Iowa Initiative for Artificial Intelligence
Final Report

Project title:	Discovering the rules that govern the response of organisms to stress.
Principal Investigator:	Johnny Cruz-Corchado, Josep Comeron and Veena Prahlad
Prepared by (IIAI):	Avinash Mudireddy
Other investigators:	
Date:	07/11/2022

Were specific aims fulfilled:	Yes
Readiness for extramural proposal?	No
If yes … Planned submission date	
Funding agency	
Grant mechanism	
If no … Why not? What went wrong?	Could not define a mechanism for labeling the clusters with the given data

Brief summary of accomplished results:

Research report:

Aims (provided by PI):

Background: Organisms function despite wide fluctuations in their environment through the maintenance of homeostasis—a process by which the internal milieu is maintained close to constant, or within a specific range, to support key cellular functions. Cells maintain homeostasis whereby certain 'critical features' of the cell are kept close to constant by changing others, yet the identity of these 'critical features' and the mechanisms responsible for monitoring and controlling them are still poorly defined. However, this question is the subject of intense research because the development and progression of the devastating neurodegenerative diseases of aging such as Alzheimer's disease and Parkinson's disease are thought to be the outcomes of the lack of the cell's ability to monitor perturbations to homeostasis and restore equilibrium. The reduced homeostasis in these disease states is supported by studies that show that the abundance of different constituents of the cell (proteins, mRNA, etc.) becomes highly variable, are more susceptible to perturbation, and do not always return to baseline following the perturbation. Although manifestly challenging, a better understanding of how cellular homeostasis is achieved is critically needed.

The traditional workflow for analyzing how cells maintain homeostasis is to measure the changes in abundance of some key constituents of the cell (typically mRNA), in response to a severe perturbation (stress), use standard statistical methods to determine which of these changes are meaningful, and subsequently, using an a priori knowledge of the function of these constituents, predict which biological or cellular functions change upon the application of this stress. This approach leverages the fact that cellular constituents that have different functions are also present at different levels of abundance, and that concordant changes in this abundance are suggestive of functional associations between them. The datasets that are typically used for this analysis are m×n 'gene expression' matrices where 'm' corresponds to the abundance of some cellular constituent of m genes (in this case, numerical values that represent mRNA levels from ~20,000 genes) upon 'n' different states (time, perturbations, etc.).

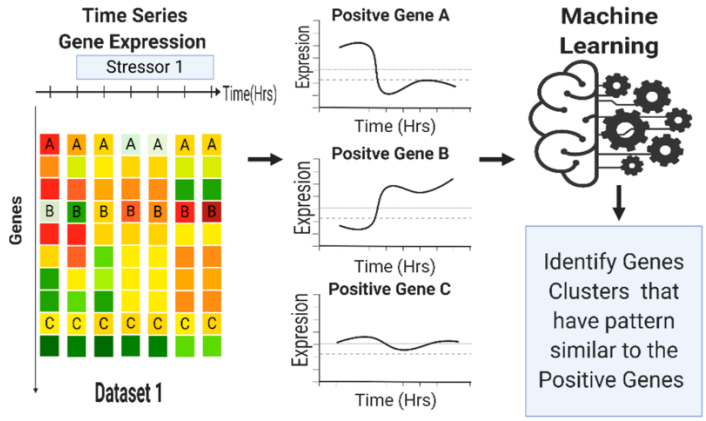
The hypothesis we aim to test here is that artificial intelligence (AI) methods applied to carefully identified datasets of gene expression changes (input variable: X) will be able to discover novel clusters of genes (output variable: Y) whose change is predictive of (a) biological functions that are altered or maintained homeostatic and (b) trade-offs between different biological tasks.

For these analyses, we will use 'gene expression' from a free-living nematode, Caenorhabditis elegans that is grown and maintained in the laboratory as a clonal population. Therefore, variability due to intrinsic differences in the identity of the animals is avoided.

Aim 1: Use gene expression data to identify clusters of genes that have similar expression patterns under stress.

We will use a detailed longitudinal (times series) dataset of gene expression, obtained from the Gene Expression Omnibus (GEO) [1, 2] to complete this aim. This time series [3] has the gene expression values of 20,000 genes measured from samples of C. elegans, immediately upon, and at different time points following, the application of a stressor (Figure 1). Among the genes in the dataset, we have identified a group of Positive genes (A-C in Figure 1), that we hypothesize based on their biological properties, will respond to the stressor, and will have a specific pattern of expression across the time course. We aim to identify other genes in our dataset that share the same temporal patterns of expression as our Positive genes and classify them into clusters according to their different temporal behavior following one stress condition, each with a distinct Stress Temporal Signal (output variable Y). The identity of the genes in these clusters will inform us as to the biological functions that are changing upon this perturbation

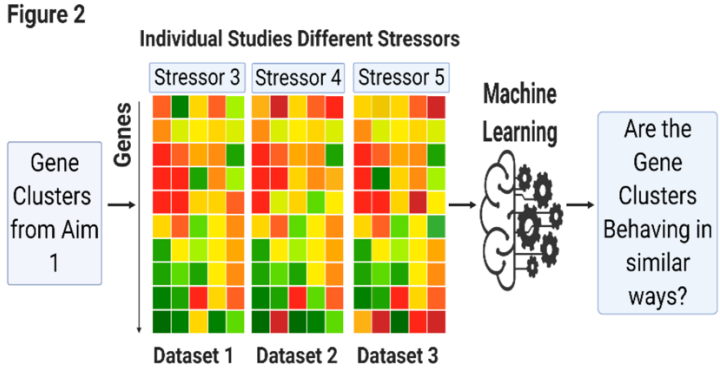
Figure 1



Aim 2: Determine if the Gene Clusters identified in Aim 1 behave similarly under different stress conditions.

