Published in final edited form as:
Pac Symp Biocomput. 2000 ; : 455–466.

PRINCIPAL COMPONENTS ANALYSIS TO SUMMARIZE MICROARRAY EXPERIMENTS: APPLICATION TO SPORULATION TIME SERIES

Soumya Raychaudhuri[*], **Joshua M. Stuart**[*], and **Russ B. Altman**[Ψ]
Stanford Medical Informatics, Stanford University, 251 Campus Drive, MSOB X-215, Stanford CA 94305-5479 {sxr, stuart, altman} @smi.stanford.edu

Abstract

A series of microarray experiments produces observations of differential expression for thousands of genes across multiple conditions. It is often not clear whether a set of experiments are measuring fundamentally different gene expression states or are measuring similar states created through different mechanisms. It is useful, therefore, to define a core set of independent features for the expression states that allow them to be compared directly. Principal components analysis (PCA) is a statistical technique for determining the key variables in a multidimensional data set that explain the differences in the observations, and can be used to simplify the analysis and visualization of multidimensional data sets. We show that application of PCA to expression data (where the experimental conditions are the variables, and the gene expression measurements are the observations) allows us to summarize the ways in which gene responses vary under different conditions. Examination of the components also provides insight into the underlying factors that are measured in the experiments. We applied PCA to the publicly released yeast sporulation data set (Chu et al. 1998). In that work, 7 different measurements of gene expression were made over time. PCA on the time-points suggests that much of the observed variability in the experiment can be summarized in just 2 components—i.e. 2 variables capture most of the information. These components appear to represent (1) overall induction level and (2) change in induction level over time. We also examined the clusters proposed in the original paper, and show how they are manifested in principal component space. Our results are available on the internet at http://www.smi.stanford.edu/projects/helix/PCArray.

1 Introduction

The study of gene expression has been greatly facilitated by DNA microarray technology (Schena et al. 1995). DNA microarrays measure the expression of thousands of genes simultaneously, and have been described elsewhere (Chee et al. 1996, Chen et al. 1998, Duggan et al. 1999, Schena et al. 1995). The anticipated flood of biological information produced by these experiments will open new doors into genetic analysis (Lander 1999). Expression patterns have already been used for a variety of inference tasks. For example, microarray data has been used to identify gene clusters based on co-expression (Eisen et al. 1998, Michaels et al. 1998), define metrics that measure a gene's involvement in a particular cellular event or process (Spellman et al. 1998), predict regulatory elements (Brazma et al. 1998), and reverse engineer transcription networks (D'Haeseleer et al. 1999, Liang et al. 1998). The success of these efforts relies on the integrity of the expression data. Both experimental noise and hidden dependencies among a set of experimental conditions may

[Ψ]To whom correspondences should be addressed.
[*]These authors contributed equally to this communication.

confound the inferential process. It is non-trivial to eliminate either of these complicating factors. One particular problem is that different experiments that seem different because of their biological context (heat shock, starvation, or oxygen deprivation, for example) may actually be identical or very similar in terms of the gene expression state that results. In such cases, a naïve analysis might associate some genes too tightly because multiple redundant measurements. Thus, it may be beneficial to pre-process the data before analysis in order to identify the independent information content of different experimental conditions.

Principal Components Analysis (PCA) is an exploratory multivariate statistical technique for simplifying complex data sets (Basilevsky 1994, Everitt & Dunn 1992, Pearson 1901). Given m observations on n variables, the goal of PCA is to reduce the dimensionality of the data matrix by finding r new variables, where r is less than n. Termed principal components, these r new variables together account for as much of the variance in the original n variables as possible while remaining mutually uncorrelated and orthogonal. Each principal component is a linear combination of the original variables, and so it is often possible to ascribe meaning to what the components represent. Principal components analysis has been used in a wide range of biomedical problems, including the analysis of microarray data in search of outlier genes (Hilsenbeck et al. 1999) as well as the analysis of other types of expression data (Vohradsky et al. 1997, Craig et al. 1997).

