Dimensionality Reduction in Modeling Transcriptome Dynamics

Thomas Wood [*], **Eli Shlizerman** [†]

[*]Wood Gesellschaft Applied Physics Laboratory, and [†]University of Washington

Submitted to Proceedings of the National Academy of Sciences of the United States of America

The number of genes in the genome of even a simple organism such as the Baker's yeast make modeling genome expression dynamics a non-trivial problem in Systems Biology. Dimensionality Reduction through Singular Value Decomposition (SVD) can help reduce the number of relevant features of genome expression and therefore aid in the fitting of macroscale dynamical models to genomic expression data. Due to the linear nature of the SVD, it is possible to extrapolate a microscale model by using a macroscale model of the genome expression dynamics to infer functional relationships between genes using time-series microarray data alone. We use our method to seek information about which genes in S. cerevisiae are affected by a combination therapy of Phenelzine and Lithium which is known to affect genome-level patterns of gene expression.

Transcriptome Dynamics | Nonlinear Optimization | Singular Value Decomposition

Abbreviations: SVD, Singular Value Decomposition; IDK, I Don't Know

Introduction

The large number of genes in the genome of the Baker's yeast make modeling the dynamics of the expression of the genome a difficult task to undertake[10]. Due to the high number of dimensions in the problem of finding a dynamical model to fit to genome expression data, even finding a linear relationship between genes in the form of a transition matrix becomes untenably difficult. Predicting the expression levels of all the genes in the genome of S. cerevisiae through the use of a transition matrix with roughly 25 million parameters can be accomplished through projecting the time series genome expression data into a space where the genes are clustered into orthogonal components that correspond to the frequency with which changes in the expression of the clusters occur.

We used genome expression data of the form

$$X = \begin{bmatrix} \vec{x}_1 & \vec{x}_2 & \cdots & \vec{x}_N \end{bmatrix}, \qquad \textbf{[1]}$$

where $\vec{x_i}$ refers to the gene expression for $M = 5619$ of the genes of S. cerevisiae which we were able to map from the Affymetrix Yeast 2.0 microarray identifiers to ENSEMBL gene ids for the i-th time series measurement.

Through applying the SVD, we were able to decompose our matrix X into the product of three matrices U, Σ, and V^*

$$X = U\Sigma V^* \qquad \textbf{[2]}$$

,which we can also write as

$$\begin{bmatrix} \vec{x}_1 & \cdots & \vec{x}_N \end{bmatrix} = \begin{bmatrix} \vec{u}_1 & \cdots & \vec{u}_M \end{bmatrix} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_N \\ 0 & 0 & 0 \\ \vdots & \vdots & \vdots \end{bmatrix} \begin{bmatrix} \vec{v}^1 \\ \vdots \\ \vec{v}^N \end{bmatrix}, \qquad \textbf{[3]}$$

where $\vec{u}_i$ is a $rank(M)$ basis column-vector which represents the topology of the i-th mode of the genome expression time series matrix and U is a M X M orthogonal matrix, Σ is a M X N matrix of singular values of the first N modes, and the N X N matrix V^* is composed of row-vectors $\vec{v}^i$ that represent the dynamics of the i-th mode in the original time-series matrix X.

From [**3**], we can determine that

$$\vec{x_t} = \sum_k \vec{u}_k \sigma_k v_t^k, \qquad \textbf{[4]}$$

where v_t^k represents the value of the k-th mode at time t. Taking the inner product of both sides of [**4**] with $\vec{u}_j$ and using the fact that $\langle \vec{u}_k, \vec{u}_j \rangle = 1$ if $j = k$ and 0 otherwise, we find

$$\langle \vec{u}_j, \vec{x}_t \rangle = \langle \vec{u}_j, \sum_k \vec{u}_k \sigma_k v_k^t \rangle \qquad \textbf{[5]}$$

$$= \sigma_j v_t^j \text{ thus,} \qquad \textbf{[6]}$$

$$v_t^j = \langle \vec{u}_j, \vec{x}_t \rangle / \sigma_j. \qquad \textbf{[7]}$$

To model the expression of the genome modes, we sought an optimal transition matrix T by using nonlinear conjugate gradient descent to minimize an $L2$ error function, thereby arriving at a difference equation

$$\vec{v}_{t+1} = T\vec{v}_t \text{ or,} \qquad \textbf{[8]}$$

$$v_{t+1}^i = \sum_j T_j^i v_t^j, \qquad \textbf{[9]}$$

where $\vec{v}_t$ represents a column vector of the expression of each mode being considered at time t. Because each dimension of the difference equation represents a single genome mode, the oscillations which arise from our difference equation are solely the result of couplings between the first-order systems which model each genome mode.