To complete this aim, we will use additional gene expression datasets that were obtained from C. elegans treated with different stressors [1, 2]. In aim 2 we want to test how the gene clusters identified in Aim 1 behave under different stressors (Figure 2). These datasets, however, are not time series, but they include samples treated with different intensities of the stressors (at one time point after exposure that can be different for different datasets) and samples without the stressors (Controls). We want to develop a machine learning model that will test how the timing and the presence of the stress affect the behavior of the gene clusters. Our ML model will analyze how each stress affects the behavior of the previously characterized genes clusters and identify clusters consistently present in all the stress conditions (common stress signal) and clusters specific to certain type of stressors (specific stress signal).

Figure 2



Significance: The project will help to develop a model that can predict fundamental rules that determine how organisms and cells respond to stress. The current research capabilities will enable experimental confirmation of any novel clusters we identify. We argue that such an effort is **highly significant** as identifying features that are maintained homeostatic under stress can provide critical biomarkers for identifying the onset and progression of neurodegenerative disease.

Data:

Our preliminary datasets are from C. elegans but expression data pertaining to other metazoans was also identified [1, 4, 5]. We obtained 9 longitudinal studies and 1304 individual experiments that represent changes in the expression of each of the ~20,000 genes of C. elegans when animals are subjected to different stressors (heat, infection, starvation, reactive oxygen species etc.) [1, 2, 4]. We have downloaded, pre-processed our datasets [2, 6], and organized them in matrices (genes x samples) to proceed with the next steps in the analysis.

AI/ML Approach:

In deep learning SOMs (Self-organizing maps) were proposed in [7] and facilitate the automatic formation of topologically correct maps of features of observable elements. These topologically correct maps are analogous to clustering the elements of similar patterns. In [8], they were used to examine the distribution of p450 genes on the map for the rat hepatocytes data set [9] and srRna genes for the yeast gene expression data set [10]. In [11], SOMs were used to find clusters of similar behavior in 282 S. cerevisiae genes found in five-time series DNA microarray databases along with two cell cycles.

Self-organizing maps are unsupervised neural networks where the multi-dimensional input is represented in one-, or two-dimensional output grid of neurons, such that similar input data are mapped to neurons closer to each other on the grid than the dissimilar ones. The SOMs abstract the input information and represent it in a simple visual way on the grid. However, the main drawback of this technique is that there is no theoretical foundation to determine its parameters like grid shape, neighborhood radius, number of iterations, learning rate, etc., which yield the best result for a given dataset. Hence, we need to perform multiple experiments and come up with novel validation techniques to determine the appropriate parameters.

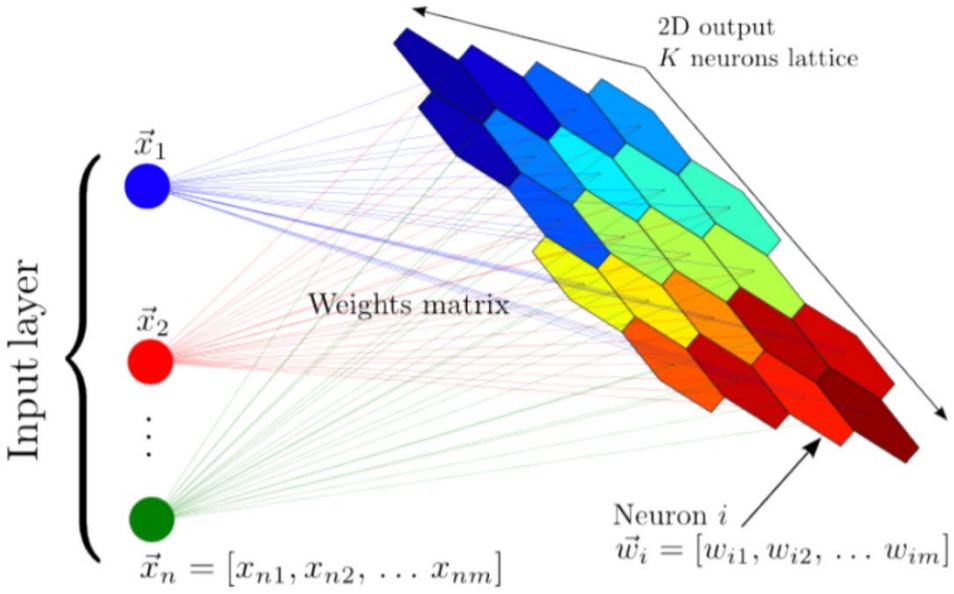


Figure 3. Self organizing maps [image source: https://www.latentview.com/blog/self-organizing-maps/]

In SOMs, all the features of each input sample form an input vector. Each input vector is connected to all the neurons on the output grid. Weights are assigned to each connection randomly. The output map weights are

gradually adjusted through a learning process to represent the input data as closely as possible. The learning process mainly contains three steps: the competitive process, the cooperative process, and the synaptic adaptation process.

