# Using eigenvalue decomposition to extract emerging system’s properties from diverse biological datasets

PI: Christine Vogel (NYU Biology) – cvogel@nyu.edu

|  |
| --- |
|  |
| **Fig. 1.** The complexity of mammalian gene expression regulation |

## The multi-dimensionality of gene expression regulation

The seemingly simple task of producing a protein molecule from a gene is in fact highly complex and regulated at multiple, diverse levels which all act in a stochastic and highly dynamic manner in what we collectively call ‘gene expression regulation’ (**Fig. 1**). The yeast and human genome encode ~6,000 and ~20,000 genes, respectively, and today’s technology allows for the measurements of the genes’ mRNA and corresponding protein concentrations, providing data matrices in which the rows correspond to genes and the columns to different experiments and time points. A simple example of RNA and corresponding protein expression patterns is shown in **Fig. 2**. Other methods allow for measurements of the synthesis and degradation *rates* of these mRNAs and proteins, providing analogous data matrices. Again different methods allow for measurements of the concentrations of hundreds of *metabolites*, and we gain values for various experiments (columns) across different metabolites (rows). In sum, current technologies enable biologists to derive large, multi-dimensional data matrices for genes (mRNAs and proteins) and metabolites to respond to various experimental treatments or to evaluate changes over time in time course experiments.

Our next task is the integrative analysis of these data matrices to extract their emerging properties. Specifically, we aim to:

1. Integrate a variety of biological datasets that have their experimental setup in common (columns, e.g. time points), but vary in the identity and number of rows (e.g. genes or metabolites).
2. Use signal processing techniques to remove measurement noise from the data.
3. Extract patterns that are similar across biological replicates across the experiments.
4. Extract patterns that are common across different data types (e.g. mRNA, protein, or metabolites).
5. Extract patterns that are specific to individual data types.

|  |
| --- |
|  |
| **Fig. 2.** Thirty-hour time course experiment |

To perform these tasks, we will use methods of eigenvalue decomposition.

## Singular value decomposition

Singular value decomposition (SVD) is a mathematical technique which, similar to principal component analysis (PCA), decomposes a matrix into eigenvectors, eigenvalues, and the projections of the eigenvectors onto the original entries in the matrix. Alter et al. used SVD for the first time for biological data, and analyzed a yeast time course experiment in which mRNA concentrations where measured in cells undergoing cell division (cell cycle)[[1](#_ENREF_1)]. SVD was able to remove the measurement noise in the data (first eigenvector), and extract patterns of genes oscillating in their expression across the cell cycle. A generalized version of SVD (GSVD) was published later and compared cell cycle data across yeast and human cells, extracting patterns common and specific to the two organisms [[2](#_ENREF_2)].

A more recent extension of the technique, called higher-order generalized singular value decomposition (HOGSVD) [[3](#_ENREF_3)] was developed in 2012 by Ponnapalli et al., and allows for analysis of diverse datasets, such as those described above. HOGSVD allows us to extract essential patterns of expression regulation that would otherwise be hidden in this high-dimensional data [[3](#_ENREF_3), [4](#_ENREF_4)]**.** HOGSVD is a generalization of singular value decomposition to two or more datasets (matrices) and performs a simultaneous diagonalization of the datasets and can successfully identify patterns (eigenvectors) that are shared or specific across an arbitrary number of dimensions.

While HOGSVD has been mathematically fully developed [[3](#_ENREF_3)], its application to biological data is still awaiting. In collaboration with the method’s author, Dr. S. Ponnapalli, we aim to tackle this challenge and propose to apply HOGSVD to a variety of biological datasets. The specific and shared eigenvectors will quantify the contributions of different time-resolved regulatory patterns that are common or specific to mRNA and protein expression, or different types of stress and across replicates, and that can be explained by the action of regulators and also stochasticity. We will be able to simultaneously analyze global characteristics of the cellular response and gene-specific levels of regulation. We will identify patterns specific to a type of stress and a level of regulation (e.g. mRNA or protein). The projections of genes onto eigenvectors will allow us to map these gene-specific patterns in their combinatorial action of putative regulators that can then be tested in future experiments.

## Biological sets

Our lab investigates the cellular response to environmental stimuli, triggered by stressors such as chemicals, heat, or starvation. In contrast to traditional biological approaches, which mostly compare treated versus untreated control, we conduct time-series analyses – which allow us to gain a deeper understanding not only of the different levels of regulation, but also crosstalk between these levels and their differential timing. We have been and are still analyzing several datasets relevant to the research in our lab. All these datasets are amenable to analysis by HOGSVD.