In order to find a microscale model for the genome expression dynamics, we needed to find a formula relating x_{t+1}^r, the expression of the r-th gene at time $t + 1$, to $\vec{x}_t$. This will allow us to find an M X M transition matrix W that can step forward in time the expression levels of the entire genome

$$\vec{x}_{t+1} = W\vec{x}_t. \qquad \textbf{[10]}$$

From considering only the r-th entry of the vector equation [**4**], we can write

$$x_{t+1}^r = \sum_i u_i^r \sigma_i v_{t+1}^i. \qquad \textbf{[11]}$$

Reserved for Publication Footnotes

Plugging [**9**] into [**11**], we can see

$$x^r_{t+1} = \sum_i u^r_i \sigma_i(\sum_j T^i_j v^j_t) \qquad [\mathbf{12}]$$

and by further substituting [**7**] into [**12**] we arrive at the expression

$$x^r_{t+1} = \sum_i u^r_i \sigma_i(\sum_j T^i_j \langle \vec{u}_j, \vec{x}_t \rangle / \sigma_j.) \qquad [\mathbf{13}]$$

On examining [**13**] we can see that $\vec{x}_t$ is the same for all i and j, thus it is now apparent that we may write the r-th row vector of the microscale transition matrix W from [**10**] as

$$\vec{w}^r = \sum_i u^r_i \sigma_i(\sum_j T^j_i \vec{u}_j / \sigma_j). \qquad \square \qquad [\mathbf{14}]$$

Results

Referencing Table **??**. Referencing Figure 2.

Simulations.

Simulation 1

I don't know why they wrote in Latin all of the text in the template.

Simulation 2

I am not sure if we are going to have a section named *Simulation*.

Real Data. The layout of the template is confusing.

Discussion

This is the section where we can give a good discussion about how the different methods of parameter estimation to fit the ODE to the genome modes worked (and didn't work) as well as plausible reasons why.

Materials and Methods

This section will contain a description of the software modules we used to.

Definition 1. *I'm going to leave this part alone for right now. A bounded function θ is a weak solution of QG if for any $\phi \,\epsilon\, C_0^\infty(\mathbb{R}/_{\mathbb{Z}} \times \mathbb{R} \times [0, \varepsilon])$ we have*

$$\int_{\mathbb{R}^+ \times \mathbb{R}/_{\mathbb{Z}} \times \mathbb{R}} \theta(x, y, t)\, \partial_t \phi\,(x, y, t) dy dx dt + \\ + \int_{\mathbb{R}^+ \times \mathbb{R}/_{\mathbb{Z}} \times \mathbb{R}} \theta\,(x, y, t) u(x, y, t) \cdot \nabla \phi\,(x, y, t) dy dx dt = 0 \qquad [\mathbf{15}]$$

where u is determined previously.

This is being left alone for the time being.

The steps taken for the mean-field method is going to be outlined above, even if it didn't work for the first-order model of each mode.

Appendix

An appendix without a title.

Appendix: Appendix title

An appendix with a title.

ACKNOWLEDGMENTS. This work was partially supported by a grant from KBRV Research and Development Division.