In the competitive process, for each input vector in the dataset, only one neuron on the grid can win and be named as its Best Matching Unit (BMU). That is, the weight vector of this BMU is closest to the input vector than any other neuron on the grid. This similarity can be measured using different functions, such as the Euclidean distance, the inner product, or the Mahalanobis distance functions. The Euclidean distance is defined by:

$$min \left|\left|\overrightarrow{x} - \overrightarrow{w_j}\right|\right|$$

where $\overrightarrow{x}$ is the input vector and $\overrightarrow{w_j}$ is the weight vector corresponding to j-th neuron on the output grid. In each iteration, through this competitive process, BMU is calculated for a given input vector.

In the cooperative process, a neighborhood radius is chosen initially and the weights of all the non-winning units/neurons within the neighborhood of the BMU are adjusted proportionally to their proximity to the BMU. The neighborhood radius will be monotonically shrinking throughout the iterations so that weights of the neurons/units outside this radius are not updated or affected. Some neighborhood functions that can be used to measure the neighborhood radius are the Mexican hat, the rectangular, and the Gaussian functions.

In the synaptic adaptation process, to keep the neural network from overfitting, the weights are updated using a learning rate (η). This term also decreases throughout the iterations. Every unit on the output map adjusts its weight using the function

$$\overrightarrow{w_j}(n+1) = \overrightarrow{w_j}(n) + \eta(n) h_{j,i}(n)\left(\overrightarrow{x} - \overrightarrow{w_j}\right)$$

where n is the iteration number, $h_{j,i}$ is the topological neighborhood value of unit j centered around the winning unit i. The greater the proximity of neuron j to the winning neuron i, the higher the value of the neighborhood function $h_{j,i}(n)$, which results in a better adjustment on the weight of the neuron, as opposed to those that are farther away from the winning neuron. In this manner, the BMU becomes increasingly similar to the input vector that is being compared. The desired behavior is that neuron values on the output map are similar to the input data, and additionally that the input vectors are placed on the output map according to their similarity.

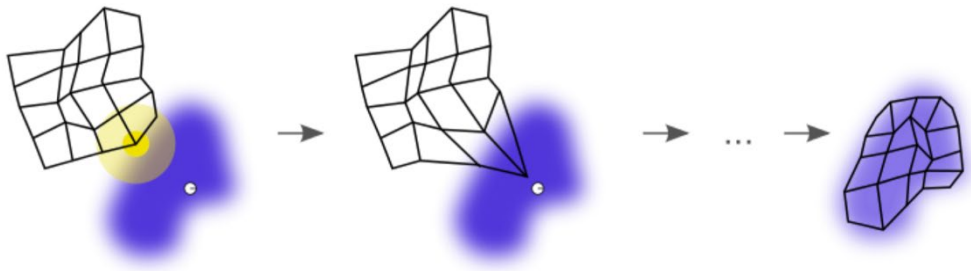


Figure 4. Synaptic adaption processes. [image source: https://www.latentview.com/blog/self-organizing-maps/]

After calculating the unit weight for all the input vectors, the learning rate η(n) and the neighborhood radius (σ) are decreased and the iteration number is increased.

Experimental methods, validation approach:

Data Size:

In this project we investigate the following stresses:

- Heatshock: This dataset contains gene expression values of 19105 genes with 3 populations measured at 9 time intervals [0, 30, 60, 120, 180, 240, 360, 480, 720] (in minutes).
- Aging: This dataset contains gene expression values of 20335 genes with 13 populations measured at 7 time intervals [4, 8, 12, 14, 16, 20, 24] (in days).
- Strarvation: This dataset contains gene expression values of 46739 genes with 3 populations measured at 7 time intervals [0, 1, 2, 3, 4, 6, 16] (in hours).

Model design:

To tackle the drawbacks of the SOMs as mentioned earlier, we developed an experimentation technique to improve the reliability in clustering results. During our experimentation, we found it difficult to decide the number of neurons to start with. This is controlled by the parameter grid shape, say (l, m) where l is the length of the grid and m is the breadth of the grid. Additionally, for each grid shape, it is challenging to determine the optimum parameters (learning rate, number of iterations, neighborhood radius).