1. **Human cancer cells responding to ER stress.** Chemicals such as tunicamycin elicit stress of the endoplasmatic reticulum (ER stress) which is characterized by a multitude of regulatory processes [[5-9](#_ENREF_5)]. This complexity is illustrated in **Fig. 2**, which shows for a subset of mammalian genes their time-resolved expression changes (columns), and highlights the vast discrepancies between mRNA and protein concentrations for the same gene (row). While it is known that ER stress causes extensive remodeling of the translation regulatory network [[10](#_ENREF_10)], and a highly specific protein degradation response [[11](#_ENREF_11)], the effect of these mechanisms on individual genes, their exact timing, and interplay, are completely unknown and still hidden in the data.Addressing this gap in our knowledge requires both the use of high-dimensional technologies to study the various aspects of gene expression regulation, and the development of sophisticated computational methods to integrate and process these data to reveal their hidden messages.
   1. Two biological replicates (1 and 3)
   2. Eight time points (columns)
   3. Two types of measurements (RNA and protein) (rows)

This dataset should serve as a starting point.

1. **Translation mutants of Candida albicans**.
   1. Three biological replicates
   2. Seven different knockout mutants (columns)
   3. Three types of measurements (RNA, protein, metabolites) (rows)
2. **Yeast responding to different treatments.** 
   1. One biological replicate
   2. Three treatments with 5-8 time points (columns) - NOTE: number of columns is NOT MATCHING across the treatments, hence the three treatments have to be analyzed separately!
   3. Two types of measurements each (RNA, protein) (rows)

Other datasets are currently being produced (e.g. of two types of treatments, three replicates, four time points, four measurement types)… Exciting times! ☺

## Workflow

The exact workflow should be subject to discussion amongst everyone involved, and it depends on the datasets.

1. Data post-processing: remove (or estimate) missing values, do basic normalization, check on gene identifiers being similar across datasets.
2. SVD or PCA on individual datasets: examine eigenvectors, match to biological functions/processes, examine for noise or biases in the data, remove ‘noise’ vectors.
3. HOGSVD across datasets: extract patterns (eigenvectors) common and specific to datasets.
4. Biological analysis: match patterns to biological functions.

# References

1. Alter, O., P.O. Brown, and D. Botstein, *Singular value decomposition for genome-wide expression data processing and modeling.* Proc Natl Acad Sci U S A, 2000. **97**(18): p. 10101-6.

2. Alter, O., P.O. Brown, and D. Botstein, *Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms.* Proc Natl Acad Sci U S A, 2003. **100**(6): p. 3351-6.

3. Ponnapalli, S.P., et al., *A higher-order generalized singular value decomposition for comparison of global mRNA expression from multiple organisms.* PLoS ONE, 2012. **6**(12): p. e28072.

4. Lee, C.H., et al., *GSVD comparison of patient-matched normal and tumor aCGH profiles reveals global copy-number alterations predicting glioblastoma multiforme survival.* PLoS ONE, 2012. **7**(1): p. e30098.

5. Bi, M., et al., *ER stress-regulated translation increases tolerance to extreme hypoxia and promotes tumor growth.* EMBO J, 2005. **24**(19): p. 3470-81.

6. Hetz, C., E. Chevet, and H.P. Harding, *Targeting the unfolded protein response in disease.* Nat Rev Drug Discov, 2013. **12**(9): p. 703-19.

7. Liu, B., Y. Han, and S.B. Qian, *Cotranslational response to proteotoxic stress by elongation pausing of ribosomes.* Mol Cell, 2013. **49**(3): p. 453-63.

8. Matus, S., et al., *ER Dysfunction and Protein Folding Stress in ALS.* Int J Cell Biol, 2013. **2013**: p. 674751.

9. Xu, C., B. Bailly-Maitre, and J.C. Reed, *Endoplasmic reticulum stress: cell life and death decisions.* J Clin Invest, 2005. **115**(10): p. 2656-64.

10. Ventoso, I., et al., *Extensive translatome remodeling during ER stress response in mammalian cells.* PLoS One, 2012. **7**(5): p. e35915.

11. Tsang, K.Y., et al., *In vivo cellular adaptation to ER stress: survival strategies with double-edged consequences.* J Cell Sci, 2010. **123**(Pt 13): p. 2145-54.