1. P. D'haeseleer, X. Wen, S. Fuhrman, R. Somogyi, Linear Modeling of mRNA Expression Levels During CNS Development and Injury , Pacific Symposium on Biocomputing, (1999), pp. 41–52.
2. F.X. Wu, W.J. Zhang, and A.J. Kusalik, Modeling Gene Expression from Microarray Expression Data with State-Space Equations, Pacific Symposium on Biocomputing, (2004), pp. 581–592.
3. O. Alter, P.O. Brown, and D. Botstein, Singular Value Decomposition for Genome-Wide Expression Data Processing and Modeling, Proc. of Nat. Acad. Sci., (2000), pp. 10101–10106.
4. C.M. Li and R.R. Klevecz, A Rapid Genome-Scale Response of the Transcriptional Oscillator to Perturbation Reveals a Period-Doubling Path to Phenotypic Change, Proc. of Nat. Acad. Sci. , 103 (2006), pp. 16254–16259.
5. J.N. Kutz, Data-driven Modeling and Scientific Computation: Methods for Complex Systems and Big Data, (2013).
6. U. Alon, An Introduction to Systems Biology: Design Principles of Biological Circuits, (2007).
7. Z. Bar-Joseph, Analyzing Time Series Gene Expression Data, Bioinformatics, 20.16 (2004), pp. 2493–2503
8. E. Shlizerman, K. Schroder, and J.N. Kutz, Neural Activity Measures and Their Dynamics, SIAM Jour. App. Math., 72.4 (2012), pp. 1260–1291.
9. C. Eckart and G. Young, The Approximation of One Matrix by Another of Lower Rank, Psychometrika, 1 (1936), pp. 211–218.
10. T. Chen, H.L. He, and G.M. Church, Modeling Gene Expression with Differential Equations, Pacific Symposium on Biocomputing, (1999), pp. 29–40.