To determine optimal parameters for a given grid shape, we adopted hyper-parameter optimization techniques [12] for SOMs. The initial learning rate and initial neighborhood radius are optimized by choosing from uniform distributions between particular ranges which is dependent on grid shape. The number of iterations is also dependent on the grid shape and the data size. However, each time we start from a different initial parameter choice, there is a difference in clustering results. In that case, we need a reliable way to select the best results. In our paper, we propose a new sampling technique and design a metric to select the best results. In this technique, for each input vector, we add 3 additional dummy vectors which are spacially very close to the input vector with a small difference δ. The expectation is that after clustering, each of these dummy vectors should fall in the same cluster that the corresponding input vector is assigned to. This way we designed a misclassification error which is defined by

$$error = \frac{\sum_{i=1}^{d} \#misclassified\ dummies}{d}$$

where d is the number of input vectors. We repeat the experiment 100 times, compute the misclassification error and choose the results of the least misclassification error as the best ones for a given grid shape.

However, selecting a grid shape is another challenge. In our experimentation, we observed that the number of clusters increases as the grid shape/number of neurons increases. This makes it difficult to choose a particular grid shape if we do not know how many clusters are present in the data beforehand. The simple way is to choose the desired number of clusters beforehand and select the grid shape which results in a number of clusters close to that desired number. However, this is not a robust strategy as we do not have control over the size of each cluster. There is a chance that we may have, say, 1 huge cluster and other clusters may contain only a small number of elements.

We hypothesize that there is congruency in the way clusters are divided as the grid shape increases. Say in a grid shape 4x4, we have 10 clusters and in 8x8, we have 20 clusters. Our hypothesis is that 2 of the clusters in the 8x8 grid are nothing but subsets of one of the clusters in the 4x4 grid. In that case, we can assign all the clusters which have a size less than a threshold to the nearest large cluster and check if the congruency is still followed across various grid shapes. This way we can ensure that small clusters do not exist. If this hypothesis is proved, we can then select the grid shape depending on the desired number of clusters.

We observed some congruency in our clusters (Figure 6). This gave us a conlusion that their lies a hierarchy among the clusters moving from smaller groups to higher groups. Hence, to improve our clusters, we modified our design to get bigger are more distributed clusters.

Modified Model design:

Once SOM outputs the cluster results. We defined an algorithm to join the smaller neighbouring clusters to grow them to a reasonable size.

Algorithm:

- Compute the centroids of each cluster.
- Find clusters with least size. Find the closest neighbour by measuring centroid to centroid distance.
- Merge the smaller cluster with the larger neighbour.
- Repeat until all the clusters are of a minimum size.
- Compute the k-means on the final joined gene groups using their centroids as the individual genes.
- Take the K-means labeling as the final labels and assigns the labels to each gene.
- Save the cluster results

Results:

Earlier design:

In our most relevant experiment, we observed the following results:

When we selected 3 grid shapes 4x4, 9x9 and 16x16. For each of the grid shapes, the hyper optimization parameters that we chose are

- Sigma/neighbourhood radius = uniform distribution(length of grid/10, length of grid//2.01)
- Learning rate = uniform distribution(0.001, 2),
- Max evaluations = min(5000,int(50*square root(# of samples)))
- Number of iterations = min(length of grid * breadth of grid *100, 5000)

After the hyper optimization, the best parameters turned out to be:
(4, 4) grid
best: {'learning_rate': 0.6648430043256534, 'sig': 1.096405348976604}
(9, 9) grid
best: {'learning_rate': 0.28535299288933247, 'sig': 2.8892781124638223}
(16, 16) grid
best: {'learning_rate': 1.8925373801471983, 'sig': 7.114064507077378}

For these optimal parameters, the number of clusters for 4x4, 9x9 and 16x16 turned out to be [16, 81, 256] respectively. Figure 5. shows the cluster distribution for all the genes.

