Funct Integr Genomics (2002) 2:92–97
DOI 10.1007/s10142-002-0047-5

ORIGINAL PAPER

Debashis Ghosh

Resampling methods for variance estimation of singular value decomposition analyses from microarray experiments

Received: 14 September 2001 / Accepted: 18 December 2001 / Published online: 14 February 2002

Abstract Microarray experiments offer the ability to generate gene expression measurements for thousands of genes simultaneously. Work has begun recently on attempting to reconstruct genetic networks based on analyses of microarray experiments in time-course studies. An important tool in these analyses has been the singular value decomposition method. However, little work has been done on assessing the variability associated with singular value decomposition analyses. In this report, we discuss use of the bootstrap as a method of obtaining standard errors for singular value decomposition analyses. We consider use of this method both when there are replicates and when no replicates exist. The proposed methods are illustrated with an application to two datasets: one involving a human foreskin study, the other involving yeast.

Keywords Bootstrap · Gene expression · Synchronized time-course study

Introduction

With the recent development of microarray technology, it has become possible to measure gene expression simultaneously on a large-scale basis. Because mRNA levels can be assessed simultaneously for thousands of genes, work has now begun on attempting to elucidate genetic networks and metabolic pathways in various model organisms. One type of experiment that is useful for understanding regulatory mechanisms is the synchronized time-course study. In these experiments, cells are suspended in a certain state and then released. At certain time points, mRNA is taken from the cells, and microarrays are run on the mRNA samples. The result is a set of measurements from chips at various points in time.

One technique proposed by several authors for analyzing microarray time-course data is the singular value decomposition (SVD; Golub and van Loan 1996). By using the SVD, the raw, high-dimensional microarray measurements are transformed into a set of independent variables in a lower-dimensional subspace. The uses of SVD analysis of microarray data have been manifold. There have been several applications of SVD analysis to gene expression data (Craig et al. 1997; Hilsenbeck et al. 1999; Dewey and Galas 2001). One proposal involved using SVD as a means of filtering and preprocessing the data (Alter et al. 2000). Another use has been to summarize microarray time-course data (Raychaudhuri et al. 2000). A third application has been to summarize the time-course data into so-called characteristic modes that are easier to study (Holter et al. 2000). In fact, these authors have suggested that a subset of these modes can explain much of the variation in gene expression dynamics and have begun investigations into dynamic modeling of gene expression data using these quantities (Holter et al. 2001).

However, there has been virtually no mention of variability assessment of SVD analyses. In Holter et al. (2000), a SVD of a randomly generated dataset was considered as a comparison with that of the real data. However, without any type of formal variance estimation, it is impossible to determine what are "real" patterns in the SVD analyses versus those that arise by chance. Similar issues arise in hierarchical clustering of microarray data (Zhang and Zhao 2000; Kerr and Churchill 2001).

Assessing this variability requires development of a statistical framework for SVD analyses of microarray time-course data. Our focus will be on performing inference for the characteristic modes. In this article, we propose a nonparametric approach for variance estimation involving the nonparametric bootstrap (Efron and Tibshirani 1993) which is applicable when there are replicate data available on time-course experiments. In many instances, however, such replicate data are not

D. Ghosh (✉)
Department of Biostatistics, School of Public Health,
University of Michigan, 1420 Washington Heights, Room M4057,
Ann Arbor, MI 48109–2029, USA
e-mail: ghoshd@umich.edu
Tel.: +1-734-6159824, Fax: +1-734-7632215

available. For this scenario, we describe the assumptions made in applying bootstrap methods. We will then apply the proposed methods to data from a human fibroblast study (Cho et al. 2001) as well as a yeast study (Gasch et al. 2000). While these data were generated using cDNA arrays, we expect that similar considerations should hold using other types of high-throughput technologies.