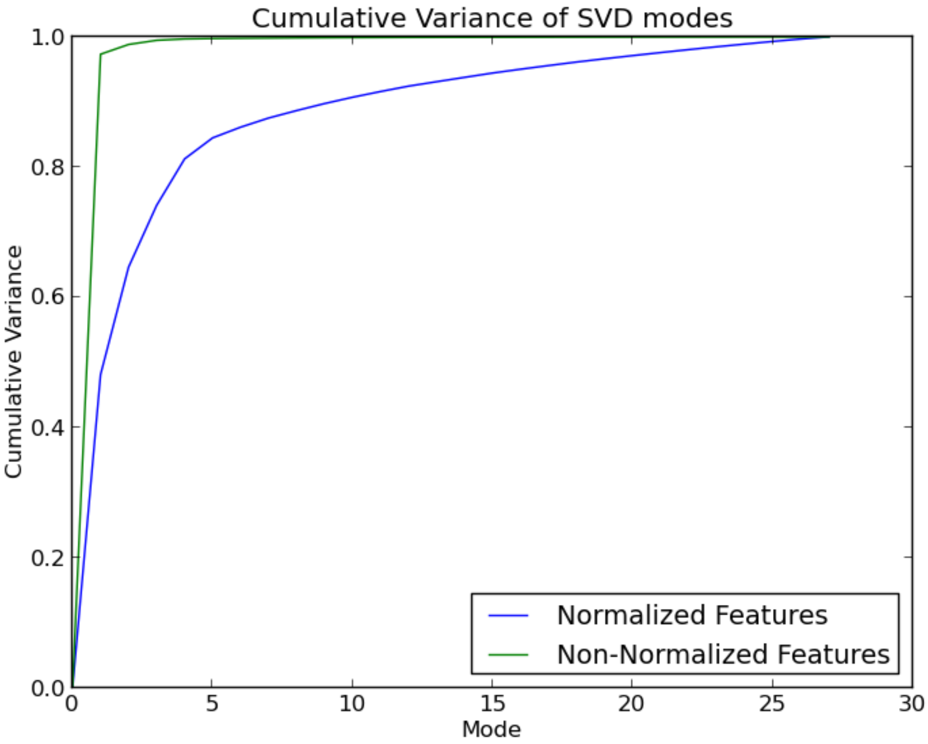

Fig. 1. This is the caption

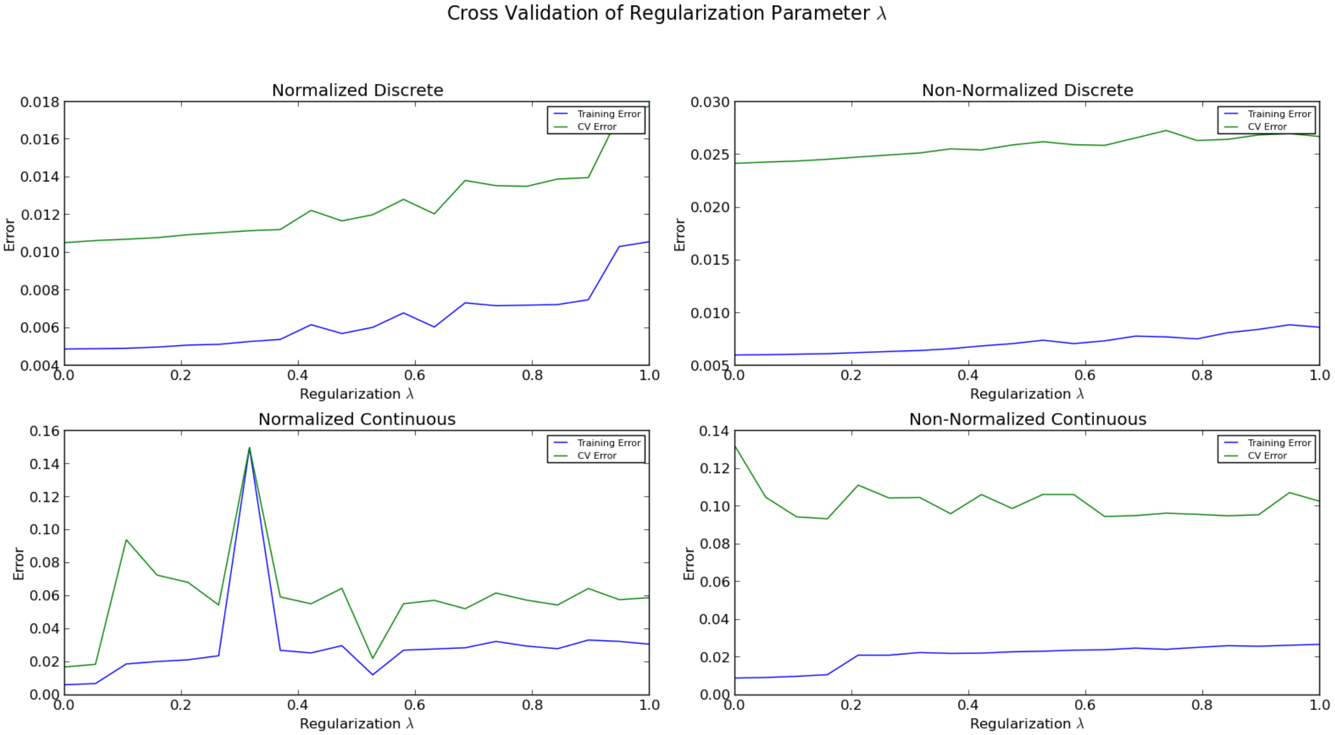