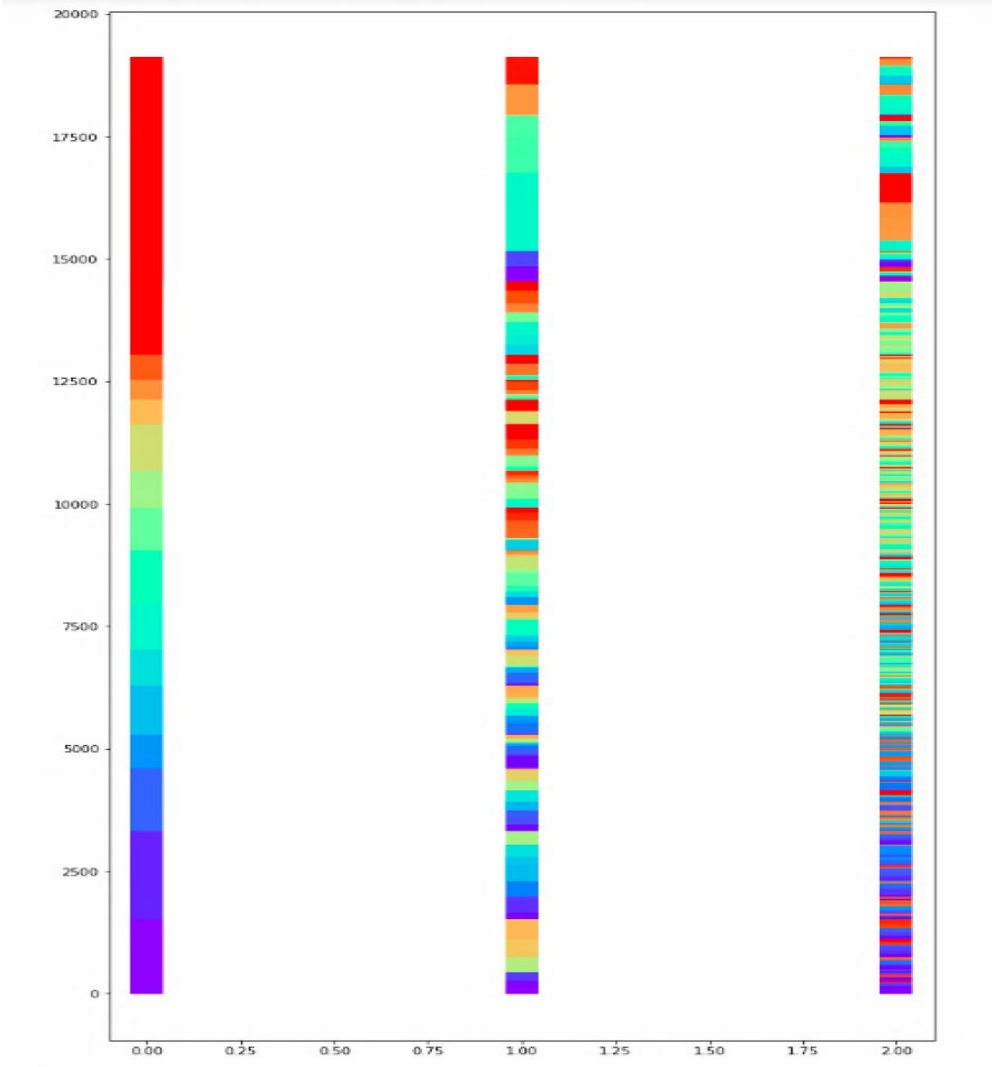


Figure 5. The cluster distribution for all the genes. X-axis represents the gene names, Y axis represents the 3 grid shapes 4x4, 9x9 and 16x16.

The color palette chosen here is to meaningfully represent each cluster. The starting reference cluster is marked in violet color. Then each of its spatially nearest clusters is marked with the nearest color in the VIBGYOR color spectrum. Hence, the nearest clusters spatially have the nearest colors in the spectrum. For the cluster-marking to follow the congruency, similar color pattern distributions should be followed for every grid shape.

However, one could see in Figure 5 that the distribution of the clusters does not seem to be congruent. Hence, we adapted the cluster refining technique mentioned in the methodology section to achieve the following result shown in Figure 6.

