LA-UR- 03-2032

Approved for public release; distribution is unlimited.

Title:

SINGULAR VALUE DECOMPOSITION AND DENSITY ESTIMATION FOR FILTERING AND ANALYSIS OF GENE EXPRESSION

Author(s):

Andreas Rechtsteiner, 174908, CCS-3
Raphael Gottardo
Luis Rocha, 121983, CCS-3
Michael Wall, 165951, CCS-3

Submitted to:

European Conference on Computational Biology (ECCB) in Paris from 09/27-09/30

Singular Value Decomposition and Density Estimation for Filtering and Analysis of Gene Expression

We present three algorithms for gene expression analysis. Algorithm 1, known as serial correlation test, is used for filtering out noisy gene expression profiles. Algorithm 2 and 3 project the gene expression profiles into 2-dimensional expression subspaces identified by Singular Value Decomposition. Density estimates are used to determine expression profiles that have a high correlation with the subspace and low levels of noise. High density regions in the projection, clusters of co-expressed genes, are identified. We illustrate the algorithms by application to the yeast cell-cycle data by Cho et.al. and comparison of the results.

Singular Value Decomposition and Density Estimation for Filtering and Analysis of Gene Expression

Andreas Rechtsteiner$^{\ddagger 1}$, Raphael Gottardo†, Luis Rocha‡, Michael E. Wall‡
‡CCS-3, Los Alamos National Laboratory
†Bioscience division, Los Alamos National Laboratory

Keywords: gene expression analysis, Singular Value Decomposition, noise filtering

1 Introduction

Due to the large amount of data and high levels of noise in the data, gene expression analysis is a challenging task. We introduce three algorithms, two of them novel, for filtering and identification of significanlty expressed genes in time series gene expression data. The two novel algorithms are based on Singular Value Decomposition (SVD) [1] and the projection of the gene expression profiles into 2 dimensional expression subspaces identified by it. We illustrate the algorithms by application to the time-series yeast cell-cycle data published by Cho *et al.* [5].

The first algorithm that we present is the Serial Correlation Test [2]. It is based on the auto-correlation function of a time series and is a common algorithm in time series analysis. To our knowledge the serial correlation test has never before been applied to gene expression data. We use this algorithm for filtering out noisy gene expression profiles. We find it in certain circumstances more useful for this task than commonly used fold-change approaches (as we were able to show when comparing our analysis to the one of Cho *et al.* who used the fold-change approach). Filtering out noisy expression profiles can significantly improve subsequent analysis.

Algorithms 2 and 3 are novel and are applied here to the gene expression data filtered with the Serial Correlation test. In algorithm 2 we project the gene exression profiles into a two-dimesnional expression subspace of interest. We use Singular Value Decomposition (SVD) to select such a sub-space. Alter *et al.* and Holter *et al.* [3, 4] have shown that the first 2 or 3 expression patterns detected by SVD, also called *eigengenes*[2], typically capture most of the interesting gene expression variation in an experiment. We therefore project the gene expression profiles into a 2-dimensional subspace of interest of the first few eigengenes. A crucial observation is that gene expression profiles that project to the boundary of that space will be highly correlated with that subspace whereas gene expression profiles that project towards the center will have a low correlation with that space. In many cases one can also observe that the low-correlated expression profiles are rather uniformly distributed around the center of the space, due to noise, whereas there is structure, *e.g.* clusters, in the projection of the expression profiles at the boundary of the sapce. Our algorithm 2 uses this observation to separate expression profiles whose projection show some structure and which are highly correlated with the two-dimensional subspace from the ones that are low correlated and whose projection is mostly uniform due to noise in the expression profiles. This performs a kind of second 'filtering' of the data. Gene expression profiles which are noisy or with patterns unrelated to the subspace will be removed. Algorithm 2 searches for a boundary in the projection plot that separates the expression profiles whose projection is uniform from the ones whose projection shows structure. The search for this boundary is automated. We calculate the one-dimensional distribution of the polar angles of the expression profiles inside a circle with radius r. The boundary $\tilde{r}$ is selected where the change in this distribution from the uniform distribution is largest. It is important to note that no parameters need to be specified a priori for this algorithm, the selection of the boundary is data driven. For example, it adapts to different levels of noise in the data.

Our 3rd algorithm is a 'clustering algorithm'. It takes the genes identified by algorithm 2 and finds clusters of co-expressed genes, also based on the distribution of polar angles.

2 Application to Yeast Cell-Cycle Data

We illustrate the algorithms by application to the time-series yeast cell-cycle data published by Cho *et al.* [5]. The original data set illustrated by Cho *et al.* contained about 6200 gene expression profiles. Our goal was to detect cell-

[1] To whom correspondence should be addressed. Contact: andreas@lanl.gov

[2] Alter *et al.* introduced the term 'eigengene' for these SVD expression patterns, we adapt this notation here.

cycle related genes, *i.e.* genes with periodic expression profiles, and compare our findings to the analysis by Cho *et al.*

First the serial correlation test was used to remove from the 6200 about 3000 expression profiles which seemed mostly random. SVD was performed on the remaining gene expression profiles. The second and third eigengenes are periodic, sine-like patterns with approximately $\pi/2$ phase difference [3]. To detect genes with periodic expression profiles we project the data into the subspace of eigengenes 2 and 3.