Fig. 2. This is the caption

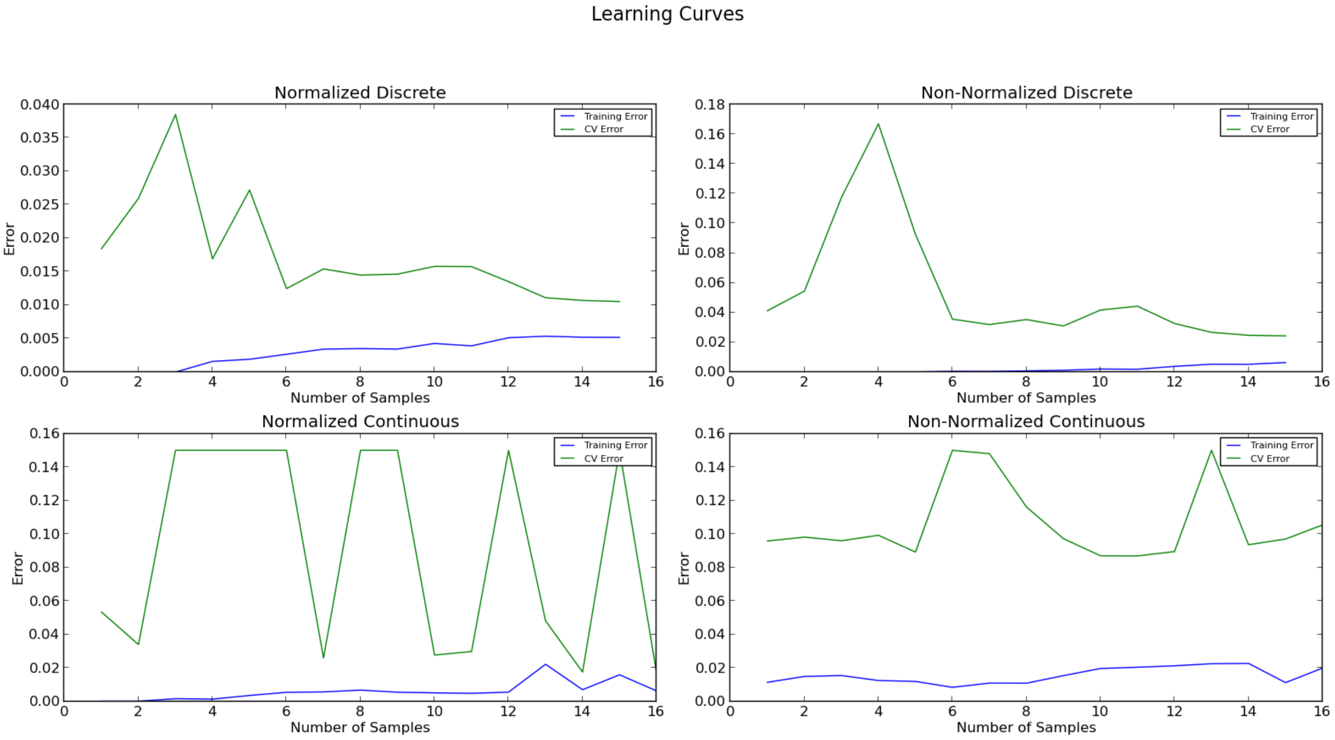

