# Analyzing Gene Expression Time Series
# CS 229 Fall 2011 Final Project

Marco Cusumano-Towner
Advised by Sofia Kyriazopoulou-Panagiotopoulou

December 16, 2011

## 1 Introduction

There are approximately twenty-thousand genes in humans, each which can be *expressed* at different levels at different times, under different conditions, and in different tissue types (in fact it is exactly differences in gene expression that determine tissue type). Broadly speaking, the expression levels of genes in a cell approximate the current state of the cell. Modern high-throughput techniques like CAGE ([7]) can comprehensively measure the expression levels for all genes at a given time point. The amount a gene is expressed can be measured by counting the number of mRNA transcripts associated with that gene (these are, very roughly speaking, copies of the gene waiting to be translated into proteins).

Most often gene expression analyses begin by measuring expression levels under two or more conditions (e.g. healthy or disease condition, or environmental or stress conditions), and aim to identify which genes are differentially expressed between the two and why (e.g. [1]). Gene expression *time series* instead measure the expression levels in a *single condition* over time, often after the introduction of a stimulus ([3]) or during a cellular process ([4]), and aim to discover the *dynamical response* of the cell to this stimulus, and possibly the underlying regulatory networks that govern this response ([2]).

In this project, we have subjected a particular gene expression time series dataset to a series of analyses, primarily involving clustering of the genes. Clustering can reveal groups of co-regulated genes (genes that are regulated by the same factors), or genes involved in the same functional pathway ([5]). Clustering of gene expression time series is a challenging task, given the notoriously noisy nature of gene expression measurements, and the often sparse and unevenly-spaced time points ([6]).

## 2 Methods and Results

### 2.1 Preprocessing

We considered two datasets for this project: a time course of human leukemia cells after introduction of hemin, and vascular endothelial cells (cells that line the blood vessels) after application of VEGFC (vascular endothelial growth factor C). Both datasets included three replicates of the same experiment. We analyzed the statistics of both datasets (such as correlations between replicates)

and determined that the endothelial cell time course likely had less noise than the leukemia time course, and we used this dataset for further analysis.

Because different genes can be expressed at different general amplitudes, gene expression analyses are usually concerned with the 'log fold changes' for each gene $g$. We use the log fold changes with respect to the first time point $(\log_2 y_t^{(g)} - \log_2 y_1^{(g)})$.

The transcriptional (gene-expression) response to any given stimulus generally involves a subset of all the genes. Therefore, an essential first step in analysis of whole-genome gene expression data is determining those genes that are likely part of the biological response and removing the rest before further analysis. To do this, we only included a gene if it reached below half or above twice its initial value in all replicates. We replaced all zero values with the value two before calculating log fold changes. After filtering, 1485 genes remained. We also discarded the last two time points because they exhibited sudden biologically uninterpretable changes in value, and further removed genes with the worst fits to the impulse model (described below). The final product was 1305 genes with 14 values, sampled at 0min, 15min, 30min, 45min, 1hr, 1hr20min, 1hr40min, 2hr, 2hr30min, 3hr, 3hr30min, 4hr, 5hr, and 6hr. We used the log fold change data from one replicate for our further analyses.

## 2.2 Validating algorithm output

We need to verify that the output of clustering algorithms is biologically meaningful, before using the output to guide hypotheses regarding *new* biological meaning. The genes within some clusters should be related to one another biologically. In whole-genome bioinformatics, biological relationships are often determined by annotations (e.g. gene X is involved in 'blood vessel development' or 'chromatin modification') that biologists have assigned to genes. We test if clusters are statistically 'enriched' for any biological functional annotations. In particular, a hypergeometric test (or one-tailed Fisher's exact test) is used to asses whether significantly more genes in a set are associated with any annotation than would be expected if the gene set was selected randomly. We use the online tool g:Profiler ([8]) for this analysis.

## 2.3 Choosing the number of clusters

We attempted to choose the number of clusters ($k$) by observing how the distortion function decreased with increasing $k$, by using hierarchical clustering, and by assessing the consistency of cluster assignments between replicates for various $k$. However, visually identifying $k$-values that struck a balance between low within-cluster variance and low between-cluster similarity proved more direct and intuitive.