Materials and methods

Notation and singular value decomposition

Before describing the singular value decomposition, we first introduce some notation. Let x_{ij} denote the gene expression measurement for the jth gene of the ith sample (collected at time t_i), $j=1,...,p$, $i=1,...,n$. Typically, p is in the order of several thousands, while n is in the order of 40–60. We define the vector $\mathbf{x}_i=(x_{i1},...,x_{ip})''$, where $\mathbf{a}'$ is the transpose of the vector $\mathbf{a}$, and the $p\times n$ matrix $\mathbf{X}=[\mathbf{x}_1\cdots\mathbf{x}_n]$. The SVD of X is defined in the following manner:

$$X = \mathbf{ADF} \quad (1)$$

where $\mathbf{A}$ is a $p\times n$ matrix of loadings, $\mathbf{D}$ is a $p\times p$ diagonal matrix of the singular values of $\mathbf{X}$, and $\mathbf{F}$ is a $p\times p$ matrix. The SVD describes the structure of the matrix $\mathbf{X}$. For example, the number of non-zero singular values on the diagonal of $\mathbf{D}$ is equivalent to the rank of $\mathbf{X}$. Furthermore, we have that $\mathbf{A'A}=\mathbf{I}$ and $\mathbf{F'F}=\mathbf{FF'}=\mathbf{I}$, where $\mathbf{I}$ is an $n\times n$ identity matrix. Typically, the authors have used $\mathbf{D}$ or $\mathbf{DF}$ as the lower-dimensional summary of $\mathbf{X}$; the rows of these $p\times p$ matrices correspond to the characteristic modes of $\mathbf{X}$ proposed by Holter et al. (2000). The computation of the SVD is typically iterative; a good summary of algorithms for performing this task can be found in Golub and van Loan (1996).

Nonparametric bootstrap with replicate data

We will first consider the scenario of replicated time-course experiments. In this instance, there will be multiple $\mathbf{X}$ matrices. One can use the bootstrap (Efron and Tibshirani 1993) to sample the individual columns of $\mathbf{X}$ (i.e. the samples) from the available replicate experiments. Alternatively, one could sample the rows of $\mathbf{X}$ (i.e. the genes) from the experiments. Say we do this B times. For each of the B bootstrapped gene expression matrices, we then apply the SVD (Eq. 1). This yields a set of B bootstrapped DF matrices. We can then plot the characteristic modes for the observed data, along with the corresponding modes for the bootstrapped datasets. We can use the bootstrapped datasets to construct pointwise confidence intervals for the modes at each of the time points. We generate B=10,000 bootstrap samples and take the confidence limits to be based on percentiles for the distribution of characteristic modes at each time point. For example, we can construct 95% pointwise confidence intervals by using the 2.5th and 97.5th percentiles for the characteristic modes at each time point. If we want to have a higher level of conservatism, we can construct 99% pointwise confidence intervals by using the 0.5th and 99.5th percentiles for the characteristic modes at each time point. It is important to note that these procedures make no assumption about the dependence of gene expression measurements within and across the n columns of $\mathbf{X}$.

Nonparametric bootstrap with no replicate data

In many situations, however, we only have data from one time course experiment and no replicate experiments. Suppose we apply the bootstrap procedures described in the previous paragraph. If we generate bootstrapped datasets by resampling from the rows (i.e. the genes) of $\mathbf{X}$, then we assume that the correlation between any two genes are the same. If we resample from the columns (i.e. the time points) of $\mathbf{X}$, then there is an assumption that for all genes, the correlation between the gene expression measurements at any two time points is the same. This does not seem to be a reasonable assumption in most instances, so we will instead focus on resampling genes later in the paper.

These are important assumptions that need to be carefully examined. By not replicating the time-course experiment, we lose the flexibility to assume arbitrary dependence among genes and among experiments across time points. In addition, any effects due to time are confounded with the between-chip variation. Such aspects needed to be taken into account when analyzing the data from a single time-course experiment.

Implementation

As was mentioned before, any of the standard software packages that fit the SVD can be used to implement the methods described in the algorithms section. All of these methods presented in the Results section use the R language, a statistical software package that can be freely downloaded at the following website: http://cran.r-project.org/. The commands used to analyze the data can be found at the author's webpage, at the following URL: http://www.sph.umich.edu/~ghoshd/COMPBIO/SVD2/index.html.

Results

Human fibroblast data

We have applied the resampling ideas described above to a series of experiments conducted by Cho et al. (2001). In these experiments, primary fibroblasts were prepared from human foreskin and then arrested in the late G_1 stage using a thymidine-block protocol (Rao and Johnson 1970). The cells were then released and collected every 2 h for 24 h. Using high-density oligonucleotide arrays, mRNA was measured at 12 time points. The authors of the study carried out the entire experiment in duplicate; we use their notation and refer to the two experiments as N2 and N3. While Cho et al. (2001) were interested in determining cell-cycle regulated transcripts, they averaged the data from the N2 and N3 experiments. We use their data simply to illustrate the techniques described in the algorithms. The data can be found at the following website: http://www.salk.edu/docs/labs/chipdata.