Fig. 3. This is the caption

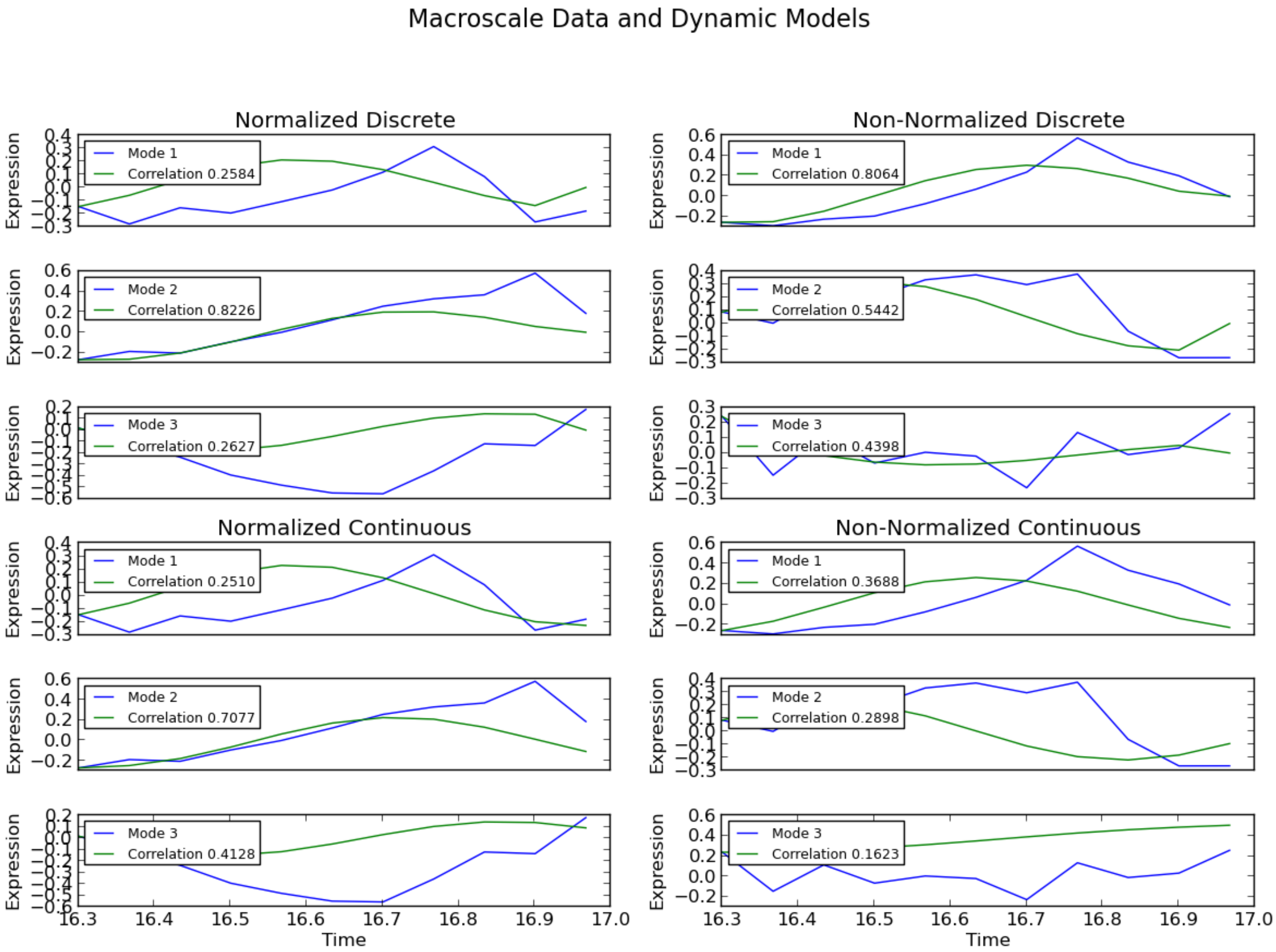

Fig. 4. This is the caption