DNA microarray data sets are now appearing in the published literature, and most initial analyses have focused on characterizing the waveform of gene epxression over time, and in clustering benes based on this waveform or other features. When clustering genes based on expression information, it can be important to determine if the experiments have independent information or are highly correlated. Chu et al (1998) measured gene expression at seven time points during sporulation in yeast, and in two mutant yeast strains. They identified 7 clusters of key genes grouped based on the approximate times during which members are up-regulated.

A PCA analysis of DNA microarray data can consider the genes as variables or the experiments as variables or both. When genes are variables, the analysis creates a set of “principal gene components” that indicate the features of genes that best explain the experimental responses they produce. When experiments are the variables, the analysis creates a set of “principal experiment components” that indicate the features of the experimental conditions that best explain the gene behaviors they elicit. When both experiments and genes are analyzed together, there is a combination of these affects, the utility of which remains to be explored. This report focuses on consideration of the experiments as variables. We first create a covariance matrix to measure how each experiment contributes information to the data set. We then summarize the information compactly with the principal experimental components. Finally, we show that this analysis clarifies the relationship between previously reported clusters and is a starting point for examining the detailed relationships and differences between genes.

2 Methods

We start with a matrix of expression data, A, where each row corresponds to a different gene and each column corresponds to one of several different conditions to which the cells were exposed. The a_{it} entry of the matrix contains the i^{th} gene’s relative expression ratio with respect to a control population under condition t. To moderate the influence of gene expression ratios above and below one, we applied the natural log transform to all ratios (Eisen et al. 1998). Up-regulated genes thus have a positive log expression ratio, while down-regulated genes have a negative log expression ratio. We did not normalize the conditions to norm 0, variance 1 as sometimes recommended when attempting PCA on

measurements that are not on a comparable scale (Everitt & Dunn 1992). The log ratios included in the analysis are comparable, no further preprocessing was necessary.

To compute the principal components, the n eigenvalues and their corresponding eigenvectors are calculated from the $n \times n$ covariance matrix of conditions. Each eigenvector defines a principal component. A component can be viewed as a weighted sum of the conditions, where the coefficients of the eigenvectors are the weights. The projection of gene i along the axis defined by the j^{th} principal component is:

$$a'_{ij} = \sum_{t=1}^{n} a_{it} v_{tj}$$

Where v_{tj} is the t^{th} coefficient for the j^{th} principal component; a_{it} is the expression measurement for gene i under the t^{th} condition. A' is the data in terms of principal components. Since V is an orthonormal matrix, A' is a rotation of the data from the original space of observations to a new space with principal component axes.

The variance accounted for by each of the components is its associated eigenvalue; it is the variance of a component over all genes. Consequently, the eigenvectors with large eigenvalues are the ones that contain most of the information; eigenvectors with small eigenvalues are uninformative.

Determining r, the true dimensionality of the data, and eliminating noisy components is often *ad hoc* and many heuristics exist. Eliminating low variance components, while reducing noise, also discards information. We chose to use one criterion that discards all components accounting for less than $(70/n)$% of the overall variability (Everitt & Dunn 1992). The **Matlab**™ software package (The MathWorks, Inc., Natick, MA) was used to conduct most of our calculations.

3 Results

The data for this analysis was obtained from a publicly accessible web site[1]. The data contains expression ratios for 6118 known or predicted genes from *Saccharomyces cerevisiae*. The data was collected for each gene at 7 different time points (0hrs, 0.5hr, 2hrs, 5hrs, 7hrs, 9hrs, 11.5hrs) during sporulation. Thus, the matrix to be analyzed has 6118 rows of genes and 7 columns of conditions corresponding to each of the measured time points. Table 1 reports the mean, median, and variance of each time point from the sporulation data over all genes. The means and medians are slightly negative but quite close to zero. Also note the relatively low variance of the t=0 time point; this is reassuring since the initial population should be similar to the background population.