The data normalization procedure we used was the same as that used by previous authors (Shedden and Cooper, personal communication). While there are measurements available on 7,129 genes, we excluded genes that had fewer than two positive expression measurements for either N2 or N3. This was done because the transcript levels for these genes were so low that it would be difficult to distinguish signal from background expression. This left a total of 5,914 genes. The next step was to apply a transformation of $\log(x+200)$ using base 2; this was done to stabilize the distribution of the gene expression measurements. Then the genes on each chip were centered and scaled to have mean 0.0 and variance 1. This was done to adjust for between-chip variation in hybridization. The second transformation corrected for difference in between-gene variation by centering and

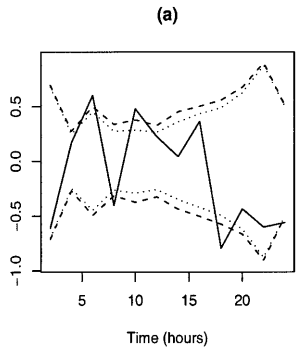


Fig. 1a–d Plot of the first four characteristic modes (*solid line*) for N2 data from Cho et al. (2001), along with 90% (*dotted*) and 95% (*dashed*) pointwise confidence intervals. Intervals were obtained by applying bootstrap to columns (time points) of data matrices from N2 and N3

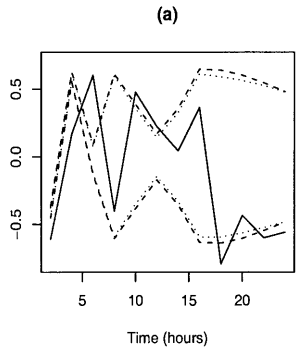


Fig. 2a–d Plot of the first four characteristic modes (*solid line*) for N2 data from Cho et al. (2001), along with 90% (*dotted*) and 95% (*dashed*) pointwise confidence intervals. Intervals were obtained by applying bootstrap to rows (genes) of data matrices from N2 and N3

scaling the measurements such that for each gene, the mean is 0.0 and variance is 1.0.

We first utilized the replicate data from both N2 and N3 and applied the bootstrap. We generated 10,000 bootstrapped datasets for each study. We took two separate scenarios; in the first, the columns (i.e. the individual time points) of the matrix were resampled while for the second, the rows (i.e. the individual genes) of the matrix were resampled. We calculated 90% and 95% confidence intervals for the modes at each time point; these values tend to be close to one another. The first four characteristic modes for the observed data (based on N2), along with the bootstrapped confidence intervals, are given in Fig. 1 for the first scenario and in Fig. 2 for the second. The modes themselves appear to suggest periodic variation; furthermore, based on resampling time points, many of the observed patterns in the modes cannot be explained by chance. If we decide to resample genes from the replicate experiments, we tend to find more significant patterns in the data.

We then used only the data from N2 so that we had no replicate data available. In Fig. 3, we show the results for the first four characteristic modes and the associated confidence intervals when the rows (i.e. the genes) of the gene expression matrix are used for resampling. Note that there is an assumption in this procedure that for each gene, the correlation between measurements at any two time points is the same. In addition, we are unable to in-