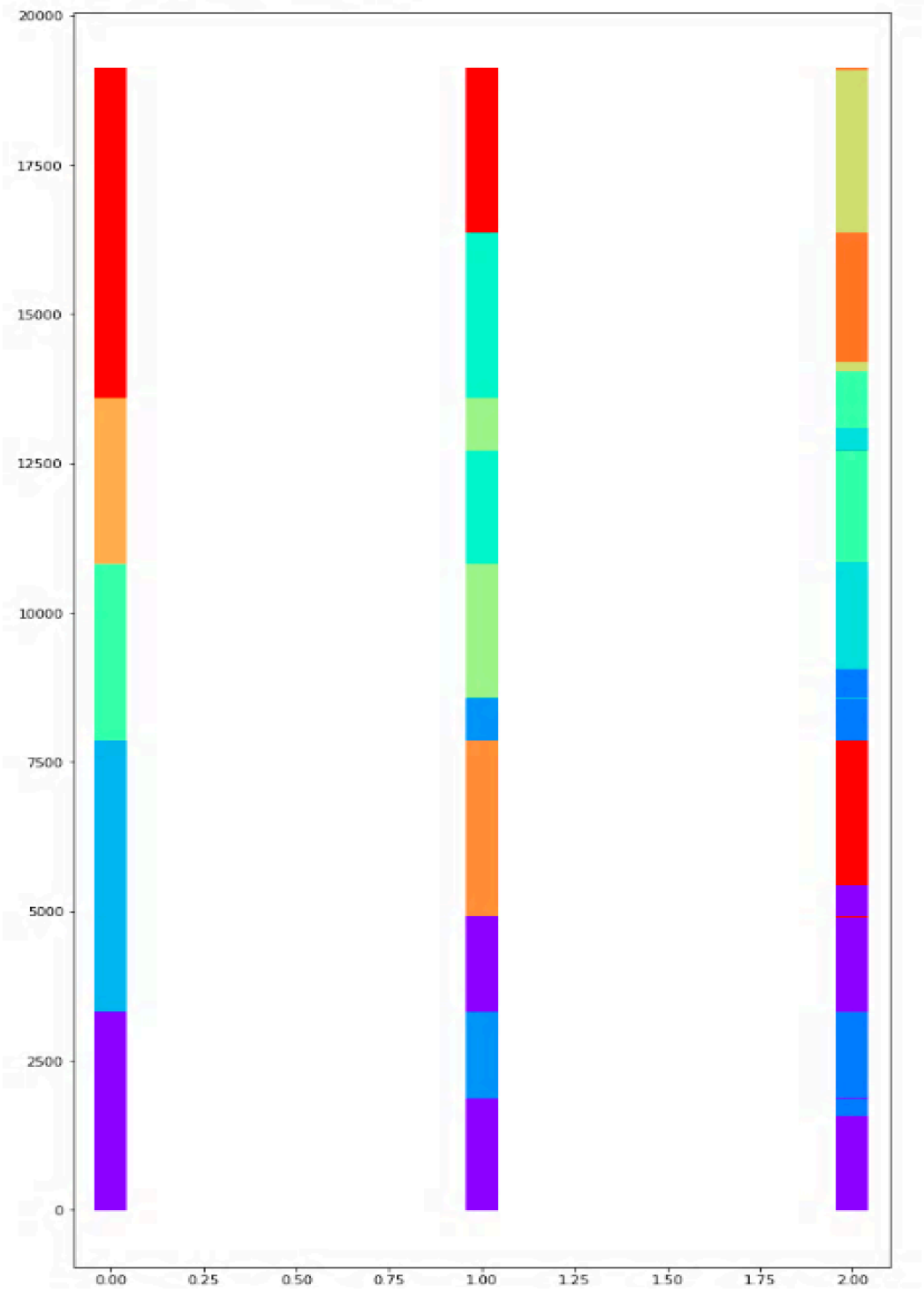
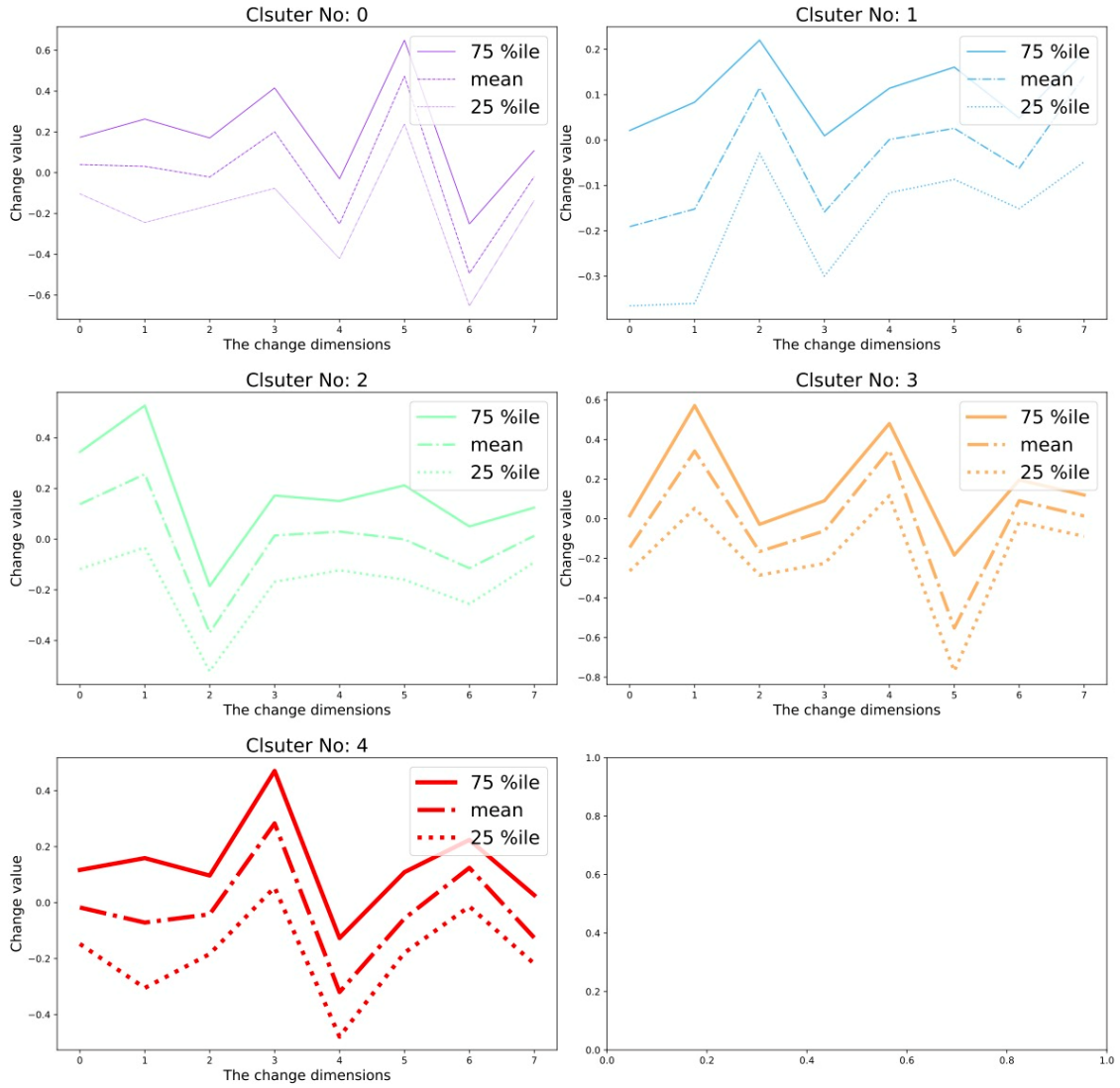


Figure 6. The cluster distribution for all the genes after hypothesis implementation. The X-axis represents the gene names, Y-axis represents the 3 grid shapes 4x4, 9x9, and 16x16.

