Three Algorithms for Filtering and Analysis of Gene Expression Data.

Raphael Gottardo[†], Andreas Rechtsteiner[‡], Luis Rocha[‡], Michael E. Wall[‡], Tom Brettin[†]
[†]Bioscience division, Los Alamos National Laboratory
[‡]CCS-3, Los Alamos National Laboratory

Abstract

Analysis of gene expression microarray data is complex due to the usually large number of genes in one assay, few measurement points per gene and noisiness of the data. We introduce three algorithms, two developed by us, for gene expression analysis. Algorithm 1 and 2 attempt to identify genes with significant changes in expression during an experiment. The methods of the algorithms are quite different, though. Algorithm 1 looks at the correlation in a gene's expression profile whereas algorithm 2 projects the data into a 2-dimensional subspace of interest. We use Singular Value Decomposition (SVD) to identify such subspaces. Algorithm 3 can be used to find clusters of co-expressed genes among the gene expression profiles identified by Algorithms 1 and 2. Clustering gene expression profiles identified by Algorithms 2 and 3 helps us avoid clustering noise and gives us higher confidence on the quality of the clusters we retrieve. The algorithms were applied to the yeast cell cycle data of Cho *et al.*[2].
Contact: raph@lanl.gov, andreas@lanl.gov, rocha@lanl.gov

1. The three algorithms

Algorithm 1: Filtering genes with the Serial Correlation Test

To improve the quality of subsequent analyses of gene expression data, the data is often *filtered*, i.e. it is attempted to remove genes with expression profiles that seem mostly due to noise.
To filter genes based on fold-change is a common approach. We introduce the serial correlation test [4], a statistical test based on the correlation of the measurements of a gene's expression levels. The technique attempts to detect if observations in a time series have random fluctuations. We find it more appropriate than the fold-change for time series gene expression data, or GE data that can be ordered by other variables, *e.g.* the varying concentration of a chemical during an experiment. Such tests take into account measurements of a gene on all arrays, and their correlation, not just measurements on single arrays (as in some fold-change approaches). To our knowledge it is the first time the serial correlation test is applied to gene expression data.
For each gene expression profile a serial correlation coefficient is calculated (see Algorithm 1). Large coefficients indicate that the fluctuations of the expression profile are unlikely to be of random nature. If a gene's coefficient is below a critical value, the gene will be removed from the data set. Critical values for the coefficient are listed in tables, see for example [4] page 201.

Algorithm 1: *Serial Correlation test for filtering noisy genes.*

1: Let x_i be the measurement on a gene x at time t_i, $i = 1, .., p$.
2: Calculate the serial correlation coefficient with:

$$s_x = \frac{p}{p-1}\left\{\frac{\sum_{i=1}^{p-1}(x_i - \bar{x})(x_{i+1} - \bar{x})}{\sum_{i=1}^{p}(x_i - \bar{x})^2}\right\}. \quad (1)$$

3: If s_x is smaller than some critical value s_c, gene x is removed from the data set.

Algorithm 2: Filtering genes by projection into interesting subspaces

Algorithm 2 is related to algorithm 1 in that it also attempts to 'filter' genes, to find the ones with significant expression profiles of interest. It is different in that we project the data into a 2-dimensional subspace of interest. More specifically, we calculate the correlation coefficient of a gene's expression profile with two orthonormal directions, or patterns of interest, and plot the coefficients against each other. Genes that project to the outside of the plot will have high correlation in the subspace and therefore will be more likely to show some real expression change than the genes at the center of the plot. A crucial assumption we make is that there is more structure among the genes at the outside of the plot, the ones that likely show some real change in expression, and the genes at the center, which most likely project where they do due to noise in the expression profiles. Our algorithm searches for the boundary between these two regions, where the change from the nearly uniform distribution at the center to a more structured distribution at the outside of the plot is largest. Note algorithm 2 does not use any parameters that need to be known a priori or tuned. The boundary is derived strictly from the data. Should a data set have more noise, the boundary will be further away from the center, as our algorithm attempts to be conservative. We want to be confident about the genes at the outside, that they represent some real change in expression. Or, in statistical terms, we are more concerned with minimizing the false positive rate than the false negative rate.
Here we used singular value decomposition (SVD) [3] to identify interesting subspaces in the data. Any other set of orthonormal directions, or patterns, of interest could be chosen.

Algorithm 2: *Algorithm to find 'interesting' genes from correlation plot.*

1: Select two orthonormal directions and compute their correlation vectors $\boldsymbol{c_{v_1}}$ and $\boldsymbol{c_{v_2}}$
2: Chose an initial value r_0, then: $r \Leftarrow r_0$
3: **while** $r < 1$ **do**
4: Find the set of genes inside the circle with radius r: $I_r \Leftarrow \{i : c_{v_1}^{(i)^2} + c_{v_2}^{(i)^2} \leq r^2\}$
5: Compute the one-dimensional density, $\hat{f}_r$, of the polar angles of the genes in I_r
6: Compute the value of $g(r) = \text{median}_j\{|\hat{f}_r(x_j) - \frac{1}{2\pi}|\}$, which is a measure of the deviation of the density $\hat{f}_r$ from the uniform density $\frac{1}{2\pi}$, over the support of the polar angles $[-\pi, \pi]$.
7: Assign a new value to r for the next iteration: $r \Leftarrow r + h$
8: **end while**
9: Find the boundary $\tilde{r}$ that maximizes the rate of change of g, *i.e.* that maximizes g'