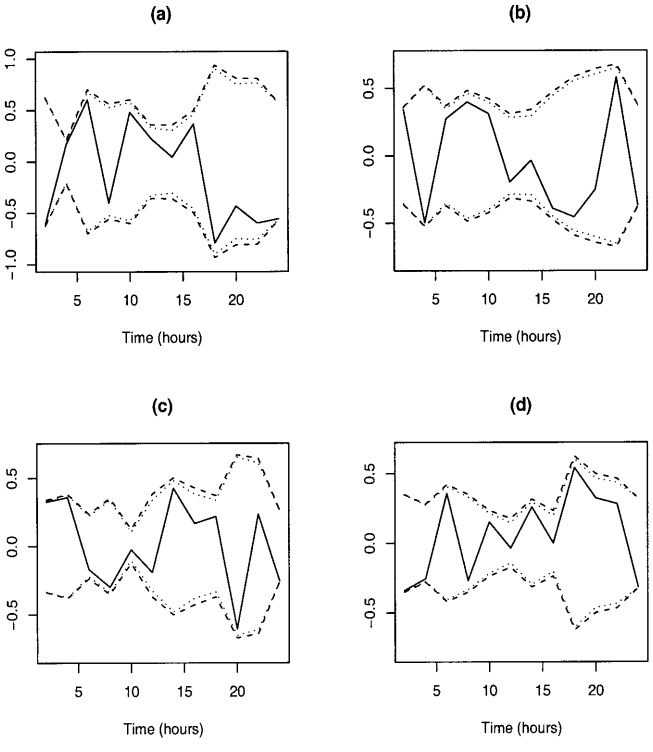
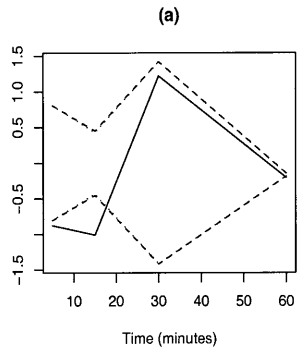
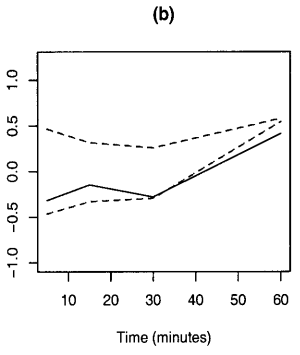
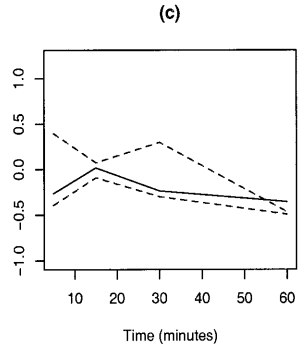


Fig. 3a–d Plot of the first four characteristic modes (*solid line*) for N2 data from Cho et al. (2001), along with 90% (*dotted*) and 95% (*dashed*) pointwise confidence intervals. Intervals were obtained by applying bootstrap to rows (genes) of data matrices from N2







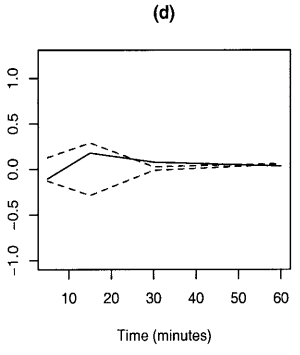
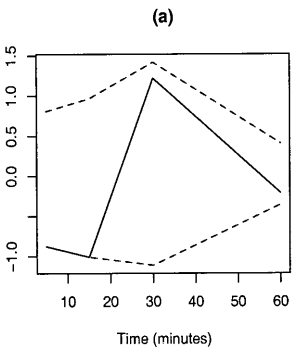
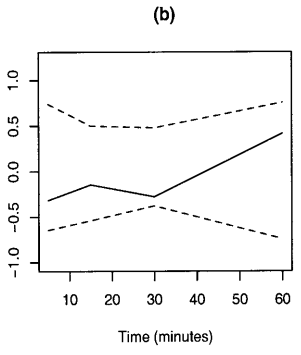
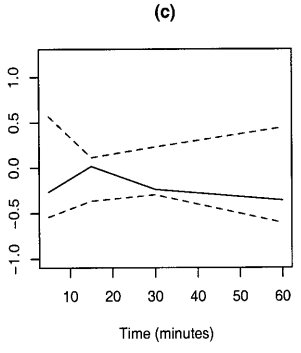


Fig. 4a–d Plot of the first four characteristic modes (*solid line*) for the first set of heat shock experiments from Gasch et al. (2000), along with 90% (*dotted*) and 95% (*dashed*) pointwise confidence intervals. Intervals were obtained by applying bootstrap to rows (genes)from two replicate heat shock experiments







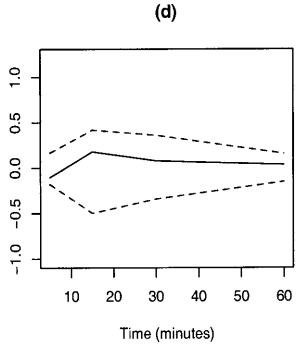


Fig. 5a–d Plot of the first four characteristic modes (*solid line*) for the first set of heat shock experiments from Gasch et al. (2000), along with 90% (*dotted*) and 95% (*dashed*) pointwise confidence intervals. Intervals were obtained by applying bootstrap to columns (time points) from two replicate heat shock experiments

corporate the variability from replicate time-course experiments. Comparing these graphs with those in Figs. 1 and 2 suggests that there is substantial variability between replicate time-course experiments. Consequently, our results appear to be much more significant by resampling only time points from one dataset.