Algorithm 2 was applied to find a boundary that separates expression profiles that are highly correlated with eigengenes 2 and 3, *i.e.* genes with periodic expression profiles, and which show structure in their projection from genes that are low correlated, *i.e.* do not show very periodic expression profiles, and which are mostly uniformly distributed around the origin of the space. See Fig. 2 a) for the projection plot and the boundary selected by Algorithm 2. About 800 genes with periodic expression profiles were found outside of the boundary. The figure also shows the 3 regions with the highest density of expression profiles identified by Algorithm 3. Fig. 2 b) shows the average expression profiles of the genes in these 3 regions. The expression profiles are clearly periodic.

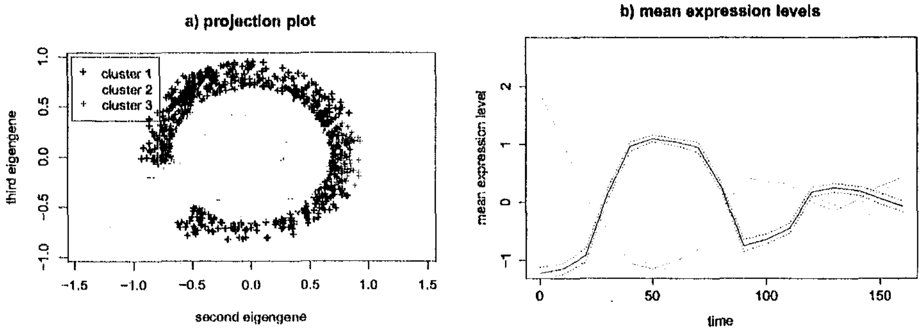

Figure 1: Figure a) shows the projection plot with the boundary selected by algorithm 2 and the 3 clusters of high-density identified by algorithm 3. Figure b) shows the average expression profiles of the genes in the thre clusters and the 95% confidence intervals.

3 Results and further work

We are currently comparing our results to Cho *et al.* and other studies on the same data and other yeast cell-cycle data [6]. The overall biological significance of each cluster can be investigated by associating the clusters with phases of the cell cycle. The biological significance of individual genes within each cluster can be explored making use of the KEGG database, seeking trends in the organization of genes into clusters. One observation we already made is that of the 184 hand-selected genes that were annotated as being involved in transcription, 32 were among the genes that our analysis reported as cell-cycle related. Of the 32 only 1 is found within one of the clusters, indicating a underrepresentation of transcription-related genes in this cluster. Another observation we made is the inclusion of genes encoding SWI6 and MBP1 among our predicted cell-cycle genes. These genes were not among the cell-cycle genes identified by Cho *et al.* , despite being known cell-cycle regulators. We have reproduced Cho *et al.* 's fold-change filtering and found that these genes were removed by this filtering approach, although they exhibit clear periodic expression profiles[4] We found that many other genes we found with periodic expression profiles were removed by Cho *et al.* 's fold-change approach.

Further work is also planned in extending the algorithms. We want to generalize our algorithms to work by projection in n dimensions, not just two. We want to explore iterative approaches, *i.e.* using different eigengenes in different iterative applications of the algorithms.

[3] The first eigengene is non-periodic, it shows a slow, linear decrease. It might capture transient expression changes due to the experimental setup, *e.g.* the release of the yeast cells from cell-cycle arrest at the beginning of the experiment.

[4] The fold-change approach does not 'pay attention' to patterns, only to some absolute change in expression at some time.

References

[1] Deprette, E. F, ed. (1988) *SVD and signal processing, Algorithms, Applications and Architectures.* (North-Holland).

[2] Kanji, G. K. (1993) *100 Statistical Tests.* (Sage).

[3] Alter, O, Brown, P. O, & Botstein, D. (2000) *PNAS* **97**, 10101–10106.

[4] Holter, N, Mitra, M, Maritan, A, Cieplak, M, Banavar, J, & Fedoroff, N. (2000) *Proc. Natl. Acad. Sci. USA* **97**, 8409–8414.

[5] Cho, R. J, Campbell, M. J, Winzeler, E. A, Steinmetz, L, Conway, A, Wodicka, L, Wolfsberg, T. G, Gabrielian, A. E, Lockhart, D. J, & Davis, R. W. (1998) *Molecular Cell* **2**, 65–73.

[6] Spellman, P, Sherlock, G, Zhang, M, Iyer, V, Anders, K, Eisen, M, Brown, P, Botstein, D, & Futcher, B. (1998) *Mol Biol Cell* **9**, 3273–97.