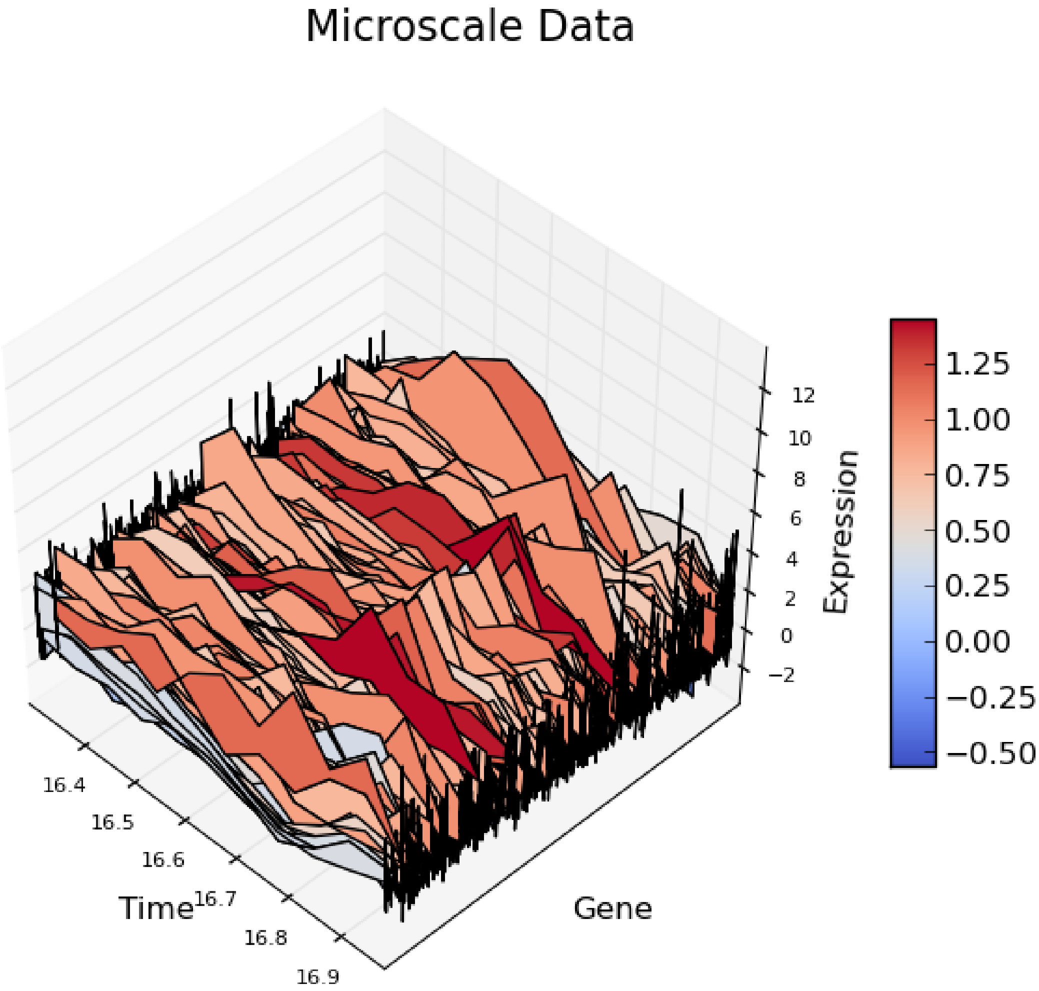

Fig. 5. This is the caption

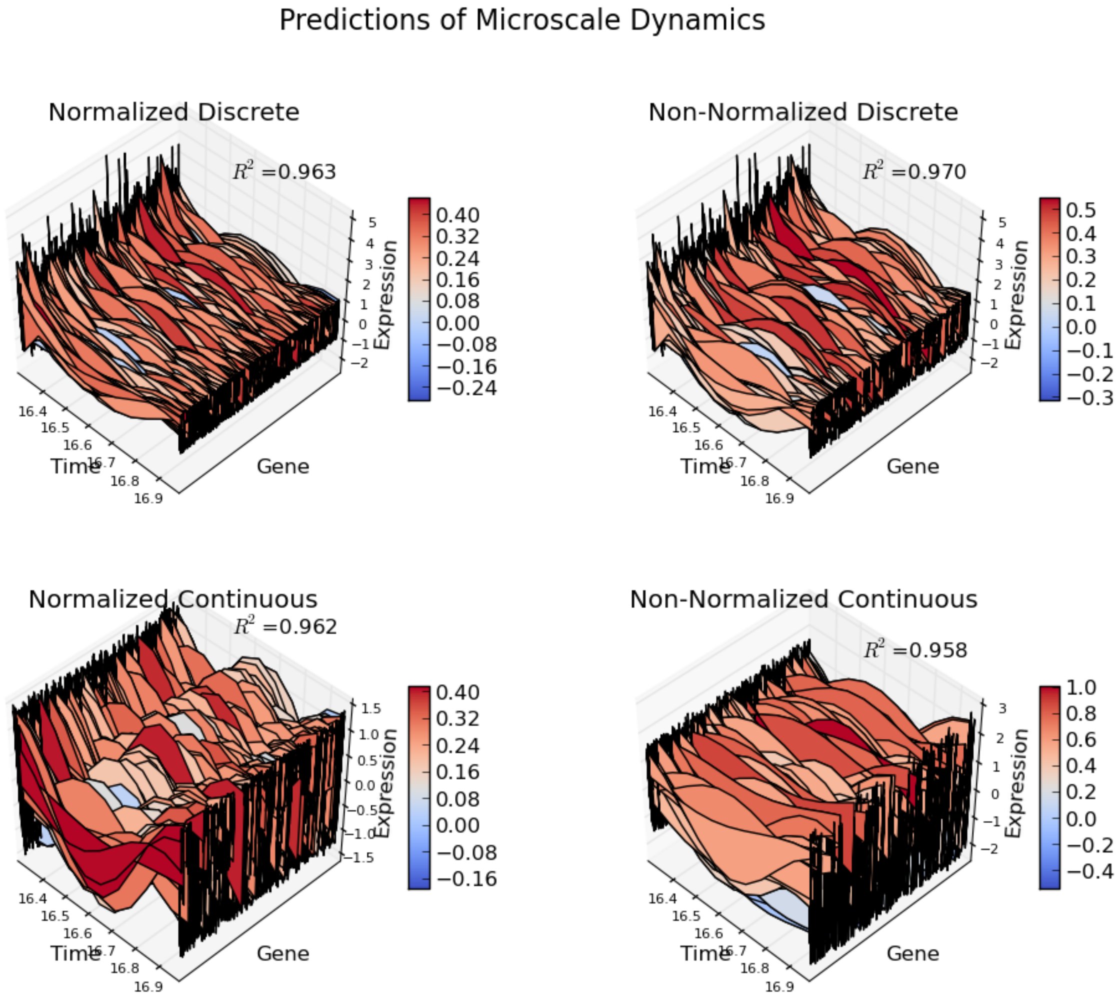