Yeast data

Our second data analysis involved a series of yeast microarray experiments conducted by Gasch et al. (2000). They were interested in characterizing the global gene expression profile response to various types of environmental stimuli, such as temperature shock, starvation and nitrogen depletion. They conducted 177 microarray experiments, of which 142 were reported in the study. Each chip contained approximately 6,200 known or predicted yeast genes that had been identified at the time of the study.

We focussed on the heat shock experiments conducted by these authors. To briefly summarize these studies, yeast cells were first grown continuously at 25°C, collected by centrifugation, resuspended and then returned to growth at 37°C (Gasch et al. 2000). While microarray data were collected at several time points, replicate chip experiments were performed at 5, 15, 30 and 60 min, so we will focus on these time points. In contrast to the study by Cho et al. (2001), these authors utilized the spotted array technology for performing microarray experiments. The reference pool, labeled with the Cy3 dye, consisted of the mRNA samples from all the time points, while the mRNA at the particular time point was labeled with the Cy5 dye.

The first set of results pertains to utilizing the replicate data for the bootstrap method. We first used the replicate data for the bootstrap; the results based on resampling time points are given in Fig. 4, while those based on resampling genes are given in Fig. 5. Based on these plots, it appears that most of the patterns in the characteristic modes can be explained by chance, with some exceptions at both early and late time points. In comparison to the data from Cho et al. (2001), we only have four data points, so any trends should be interpreted cautiously.

We then applied the bootstrap to the first experiment, resampling the genes. The plots of the confidence intervals of the characteristic modes are given in Fig. 6. Based on these graphs, there appear to be more significant patterns in the characteristic modes relative to those found in Figs. 4 and 5. However, it is important to remember that the between-experiment variability is not incorporated in this procedure so that any patterns found here might be particular to the experiment.

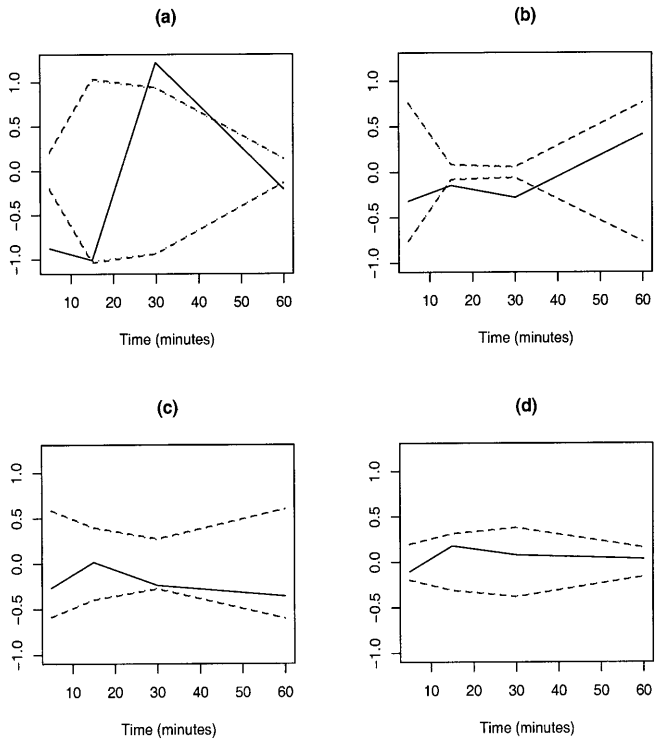


Fig. 6a–d Plot of the first four characteristic modes (*solid line*) for the first set of heat shock experiments from Gasch et al. (2000), along with 90% (*dotted*) and 95% (*dashed*) pointwise confidence intervals. Intervals were obtained by applying bootstrap to rows (genes) of data from the first heat shock experiment

Discussion

The goal of this paper is to bring out the importance of assessment of variability of SVD analyses in microarray experiments. Since data from high-throughput technologies are extremely multivariate, there will be a large number of trends that will arise due to randomness. It is important not to treat these findings as confirmatory without statistical and/or external experimental validation.

