Update

As per the proposal milestones, I've managed to load a subset of the data into D3 and draw the heatmap. To do so I had to complete the following subtasks.

1. Understanding how to download and parse the data from the GEO repository.
2. Subset the data into a manageable size and serialize it into a D3 consumable format (JSON in this case).
3. Read through several D3 tutorials to understand the framework.
4. Adapt an existing heatmap implementation to draw the microarray data.

I overestimated the pace at which I could carry out development due to my lack of familiarity with both Javascript and D3. Debugging takes me much longer than expected and most operations need to be looked up. Data transformations became much faster after the discovery of the underscore library in Javascript which provides a host of familiar functional tools like zip/filter/map/reduce etc.

I'm slightly worried about the loading the full dataset (~25 Mbytes), for the vis. It's possible that a pre-filter step or a sharding strategy will have to be implemented in order to keep the response time within reasonable bounds.

Overall my first milestone was hit on the nose and I will be forging ahead as planned.

Related Works

Other tools aimed at visualizing the expression profiles of cells are concerned with how the expression patterns are related. The tools are aimed at finding genes that are related to each other, both between cell types and within a single cell type. In a single cell type, the other tools are trying to surface relationships between genes. Through the noise, we'd like to identify whether they work together, in which case they will be expressed together, or inhibiting, in which case one will be highly expressed when the other is not. The other task is to find relationships between samples for the same gene. In this setting, the sample is examined under several treatments or exist in different disease states. The task then is to identify how the patterns of expression for a single gene changes across treatments or states.

ACTGes is different because it facilitates a different goal, the identification of genes that have a profile unique to a particular cell type. These so-called biomarkers would be identified as outliers when approached with the perspective described above. Concretely, the other tools would identify a gene with a characteristic partition of samples where one sample expresses the gene differentially from the rest. While it would be theoretically possible to use the other tools to carry out such a task, in practice the resolution of the other tools make the such identification infeasible. Additionally, ACTGes supports the production of a the

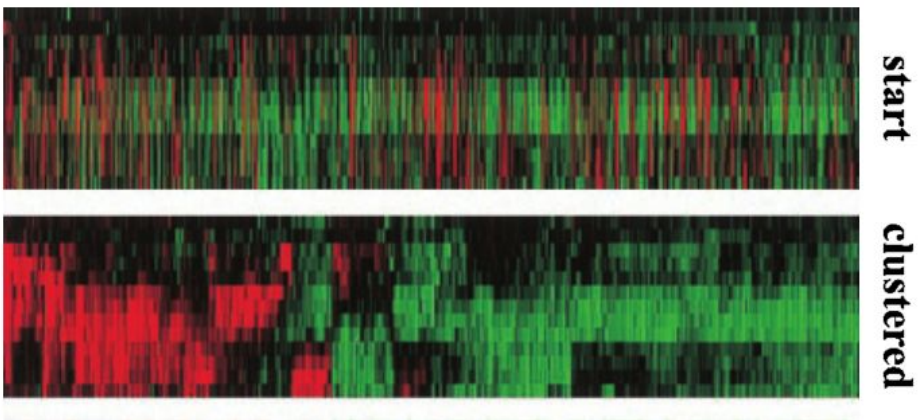
signature matrix that uniquely identifies each cell under inspection. Such an operation would require an external recording mechanism and doesn't allow the gestalt of the entire matrix to be examined.

We briefly summarize the main approaches to visualizing gene expressions below. We could not find prior visualization work aimed at identifying gene expression biomarkers.