Algorithm 3: Detection of clusters of co-expressed genes

Algorithm 3 attempts to detect clusters of co-expressed genes in the set of genes returned from algorithm 2, the genes with high correlation with the 2-dimensional subspace. Because the genes have been 'filtered' by algorithm 2, our confidence about the quality of the clusters can be high. Algorithm 3 searches for clusters in the 2-dimensional subspace used by algorithm 2.

Algorithm 3: *Find clusters of co-expressed genes*

1: Apply Algorithm 2: Select two orthonormal directions and compute the value of the boundary with radius $\tilde{r}$.
2: Find the set of genes outside the boundary with radius $\tilde{r}$: $I_{\tilde{r}} \Leftarrow \{i : c_{v_1}^{(i)^2} + c_{v_2}^{(i)^2} \geq \tilde{r}^2\}$
3: Compute the density, $\hat{f}_{\tilde{r}}$, of the polar angles for the genes in $I_{\tilde{r}}$
4: $S_c \Leftarrow \{j : x_j$ is a local maximum for $\hat{f}$ and $\hat{f}(x_j) > 1/2\pi\}$
5: Let $s_{(j)}$ be the ordered values of S_c (x-values at the peaks)
6: $nb_c \Leftarrow \text{card}(S_c)$(Number of peaks)
7: $h_1 \Leftarrow 1/2\pi$ (Start with the uniform density as minimum height)
8: **for** $j = 1$ to nb_c **do**
9: $lwr \Leftarrow \min\{m : m < s_{(j)}, \hat{f}(m) > k_j$ and $\hat{f}(m) < \hat{f}(m+)\}$ (Left boundary)
10: $upr \Leftarrow \max\{m : m > s_{(j)}, \hat{f}(m) > k_j$ and $\hat{f}(m) > \hat{f}(m+)\}$ (Right boundary)
11: $cluster_j \Leftarrow \{k : lwr < \theta_k < upr$, where θ_k is the polar angle for gene k $\}$ (Genes in the cluster)
12: $h_j \Leftarrow 1/(2\pi - upr + lwr)$ (Increase the minimum height)
13: **end for**

2. Application of the algorithms to cell-cycle data

We applied all three algorithms to the *S. cerevisiae* cell-cycle data [2]. 6220 genes were monitored every 10 mins for 17 time points, covering nearly two cell cycles.
We removed 3000 genes that seemed to have quite random expression profiles with algorithm 1. We applied SVD to the remaining gene expression data. Figure 1 shows the profiles of the first three eigengenes (a notation introduced by Alter *et al.*[1] for the patterns of the right-singular vectors in gene expression analysis with SVD) and the singular values. The second and third eigengenes show periodic patterns very close to a sine function. Those two eigengenes were used in algorithm 2 and 3 to detect the cell cycle related genes.

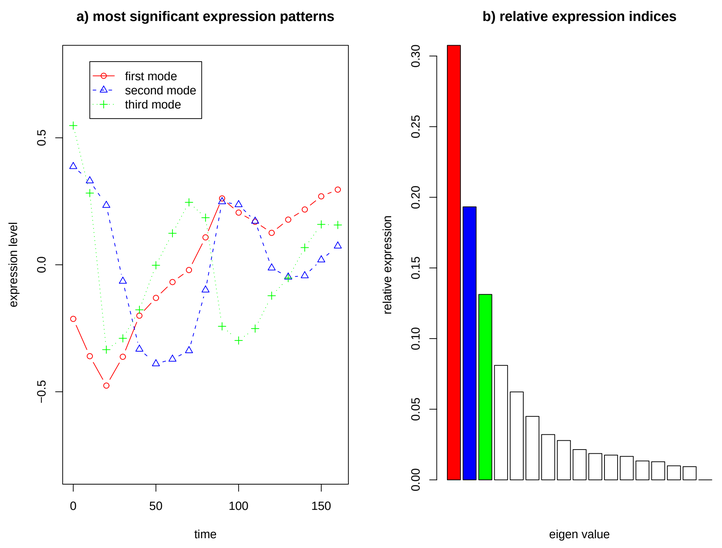

Figure 1: *a) The first three eigengene profiles. b) Relative expression level for the SVD eigengenes.*

Figure 2 shows the boundary algorithm 2 selected and the three high density regions of co-expressed genes that were selected by algorithm 3. Figure 3 shows the expression profiles of the genes in these 3 clusters. The periodic expression profiles of the genes in these clusters are evident. The different phases of the periodic patterns of the three clusters are evident as well.
The number of genes outside the circle with radius 0.67, *i.e.* potentially cell cycle regulated, is 895. Cho *et al.*[2] reported 416 cell-cycle regulated genes of which 231 agree with ours. However, they first filtered the genes by a fold-change approach and inspected the remaining 1300 genes visually. Among the 895 genes that were detected as potential cell cycle regulated by our method about 600 were removed by the fold change criteria used by Cho *et al.*. Most of these 600 genes showed clear periodic expression patterns. Two genes that stood out as particularly interesting were SWI6 and MBP1 which are known to be involved in the cell-cycle regulation and were detected as exhibiting clear periodic patterns by our methods but were removed by the fold-change approach of Cho *et al.*. Figure 4 shows the expression profiles of the two genes.

