeast *Saccharomyces cerevisiae* under exposure to hypoxia. During hypoxia exposure, yeast sense an inability to maintain oxygen-dependent heme biosynthesis. A heme-dependent transcriptional factor in yeast, heme-HAP1, is responsible for suppressing multiple hypoxia-response genes in yeast. It does this by activating ROX1, which in turn represses hypoxic genes by binding to them (Zitomer et al.). Thus, when an absence of oxygen leads to an absence of heme, ROX1 expression falls dramatically and hypoxic genes begin expressing. Some important examples of hypoxic genes are COX5b (electron transport chain), SUT1 (sterol uptake) and ERG11 (sterol synthesis) (Zitomer et al.).

We obtained time-course RNA-seq data for *S. cerevisiae* exposed to hypoxia for 0 (no hypoxia exposure), 5, 10, 30, 60, 120, 180 and 240 minutes (Bendjilali et al.). See Section 2.1 for further details about the data.

Our goal was to infer regulatory relationships between genes using time-course RNA-seq data. Through a number of simplifications and assumptions (see 2.5), we obtained a model to classify genes as regulators and regulated. While these results come at significant risk of false positives, we believe that this is an exciting first step to constructing regulatory networks through RNA-seq time-course data.

In this report, we have only included the analysis for upregulated genes. The same pipeline can be applied to study downregulated genes too, but that is not shown here for the sake of conciseness.

2 Methods

2.1 Data collection

Time course RNA-seq data of the yeast *Saccharomyces cerevisiae* was obtained from the Saccharomyces Genome Database (SGD). The durations of hypoxia the yeast was exposed to were, in minutes, 0 (no exposure), 5, 10, 30, 60, 120, 180, 240. The database included 5320 genes and contained TPM (transcripts per million) values for all of them across all time points. Since genes express at different baseline levels, we normalized all TPM values to a scale of 0-1 by dividing TPM across time points by the greatest TPM value for the gene.

2.2 Data Refinement

To narrow down the list of genes to consider, we retained only those genes that displayed adequate variation through the time points. This was due to the assumption that genes that differentially express during hypoxia will demonstrate a dramatic change in expression, leading to highly varying TPM values across time points. Therefore, we filtered out all genes with a standard deviation less than the 75th percentile, which was 0.225. This reduced the gene list to 1267 potentially hypoxia-regulated genes.

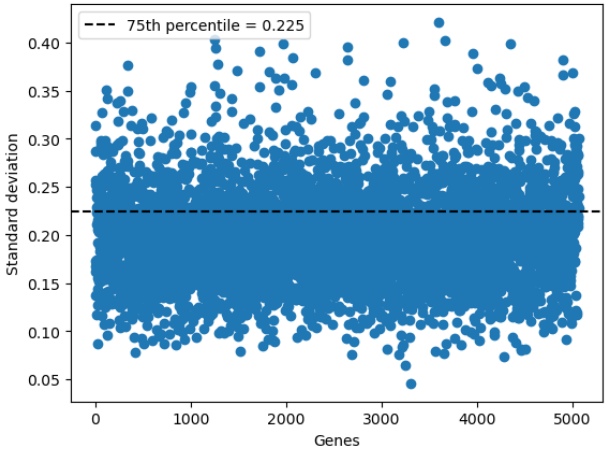

Figure 1: Genes with high enough TPM variation (standard deviation > 75th percentile) were retained

2.3 Classifying Gene Expression

Log2-fold-change (log2-FC) was used as a metric to characterize differential expression. The fold change was calculated with respect to two different references: 0 minutes (log2-FC$_0$) and the previous time-point (log2-FC$_{i-1}$). The log2-FC$_0$ reflected the difference in expression under normal oxygen concentrations and a specific duration of hypoxia exposure. On the other hand, the log2-FC$_{i-1}$ reflects the time point at which expression changes dramatically. Since our aim was to elucidate gene expression pathways using time-course data, the former metric was useful in determining the overall effect of hypoxia on the gene to categorize it as upregulated or downregulated, while the latter metric would allow us to investigate whether this change occurs consistently from the beginning, or is triggered after a delay. The latter case would suggest that the gene's expression is dependent on another regulator gene, thus making it an indirect target of hypoxia in yeast.