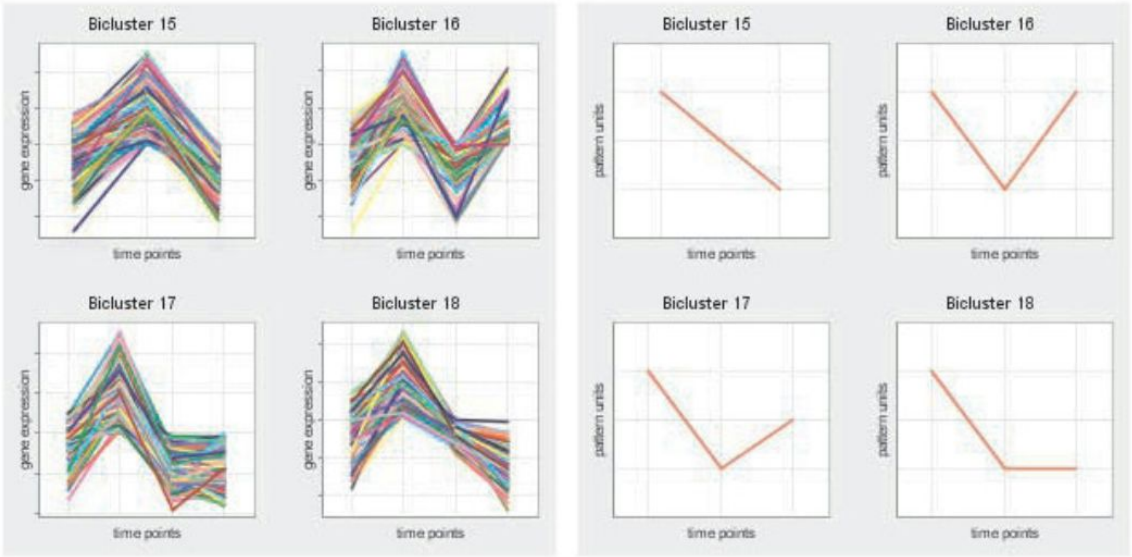
Clustering + Heatmaps

Clustering is a technique for grouping objects based upon a similarity score. In this case, we calculate the similarity based upon gene expression. Clustering has been shown to group together genes that are co-functional. Such clustering has been shown to cluster genes of similar functions together. Usually the cluster is encoded as a heatmap with a red-green diverging colour map, a staple of biological visualizations.

Eisen (1998), demonstrated a pairwise linkage clustering visualization that is adapted from phylogenetic tree reconstruction algorithms. A tree of similarity is built between the genes and then they are color coded to reflect their normalized expression levels.



Another method of clustering, called bi-clustering, attempts to cluster both the genes and experimental conditions simultaneously.
Gonçalves (2009), introduced a bi-clustering technique that could be used to understand gene expression time-series data. The output is a filtered block of genes whose expression patterns move in sync over time. This is all displayed in a SPLOM-like fashion, with each component being a line chart representing a particular cluster of genes.



(a)

Self Organizing Maps

Another technique that is used to understand expression profiles is the self organizing map. To create a self organizing map, one starts with a simple geometry like a rectangular grid. These points are then projected k-dimensional space occupied by the gene expression data and then shifted iteratively towards the data points. Each point moves towards the data point in proportion to the the distance away from the data point. After many thousands of iterations, the map will identify clusters of genes with similar expression patterns. Once again, this can be displayed in a SPLOM-like fashion to visualize the per-cluster expression patterns.