## 2.4 K-means clustering

Our first attempt at clustering used plain k-means. We used the standard Euclidean metric as well as the Pearson correlation coefficient metric in which the distance between genes $i$ and $j$ is $d(y^{(i)}, y^{(j)}) = 1 - r(y^{(i)}, y^{(j)})$, and $r(y^{(i)}, y^{(j)})$ is the Pearson sample correlation coefficient. The correlation metric is a common choice for this data domain, because any fine-grained fluctuations in the data are likely masked by noise. Clusters from k-means with the correlation metric are shown on the left in Figure 1. The clusters returned from k-means were not significantly enriched for any relevant biological annotations specific to the stimulus condition. This is likely because k-means fit

to random patterns in the data, which are present because of the large number of genes and the small number of time points, coupled with the high level of noise ([10]). A sign that this is occuring in our results is that many of the clusters are distinguished by sudden changes in value during the middle or end of the time course, which are unlikely to be true biological events.

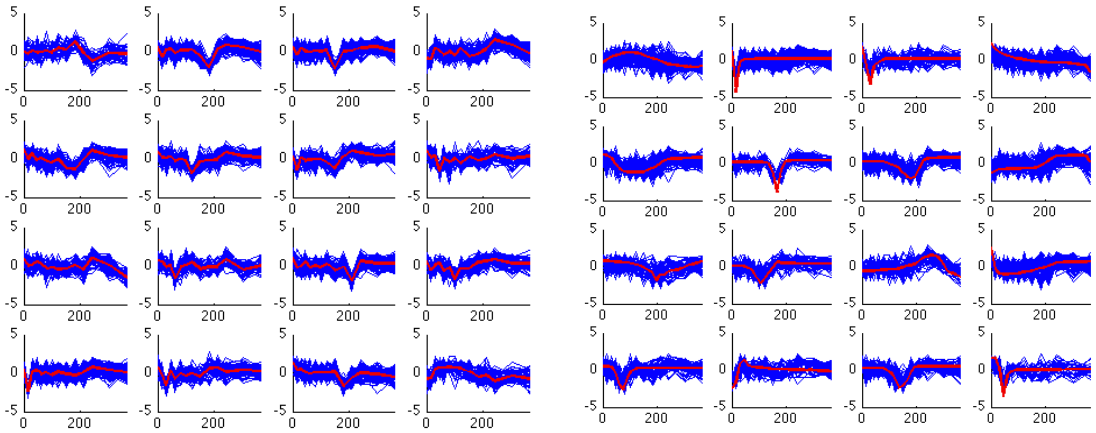

Figure 1: Clusters from k-means run on the raw data (left) and the impulse-fitted data (right). Centroid shown in red, individual gene time series shown in blue. The x-axis measures minutes after stimulus, and the y-axis measures (zero-mean unit-variance transformed) log-fold-change values.

## 2.5 Clustering using least-squares fits to a continuous model

To prevent fitting to random noise or systematic nonbiological fluctuations, we would like to restrict the clustering algorithm to only detect biologically meaningful patterns in the data. In addition, the basic k-means distance metrics described above are limited in that they ignore the temporal dependencies in the data, and treat each time point as an independent dimension.

To deal with the noise problem, prior work fits a parameterized continuous model to the data, which is then used as input for enrichment analysis or clustering algorithms. In particular, [3] introduce a model for gene expression time series that parameterizes a set of 'impulse' shaped curves with six biologically meaningful parameters (initial steady-state value, onset-time, peak-value, offset-time, final stead-state value, and a rate parameter). The authors claim this model captures the observed typical reaction of gene expression levels to many stimuli. We fit this model to our data by minimizing the sum-of-squares reconstruction error with respect to the parameters. We also fit a variant given in [10] that places more constraints on the parameters (e.g. offset-time must be greater than onset-time).

We ran k-means on these model fits, where we also sampled the continuous fitted functions more densly than the original time series. The resulting clusters and centroids are shown on the right in Figure 1. The only cluster that shows enrichment for biological functions related to our experimental condition (vascular endothelial cells, stimulated with a vascular endothelial growth factor), is the top-left cluster, which also corresponds to a cluster when fitted using the constrained impulse model. This cluster is enriched for 'blood vessel development', among a host of other less

specific terms. If this cluster is representative of the true signal in the data, the amount of within-cluster variance around the centroid illustrates the predominance of noise in the data. We did not see major improvements in enrichment analysis using the more constrained model.

### 2.6 Analyzing CDFs of fitted parameters

As a complementary analysis to testing the enrichment of clusters, we followed [10] and investigated whether genes with a certain annotation had any significantly different impulse model parameters from genes without the annotation (e.g. perhaps the 'onset-time' parameter is earlier than on average for genes involved with 'cell cycle'). For this we used the KS (Kolmogorov-Smirnov) test to compare the empirical distributions (empirical CDF) of the parameter for genes with and without the annotation. We did this for all annotations. We found several annotations related to our condition that were enriched for parameter distribution, although a number of unrelated terms were also enriched, so the significance of these is not yet clear. However, this analysis did lead to discovery of a set of nine genes related to a certain phase of the cell cycle that all have a very similar time profile (decrease rapidly in the first time sample, then rebound). These genes were located in the same cluster in our model fits, but their common annotations were not identified as statistically significant due to the large size of the cluster.