Therefore, these two strategies were used in tandem to offer insight into not only the general response of a gene to hypoxia, but also the directness of the response. However, some analyses were also performed separately for these two strategies to compare their results and determine which one was optimum for different situations. Finally, a gene with log2-FC (either definition) of greater than 1.5 was classified as an upregulated gene and less than -1.5 at any of the 7 time points was classified as a downregulated gene.

2.4 Classifying Relative Temporal Gene Expression

To determine when a gene was expressed after hypoxia exposure, we compared two metrics. First, we found the time point at which a gene has maximum expression (highest TPM value). An earlier peak suggests a direct response to hypoxia, whereas a later peak would suggest dependence on upstream regulators. The second metric was finding the time point at which the log2-FC$_0$ first crosses the threshold log2-FC value of ±1.5. This tells us how long does a gene take after hypoxia exposure to be induced or repressed.

2.5 Model and Assumptions

The goal of our experiment after obtaining the filtered gene list with 1267 genes was to separate them into potential regulators of hypoxia response and potential downstream effectors (regulated genes). Of course, it is expected that this simplistic model can lead to false positives as there is no category for genes neutral to hypoxia response. This assumption was made as we were working on a smaller dataset with only the most varying genes that showed significant change in expression ($|\text{log2-FC}| > 1.5$), meaning that all the genes showed sufficiently significant variation to be regarded as being differentially expressed during hypoxia exposure. The validity of this assumption can be optimized by adjusting the percentile of standard deviation we use as our cutoff, but that is outside the scope of our investigation in this project.

The second assumption we made in our analysis, as mentioned earlier, is that genes that achieve maximal expression earlier are likely to be regulators, while those that do so later are likely to be regulated. The primary challenge for this was defining a cutoff time for 'early' and 'late' expression. While an arbitrary decision could be made, we chose to base our decision on Kmeans clustering of the normalized TPM data and supplementing that with the heatmap visualizations. Since we used different metrics to quantify fold change (see 2.3/2.4), the quality of clustering became an important result to compare the efficacy of the different strategies (for results, see 3.2).

2.6 Packages and Tools

Numpy and Pandas libraries were used in Python for data processing. Heatmaps were generated using the *clustermap* function in Seaborn, while PCA (principal component analysis), TSNE and UMAP visualizations were done using Sklearn.

3 Results

3.1 Comparison of Differential Expression Metrics

The log2-FC$_0$ and log2-FC$_{i-1}$ were compared using the heatmaps they generated. The heatmap of time-course expression (normalized TPM) of genes classified as upregulated using log2-FC$_0$ demonstrated a clear division of genes reaching maximal expression at 30 minutes and earlier (regulators) and between 30 and 240 minutes (regulated), as shown in Figure 3. However, this clear differentiation was lost when using log2-FC$_{i-1}$ (Figure 4). Therefore, the log2-FC$_0$ metric was deemed more suitable for our n analyses.

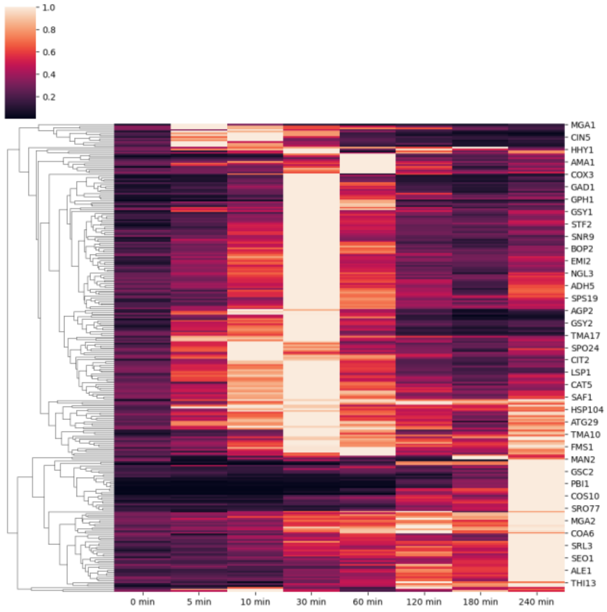

Figure 2: Heatmap of normalized TPM for upregulated genes determined using log2-FC(0)

3.2 Comparing Metrics for Maximal Expression

As described earlier, we considered two metrics to determine when a gene was maximally expressed. The first was simply finding the time point with the maximum TPM (normalized TPM = 1), while the second was to find the first time point at which the log2-FC$_0$ or log2-FC$_{i-1}$ exceeded the threshold of 1.5. The number of genes with maximal expression at each time point using these different strategies is summarized in Table 1.