Our primary interest in applying SVD was to study patterns in the characteristic modes. In particular, this method studies the aggregate variation over all the genes on the microarray and it is not possible to make inferences on individual genes based on the results of the SVD analysis.

The methods proposed here would be useful in the situation where replicate time-course data exist. While replicating microarray time-course experiments can be expensive, such an analysis could still be performed if one were performing a meta-analysis of time-course experiments using a publicly available microarray database. However, in many situations, it is not feasible to perform replicate time-course experiments. It would be desirable to develop model-based techniques for SVD analyses of microarray time-course data in the absence of replicate experimental data. In addition, such an approach could allow for performing inference on the number of characteristic modes.

The choice of which method to use in practice depends on two criteria. The most desirable method to use is the nonparametric bootstrap with replicate experiments. This method makes the fewest assumptions on correlation between genes and between experiments. If replicate data are available, there is no clear winner between resampling genes or resampling experiments. If no replicate data are available, then the analyst should determine what is a more realistic assumption, exchangeability between genes or exchangeability between time points. If the answer is the former, then one should resample genes; otherwise, one should resample experiments.

If the measurement error changes over time, then this will not affect the validity of the bootstrap with replicate data. On the other hand, the confidence intervals for the bootstrap with no replicate data will not be valid.

Another issue has to do with the spacing of the time-course experiments. The information on the time points at which the measurements are taken is not utilized in the SVD analysis. However, in practice, measurements will be taken at potentially irregularly spaced time points. Aach and Church (2001) have proposed using dynamic programming-type algorithms for aligning gene expression time-series data. An alternative approach would be to treat the sequence of gene expression measurements over time as a function and to consider functional SVD techniques for analyzing gene expression dynamics. This is another interesting area for research that we are currently exploring.

References

Aach J, Church GM (2001) Aligning gene expression time series with time warping algorithms. Bioinformatics 17:495–508

Alter O, Brown PO, Botstein D (2000) Singular value decomposition for genome-wide expression data processing and modeling. Proc Natl Acad Sci USA 97:10101–10106

Cho RJ, Huang M, Campbell MJ, Dong H, Steinmetz L, Sapinoso L, Hampton G, Elledge SJ, Davis RW, Lockhart DJ (2001) Transcriptional regulation and function during the human cell cycle. Nat Genet 27:48–54

Craig JC, Eberwine JH, Calvin JA, Wlodarczyk B, Bennett GD, Finnell RH (1997) Developmental expression of morphoregulatory genes in the mouse embryo: an analytical approach using a novel technology. Biochem Mol Med 60:81–91

Dewey TG, Galas D (2001) Dynamic models of gene expression and classification. Funct Integr Genomics 1:269–278

Efron B, Tibshirani E (1993) An introduction to the bootstrap. Chapman and Hall, New York

Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO (2000) Genomic expression programs in the response of yeast cells to environmental changes. Mol Biol Cell 11:4142–4157

Golub GH, van Loan CF (1996) Matrix computations. John Hopkins University Press, Baltimore

Hilsenbeck SG, Friedrichs WE, Schiff R, O'Connell P, Hansen RK, Osborne CK, Fuqua SA (1999) Statistical analysis of array expression data as applied to the problem of tamoxifen resistance. J Natl Cancer Inst 91:400–401

Holter NS, Mitra M, Maritan A, Cieplak M, Banavar J, Fedoroff NV (2000) Fundamental patterns underlying gene expression profiles: simplicity from complexity. Proc Natl Acad Sci USA 97:8409–8414

Holter NS, Maritan A, Cieplak M, Fedoroff NV, Banavar JR (2001) Dynamic modeling of gene expression data. Proc Natl Acad Sci USA 98:1693–1698

Kerr MK, Churchill GA (2001) Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. Proc Natl Acad Sci 98:8961–8965

Rao PN, Johnson RT (1970) Mammalian cell fusion: studies on the regulation of DNA synthesis and mitosis. Nature 225:159–164

Raychaudhuri S, Stuart JM, Altman RB (2000) Principal components analysis to summarize microarray experiments: application to sporulation time series. Pac Symp Biocomput 5:452–463

Zhang K, Zhao H (2000) Assessing reliability of gene clusters from gene expression data. Funct Integr Genomics 1:156–173