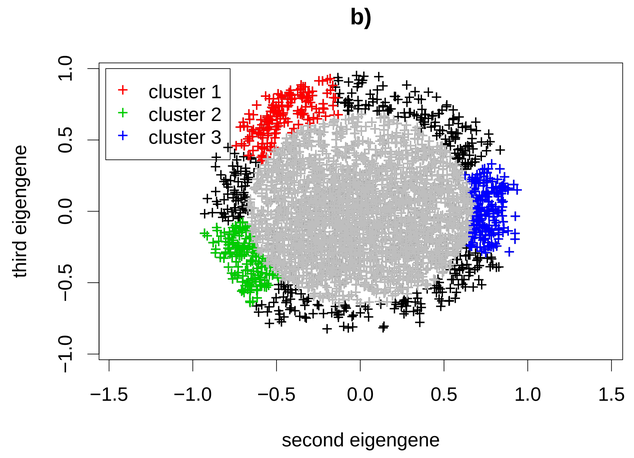

Figure 2: *Correlation plot of the cell-cycle gene expression data. The boundary detected by algorithm 2 and the clusters detected by algorithm 3 are shown.*

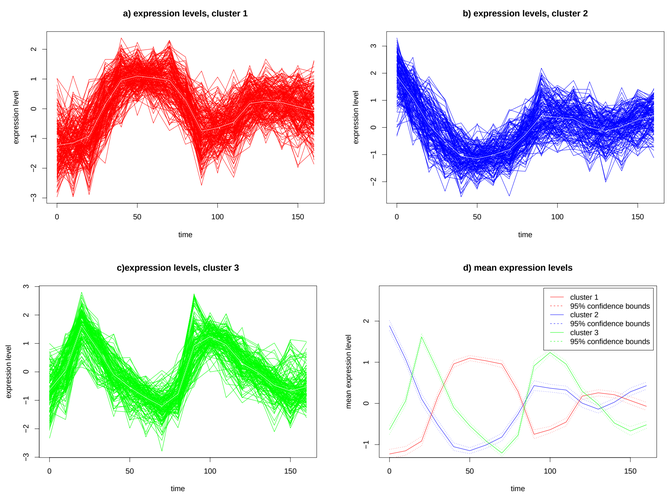

Figure 3: *Three different clusters related to the cell-cycle identified by algorithm 3. Sub-figures a-c) show the expression pattern of each gene in the cluster. Sub-figures d) shows the average expression patterns for the three clusters.*

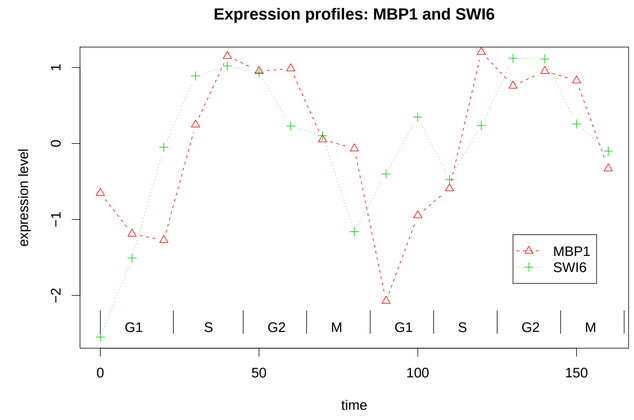

Figure 4: *Expression profiles of the two genes coding for SWI6 and MBP1 respectively.*

3. Conclusion

We introduced three new algorithms for identifying genes with interesting and significant variance in their expression profiles. Algorithm 1 helps to remove noisy gene expression profiles. Algorithm 2 then attempts to identify the genes with the interesting and significant expression profiles. Because we 'filter' genes in these two steps, we avoid clustering too much noise with algorithm 3. It should be noted that algorithm 2 is parameter free, it is completely data driven. Furthermore, algorithm 3 makes no assumption about the number of clusters in the data, such knowledge is not required a priori.

References

[1] Orly Alter, Patrick O. Brown, and David Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *PNAS*, 97(18):10101–10106, 2000.

[2] Raymond J. Cho, Michael J. Campbell, Elizabeth A. Winzeler, Lars Steinmetz, Andrew Conway, Lisa Wodicka, Tyra G Wolfsberg, Andrei E. Gabrielian, David J. Lockhart, and Ronald W. Davis. A genomic-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell*, 2:65–73, 1998.

[3] Ed. F. Deprette, editor. *SVD and signal processing, Algorithms, Applications and Architectures*. North-Holland, 1988.

[4] Gopal K. Kanji. *100 Statistical Tests*. Sage, 1993.