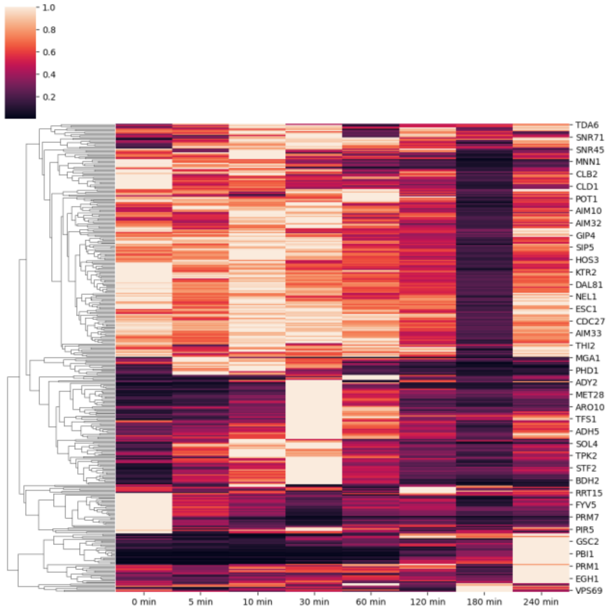

Figure 3: Heatmap of normalized TPM for upregulated genes using log2-FC(i-1)

Since the method of finding the time point with the greatest TPM value agreed best with the results from the log2-FC$_0$ heatmap, we chose to use this as the optimal strategy for downstream analyses. However, it should be noted that since the log2-FC$_0$, by definition, occurs at the time point with the peak TPM value, this metric does not add any new information to our analysis. Since the other metrics gave haphazard results similar to the log2-FC$_{i-1}$ heatmap, this part of the analysis turned out to be redundant, unfortunately.

3.3 Clustering and Validation

Based on the above analyses, there were 265 upregulated genes (log2-FC$_0$ > 1.5). Our model says that these genes can either be regulators or regulated downstream in response to hypoxia. Thus, Kmeans clustering was performed to cluster these 265 genes, which can be seen in the PCA plot in Figure X. Cluster 0 (P0) was regarded as potential regulators and cluster 1 (P1) as potential regulated genes.

Our proposed list of regulators were genes with maximal expression before or at 30 minutes (R0), while all others were proposed regulated genes (R1). The intersection of the respective groups was used to determine the success of the model.

$$P0 \cap R0 = 0.90\left(\frac{170}{188}\right)$$

$$P1 \cap R1 = 1.0\left(\frac{77}{77}\right)$$

Time point (min)	**Number of genes (peak)**	**Number of genes (log2-FC$_0$ > 1.5)**	**Number of genes (log2-FC$_{i-1}$ > 1.5)**
5	8	57	34
10	24	96	35
30	138	13	24
60	15	19	18
120	7	8	7
180	3	22	214
240	70	0	0

Table 1: Distributions of gene lists produced by different strategies

Since both R0 ∩ P0 and R1 ∩ P1 had very high values, our hypothesis that genes that reach maximal expression sooner are regulators and those that reach it later are regulated is supported.

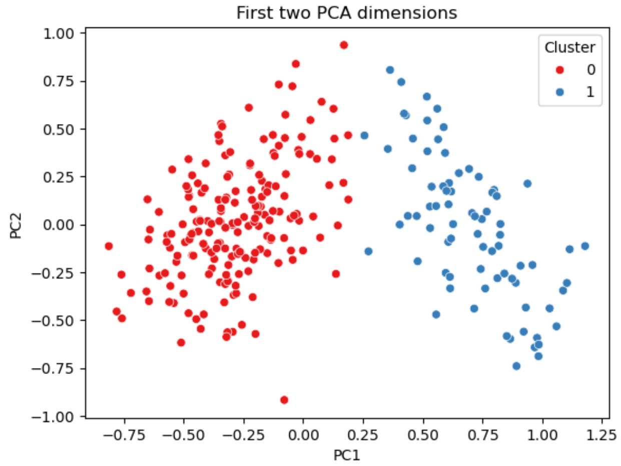

Figure 4: PCA plot of upregulated genes split into two clusters (P0 and P1)