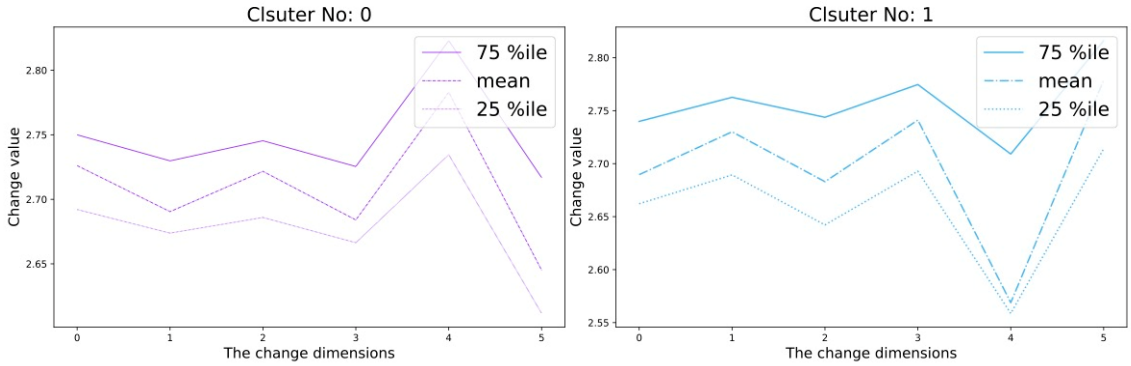
In Figure 6., one can observe that the color palette starts to look similar for all the grid shapes.

Modified Model design results:

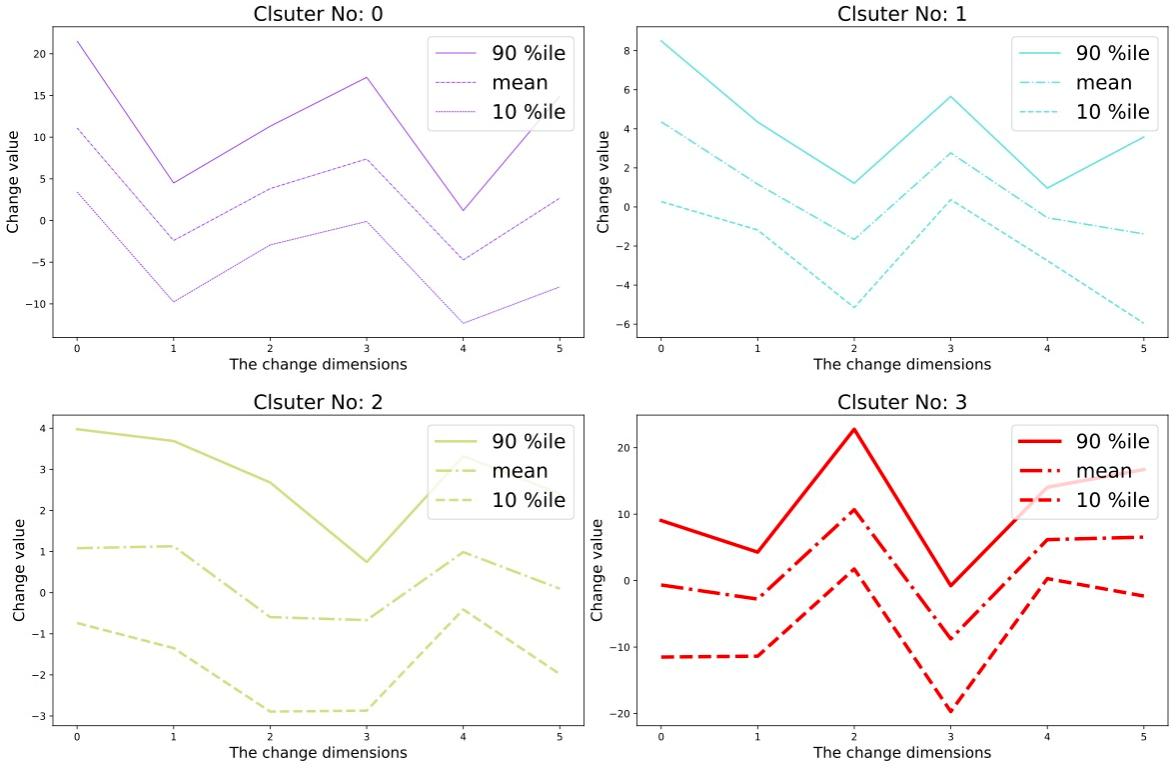
The cluster pattern on each dimensions for the Heat shock data is:



The cluster pattern on each dimensions for the Aging data is:



The cluster pattern on each dimensions for the Starvation data is:



The above diagrams illustrate the values of genes in each dimension in a given cluster. It is not feasible to represent every gene. Hence, we report the values at various percentiles.