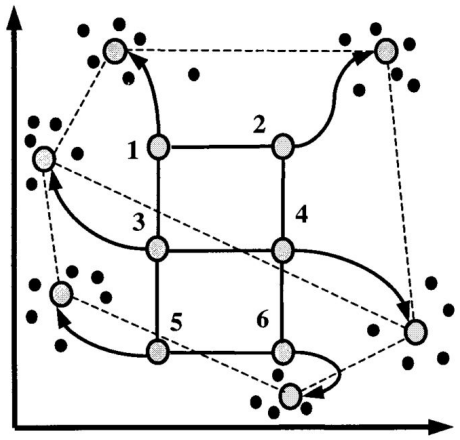
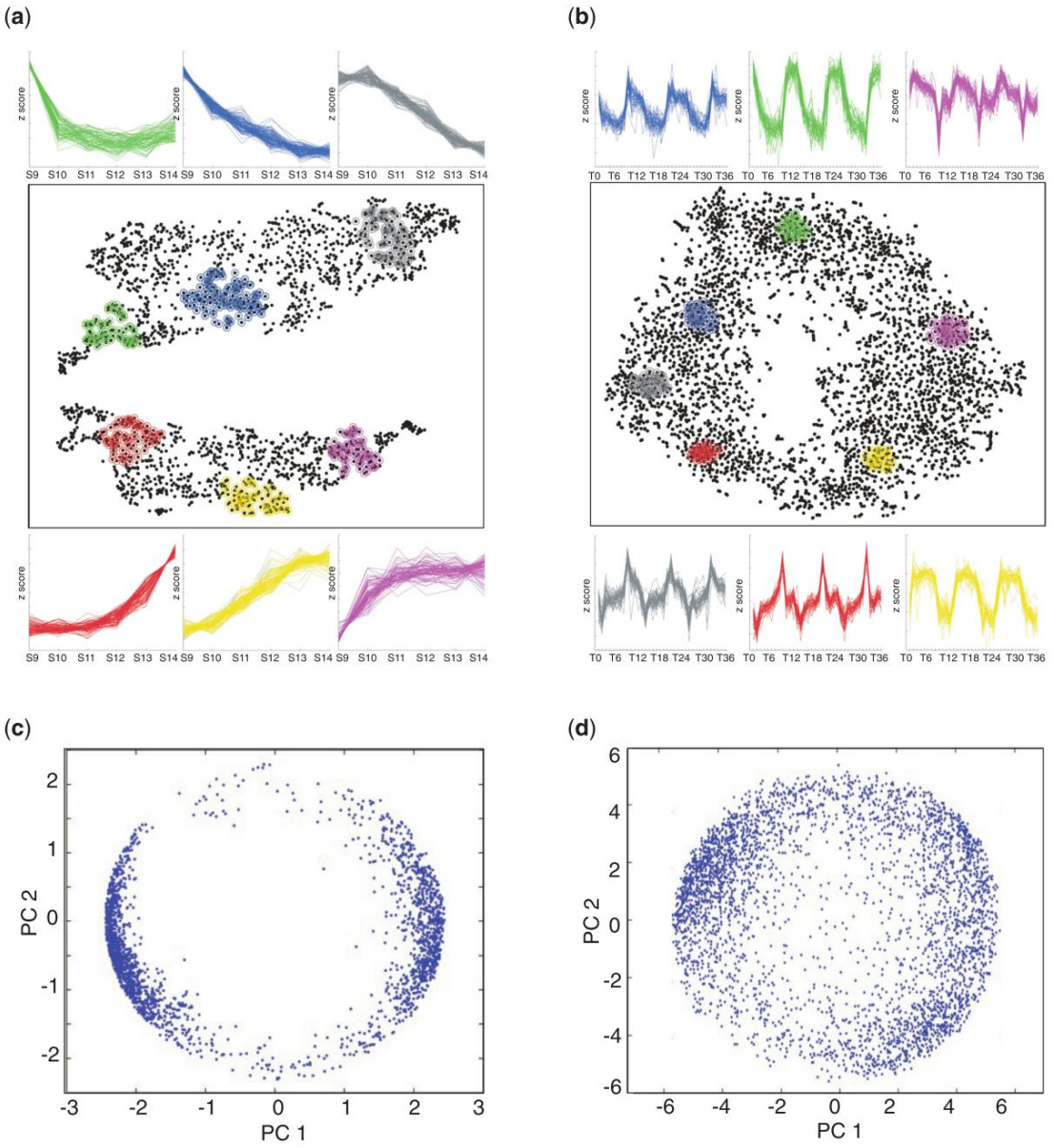


FIG. 1. Principle of SOMs. Initial geometry of nodes in 3×2 rectangular grid is indicated by solid lines connecting the nodes. Hypothetical trajectories of nodes as they migrate to fit data during successive iterations of SOM algorithm are shown. Data points are represented by black dots, six nodes of SOM by large circles, and trajectories by arrows.

Dimensionality Reduction

Since the expression of k genes can be interpreted as a vector in k-dimensional space, it is natural to attempt dimensionality reduction. Bushati (2011) demonstrates the use of t-SNE and PCA in projecting the k-dimensional space down to 2D that is amenable to a scatter plot. The data is grouped by treatment, in this case embryo development stage, and overlaid onto the graph. There are also embedded line charts showing the expression of particular gene subsets in different experimental conditions.



Citations

Abbas AR, Baldwin D, Ma Y, Ouyang W et al. Immune response in silico (IRIS): immune-specific genes identified from a compendium of microarray expression data. Genes Immun 2005 Jun;6(4):319-31.

Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. Proc. Natl Acad. Sci. USA, 95, 14863–14868

Tamayo,P., Slonim,D., Mesirov,J., Zhu,Q., Kitareewan,S., Dmitrovsky,E., Lander,E.S. and Golub,T.R. (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc. Natl Acad. Sci. USA, 96, 2907–2912

Bushati N, Smith J, Briscoe J, Watkins C. An intuitive graphical visualization technique for the interrogation of transcriptome data.Nucleic Acids Res 2011;39:7380–7389.
Gehlenborg,N., O'Donoghue,S.I., Baliga,N.S., Goesmann,A., Hibbs,M.A., Kitano,H., Kohlbacher,O., Neuweger,H., Schneider,R., Tenenbaum,D. et al. (2010) Visualization of omics data for systems biology. Nat. Methods, 7, S56–68.

Ostrand-Rosenberg S (2008) Immune surveillance: a balance between protumor and antitumor immunity. Curr Opin Genet Dev 18:11–18

Joana P. Gonçalves, Sara C. Madeira and Arlindo L. Oliveira, BiGGEsTS: integrated environment for biclustering analysis of time series gene expression data, BMC Research Notes 2009, 2:124.