143. Singular Value Decomposition for Analysis of Gene Expression

Andreas Rechtsteiner [1] [2], **Raphael Gottardo** [3], **Luis Rocha** [2], **Michael E. Wall** [2]

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

1 Introduction

Three algorithms, 2 of them novel, are introduced for gene expression analysis. Two of them use Singular Value Decomposition (SVD) [1], or 2 dimensional subspaces in gene expression space identified by it.

The first algorithm that we present is the Serial Correlation Test [2]. We find it to be more useful for filtering out very noisy gene expression profiles in time-series experiments than the commonly used fold-change approaches. Filtering out noise from gene expression data can significantly improve subsequent analysis. To our knowledge the serial correlation test has never been applied to gene expression data.

Our second algorithm is applied to the filtered gene expression data set. Alter *et al.* and Holter *et al.* [3, 4] have shown that the first 2 or 3 expression patterns detected by SVD, linear combinations of the gene expression profiles, typically capture most of the variance in the data. Alter *et al.* introduced the term 'eigengene' for these patterns, we adapt this notation. In algorithm 2 we project the gene expression profiles into a 2-dimensional subspace of interest of the first few eigengenes. We then search for structure, or clusters, in that projection among the genes with high correlation with that subspace, *i.e.* away from the origin of the plot. 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. The search for structure is automated. It is important to note that no parameters need to be specified a priori, the algorithm is completely data driven. For example, it adapts to different levels of noise in the data.

Our 3rd algorithm is a 'clustering algorithm'. It clusters the genes identified by algorithm 2, the genes with high correlation with the subspace and some structure in the projection plot.

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]. Our goal was to detect cell-cycle related genes and compare our findings to the analysis by Cho *et al.* . We therefore attempt to detect periodic expression patterns (see [5]).

First the serial correlation test was used to remove from the 6200 about 3000 gene 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 $\pi/2$

[1] To whom correspondence should be addressed. Contact: andreas@lanl.gov
[2] CCS-3, Los Alamos National Laboratory
[3] Bioscience division, Los Alamos National Laboratory

phase difference [4]. To detect genes with periodic expression profiles we project the data into the subspace of eigengenes 2 and 3.

Algorithm 2 detected about 800 genes with periodic expression profiles that seemed significant. 235 of these were also detected by Cho *et al.* . We found that most of the others were probably removed by Cho *et al.* 's fold-change approach. Algorithm 3 is applied to the expression profiles of these genes and 3 clusters with different phases in their periodic profiles are found (see Figure 1).

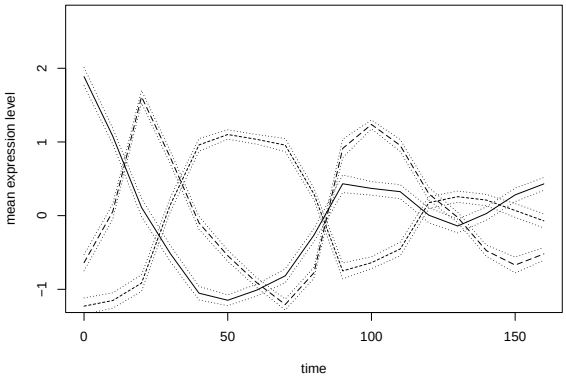

Figure 1: The three averaged expression profiles of the three clusters of periodic genes identified by the application of the 3 algorithms. (95% confidence intervals for the mean values are also plotted.)

The clusters can be associated with different phases of the cell-cycle. One observation was 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 report 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.

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.

[4]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.