Observations:
We observed that joining the genes from larger grid shapes gave more distributed clusters with better silhouette and Davis- Bouldin scores. However, due to computation limitations, we could not investigate more.

Ideas/aims for future extramural project:

V.P and J.C are currently funded (NIH, NSF) but not for this specific project.

Plans to receive extramural funding: If successful, the results will be used to identify genes of similar function as the clusters in human disease and aging databases, e.g., Human Protein atlas (https://www.proteinatlas.org) [7, 8], Allen Brain Map (https://portal.brain-map.org/) [7, 9] and Gene Expression Nervous System Atlas (http://www.gensat.org/) [10].

Publications: Combined, the results and methodology developed in this pilot project will be submitted for publication and serve as strong preliminary data for an NIH R01 or NIH R21 mechanism with the goal of identifying novel biomarkers in aging and disease.

References:

1. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M *et al*: **NCBI GEO: archive for functional genomics data sets—update**. *Nucleic Acids Research* 2012, **41**(D1):D991-D995.
2. Zhu Y, Davis S, Stephens R, Meltzer PS, Chen Y: **GEOmetadb: powerful alternative search engine for the Gene Expression Omnibus**. *Bioinformatics* 2008, **24**(23):2798-2800.

3. Harvald EB, Sprenger RR, Dall KB, Ejsing CS, Nielsen R, Mandrup S, Murillo AB, Larance M, Gartner A, Lamond AI *et al*: **Multi-omics Analyses of Starvation Responses Reveal a Central Role for Lipoprotein Metabolism in Acute Starvation Survival in C. elegans**. *Cell Systems* 2017, **5**(1):38-52.e34.
4. Ziemann M, Kaspi A, El-Osta A: **Digital expression explorer 2: a repository of uniformly processed RNA sequencing data**. *GigaScience* 2019, **8**(4).
5. Lachmann A, Torre D, Keenan AB, Jagodnik KM, Lee HJ, Wang L, Silverstein MC, Ma'ayan A: **Massive mining of publicly available RNA-seq data from human and mouse**. *Nature Communications* 2018, **9**(1):1366.
6. Zhu Y, Stephens RM, Meltzer PS, Davis SR: **SRAdb: query and use public next-generation sequencing data from within R**. *BMC Bioinformatics* 2013, **14**(1):19.
7. Kohonen, T: **Self-organized formation of topologically correct feature maps**, Biol. Cybern. 43, 59–69 (1982). https://doi.org/10.1007/BF00337288
8. Xiang Xiao, E. R. Dow, R. Eberhart, Z. B. Miled and R. J. Oppelt: **Gene Clustering Using Self-Organizing Maps and Particle Swarm Optimization**, Proceedings International Parallel and Distributed Processing Symposium 2003, pp. 10 pp.-, doi: 10.1109/IPDPS.2003.1213290.
9. Thomas K Baker, Mark A Carfagna, Hong Gao, Ernst R. Dow, Qingqin Li, George H. Searfoss, and Timothy P. Ryan: **Temporal Gene Expression Analysis of Monolayer Cultured Rat Hepatocytes**, Chem. Res. Toxicol. 2001, Vol 14, No. 9.
10. Cho, R. J., Campbell, M. J., Winzeler, E. A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T. G., Gabrielian, A. E., Landsman, D., Lockhart, D. J. and Davis, R. W.: **A Genome- wide Transcriptional Analysis of the Mitotic Cell Cycle**, Molecular Cell 1998, Vol. 2, pp. 65-73.
11. Chavez-Alvarez R, Chavoya A, Mendez-Vazquez A: **Discovery of Possible Gene Relationships through the Application of Self-Organizing Maps to DNA Microarray Databases**, PLoS ONE 2014, 9(4): e93233.
12. Bergstra, J., Yamins, D., Cox, D. D.: **Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures.** Proc. of the 30th International Conference on Machine Learning (ICML 2013), PMLR 28(1):115-123.
13. Sjöstedt E, Zhong W, Fagerberg L, Karlsson M, Mitsios N, Adori C, Oksvold P, Edfors F, Limiszewska A, Hikmet F *et al*: **An atlas of the protein-coding genes in the human, pig, and mouse brain**. *Science* 2020, **367**(6482):eaay5947.
14. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson Å, Kampf C, Sjöstedt E, Asplund A *et al*: **Tissue-based map of the human proteome**. *Science* 2015, **347**(6220):1260419.
15. Hawrylycz MJ, Lein ES, Guillozet-Bongaarts AL, Shen EH, Ng L, Miller JA, van de Lagemaat LN, Smith KA, Ebbert A, Riley ZL *et al*: **An anatomically comprehensive atlas of the adult human brain transcriptome**. *Nature* 2012, **489**(7416):391-399.
16. Heintz N: **Gene Expression Nervous System Atlas (GENSAT)**. *Nature Neuroscience* 2004, **7**(5):483-483.