4 Discussion

The results of our heatmaps demonstrate that there are expression peaks detected in distinct time intervals for up-regulated genes. However, the expression is much more spread out and show a less noticeable pattern when considering fold-change with respect to expression at previous time points. To predict regulatory pathways using the more basic approach, Pathfinder currently takes it as necessary to have discernible fold-change peak patterns. Therefore, we found that, for the development of Pathfinder, it would be more sensible to use fold-change values with respect to the initial time point's expression level (log2-FC$_0$).

There are several limitations to the current status of Pathfinder. One limitation is that there is a lack of a validation dataset with the hypoxic response. Although *SGD* contains data regarding some pathways, there is a lack of distinct genetic pathways as part of hypoxic response. According to *SGD* data, it seems that hypoxic response does not always follow a model of discernible genetic pathways, so it is challenging to confirm whether all pathways Pathfinder detects are true. Due to this lack of data, it is possible to reduce gene-lists to potential regulators and regulated genes, but it is challenging to infer interactions with high accuracy without more advanced methods.

It is necessary to obtain additional data regarding yeast gene expression, as this data may be usable with more advanced methods to construct regulatory networks. Previous studies have demonstrated that useful data might be chromatin data, which can indicate which regions of chromatin are transcribed at given time points (Duren et al.).

Once this additional data is obtained for *S. cerevisiae* under hypoxic response, it may be possible to revisit the results demonstrated to more accurately infer expression pathways. One component of accuracy that can be improved with new data is the initial filtering criteria for genes. This work employed filtering from the 75th percentile standard deviation, but new data may help filter based on other more sophisticated criteria. Furthermore, with additional training data, it may be possible to train machine learning models to further improve the statistical validity of predictions.

5 Acknowledgements

We would like to thank Professor Michael Schatz, who guided our inquiry throughout this project through helpful feedback and provided us with the knowledge and skills we needed through the Computational Genomics: Applied Comparative Genomics course. We would also like to thank Alex Sweeten, who was the TA for the course and also served as a guide through this project.

References

Bendjilali, Nasrine, Samuel MacLeon, Gurmannat Kalra, Stephen D. Willis, A. K. M. Nawshad Hossian, Erica Avery, Olivia Wojtowicz, and Mark J. Hickman. 2017. “Time-Course Analysis of Gene Expression During the Saccharomyces Cerevisiae Hypoxic Response.” G3 7 (1): 221–31.

Duren, Zhana, Xi Chen, Jingxue Xin, Yong Wang, and Wing Hung Wong. 2020. “Time Course Regulatory Analysis Based on Paired Expression and Chromatin Accessibility Data.” Genome Research 30 (4): 622–34.

Monti, Michele, Jonathan Fiorentino, Edoardo Milanetti, Giorgio Gosti, and Gian Gaetano Tartaglia. 2022. “Prediction of Time Series Gene Expression and Structural Analysis of Gene Regulatory Networks Using Recurrent Neural Networks.” Entropy 24 (2). https://doi.org/10.3390/e24020141.

Mutarelli, Margherita, Luigi Cicatiello, Lorenzo Ferraro, Olì Mv Grober, Maria Ravo, Angelo M. Facchiano, Claudia Angelini, and Alessandro Weisz. 2008. “Time-Course Analysis of Genome-Wide Gene Expression Data from Hormone-Responsive Human Breast Cancer Cells.” BMC Bioinformatics 9 Suppl 2 (Suppl 2): S12.

Wang, Zhong, Mark Gerstein, and Michael Snyder. 2009. “RNA-Seq: A Revolutionary Tool for Transcriptomics.” Nature Reviews. Genetics 10 (1): 57–63.

Zitomer, R. S., P. Carrico, and J. Deckert. 1997. “Regulation of Hypoxic Gene Expression in Yeast.” Kidney International 51 (2): 507–13.