### 2.7 Integrated clustering and modeling algorithm

The final algorithm we evaluated ([10]) is a recent impulse-based iterative clustering algorithm for gene expression time series related to k-means, in which cluster centroids are represented as priors over the parameters of the impulse model described above. In the assignment step, the genes are fitted to the impulse model while regularizing the parameters with respect to each of the priors. The prior that results in the least total regularized cost for a gene becomes that gene's cluster. In the 'M' step, the priors are fit to the average of the genes in their respective cluster. This model aims to further prevent overfitting of the impulse fit to the data.

Due to the high computational demands for this algorithm, we were only able to run it for five iterations. However, the resulting clusters after five iterations are very similar to the clusters obtained by running k-means on the fits, as described above (again, there was one enriched cluster of genes that rise then fall, with many of the same genes). This suggests that overfitting was not a problem with the fit-then-cluster approach for our data. However, this algorithm also has a number of regularization parameters which could be further tweaked, and could potentially improve the results.

## 3 Discussion

Although the impulse model has been used successfully in prior works ([3], [10]), our results suggest that it might not appropriate for use with our data. Perhaps a subset of the gene expression profiles are indeed captured by the impulse model but others follow a multimodal pattern, perhaps related to the (periodic) cell cycle. In addition, the discovery of a small set of nine genes with highly related functional annotations and very similar time profiles that were not identified in enrichment analysis of clusters suggests that our cluster sizes may be too large. However, decreasing cluster sizes risks splitting related genes.

# 4 Future work

There are a number of other clustering algorithms for gene expression time series that are worth exploring. Several algorithms use iterative soft EM-like assignments. Of note is a method that uses mixtures of HMM's ([9]). As these authors point out, hard assignments of genes to clusters is not entirely biologically meaningful, because the interaction network between genes is not composed of separate connected components, and a mixture model may be more appropriate. This model also does not constrain the time profiles to have an impulse shape, but instead learns profile templates from the data, and can represent both impulse profiles and periodic profiles. Integrating a soft-assignment mixture model with enrichment analysis would be an interesting course of research that could potentially identify very small groups of related genes without over-partitioning the genes and breaking apart other related gene groups.

# References


[1] Oleg M Alekseev, Richard T Richardson, Oleg Alekseev, and Michael G O'Rand. Analysis of gene expression profiles in HeLa cells in response to overexpression or siRNA-mediated depletion of NASP. *Reproductive biology and endocrinology : RB&E*, 7:45, January 2009.

[2] Mukesh Bansal, Giusy Della Gatta, and Diego di Bernardo. Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. *Bioinformatics (Oxford, England)*, 22(7):815–22, April 2006.

[3] Gal Chechik and Daphne Koller. Timing of gene expression responses to environmental changes. *Journal of computational biology a journal of computational molecular cell biology*, 16(2):279–290, 2009.

[4] Raymond J. Cho, Michael J. Campbell, Elizabeth A. Winzeler, Lars Steinmetz, Andrew Conway, Lisa Wodicka, Tyra G. Wolfsberg, Andrei E. Gabrielian, David Landsman, David J. Lockhart, and Ronald W. Davis. A Genome-Wide Transcriptional Analysis of the Mitotic Cell Cycle. *Molecular Cell*, 2(1):65–73, July 1998.

[5] P. D'haeseleer, S. Liang, and R. Somogyi. Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics*, 16(8):707–726, August 2000.

[6] Jason Ernst, Gerard J Nau, and Ziv Bar-Joseph. Clustering short time series gene expression data. *Bioinformatics*, 21 Suppl 1(Suppl 1):i159–i168, 2005.

[7] Charles Plessy et al. Linking promoters to functional transcripts in small samples with nanoCAGE and CAGEscan. *Nature methods*, 7(7):528–34, July 2010.

[8] Jüri Reimand, Tambet Arak, and Jaak Vilo. g:Profiler–a web server for functional interpretation of gene lists (2011 update). *Nucleic acids research*, 39(suppl_2):W307–315, June 2011.

[9] A. Schliep. Using hidden Markov models to analyze gene expression time course data. *Bioinformatics*, 19(90001):255i–263, July 2003.

[10] Julia Sivriver, Naomi Habib, and Nir Friedman. An integrative clustering and modeling algorithm for dynamical gene expression data. *Bioinformatics*, 27(13):i392–i400, 2011.