Fig. 6. This is the caption

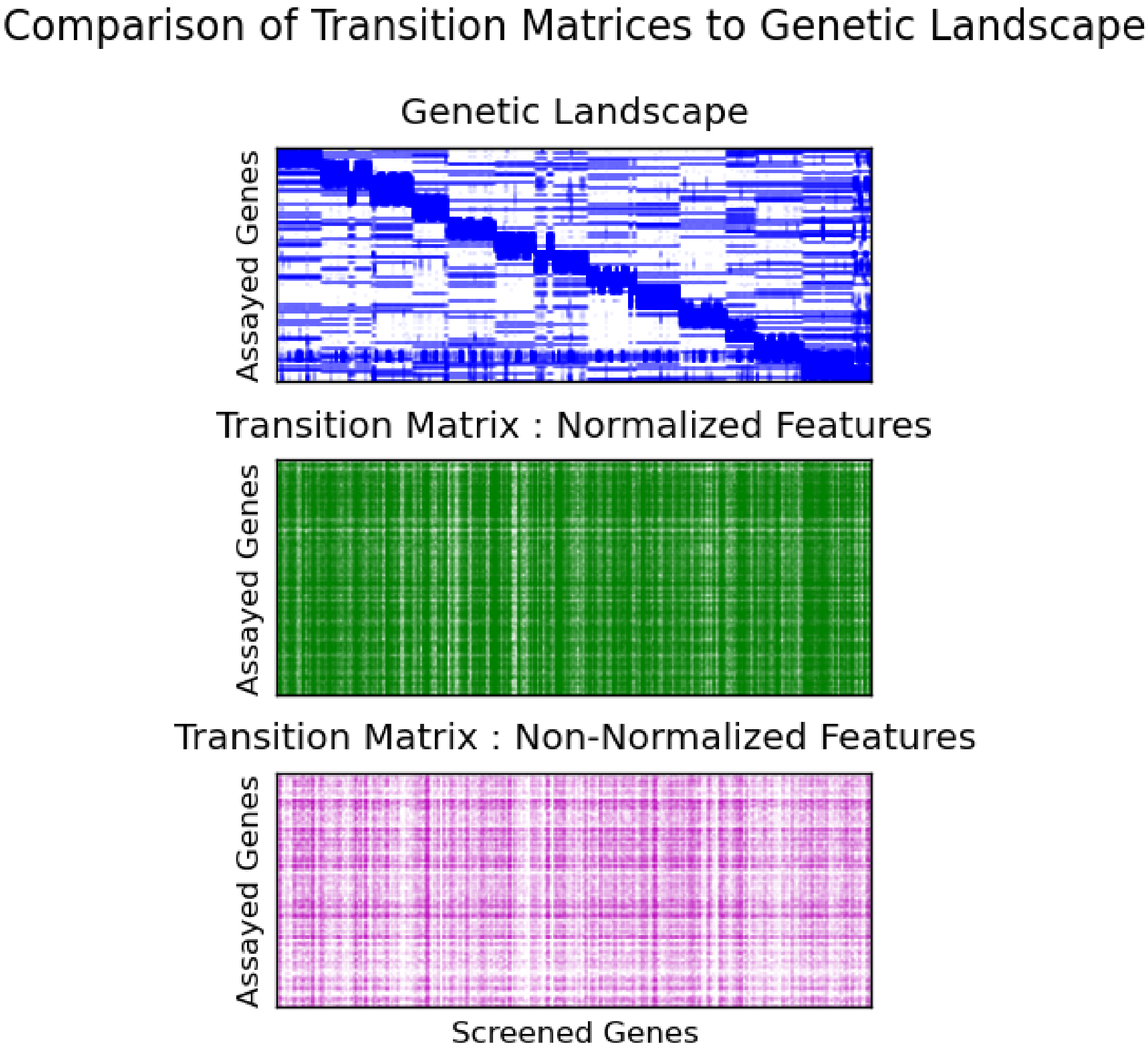

Fig. 7. This is the caption