Our analysis indicates that we can summarize the data with just two variables. Table 2 contains all 7 principal components and their corresponding eigenvalues. Figure 1 is a plot of the eigenvalues of the components. Two eigenvalues lie above the 10% (70/7) cutoff, suggesting two dimensions for the sporulation data. The first two principal components account for over 90% of the total variability; including the third component accounts for almost 95%. The meaning of these components can be distilled from their respective coefficients.

[1] http://cmgm.stanford.edu/pbrown/sporulation/index.html

The first component represents a weighted average and distinguishes genes by their average overall expression. Ignoring the t=0 coefficient (it has negligible magnitude), it can be seen in Figure 2A that the remaining coefficients are positive (see also Table 2). The coefficients are proportional to the variance of the time points they are associated with (correlation = 0.97). The first component is an average expression weighted by the information content (i.e. variance) of a particular experiment. Genes with highly positive values along this component are up-regulated during sporulation, whereas genes with highly negative values are down-regulated.

The second component represents change in expression over time; it distinguishes genes by their first derivatives. In Figure 2B the coefficients linearly increase with time from negative to positive values. Again, the exception to the rule is the low variance t=0 observation which has a negligible coefficient. Consider a gene *i* that is repressed (negative log expression ratio) in the early time points and highly induced (positive log expression ratio) in the final time points. The coefficient multiplied by the log expression score will be positive for each time point. Gene *i*'s value along the second component, a'_{i2}, is large and positive since every product in the sum is positive. Alternatively the second component for a gene that is induced early and repressed later will be large and negative. The expression scores are multiplied with coefficients of the opposite sign, yielding a large negative score. This component is positive for genes whose relative expression increases through time, and negative for those whose relative expression decreases; it measures positive trend in expression.

The third component measures concavity—notice the parabolic nature of the coefficients in Figure 2C (again ignore the negligible t=0 coefficient). Consider a gene *i* that is expressed at background level in the early and middle time points, but induced in the final time points—it has an expression profile that is concave up. Since the only non-zero expression levels occur at the final time points, only the later negative coefficients contribute to the sum, a'_{i3}. Consequently this gene will have a negative third component. Alternatively consider a gene with a similar profile, but that is expressed in the middle time points also (concave down); in this case the middle time points with positive coefficients increase the score along this component. The score of the second gene will be less negative.

The first two components account for over 90% of the variance allowing most of the information to be visualized in two dimensions. Thus, even though there were seven experiments in the time series, there were only two or three independent features for each gene. All yeast genes are plotted in Figure 3 against the first two principal components; an ellipse enclosing 95% of the genes is drawn to distinguish between high and low variance genes. The genes appear to be distributed in a unimodal distribution. The data has been made available as a VMRL source at http://www.smi.stanford.edu/projects/helix/PCArray; the user can quickly navigate through two or three dimensional component space. Each data point is linked to its corresponding entry in the Saccharomyces Genome Database (Cherry et al. 1998).

The transcription factor NDT80 is key to the induction of many genes expressed in the middle of the sporulation process (Xu et al. 1995). The original dataset also includes measurements of gene expression for a NDT80 knockout microarray experiment and an ectopic NDT80 over-expression experiment. Including these extra experiments in the analysis offers an opportunity to test the robustness of our analysis. The coefficients for the first two components are consistent with our understanding of the phenotype of these cells. In particular, the NDT80 knockout experiment traps cells in an early stage of sporulation; correspondingly, the coefficients in the first two components are most similar to the t=2 hour coefficients from the sporulation time series. Similarly, the NDT80 over-expression data

yields coefficients most similar to the t=11 hour coefficients. Since NDT80 is a sporulation promoting factor, the effects of over-expression may cause a phenotype that mimics a late time point.

4 Discussion

Our results with the sporulation data confirm that PCA can find a reduced set of variables that are useful for understanding the experiments. The application of PCA to time series is somewhat controversial because of the problems with uneven time intervals and the dependencies between data points. In this case, PCA identifies basic temporal patterns, such as magnitude, change, and the concavity of overall expression as the important features that characterize genes. Application of PCA (unpublished result) to the publicly available cell division cycle data[2] also reveals that PCA can also identify periodic patterns in time series data (Spellman et al. 1998). For example, the cell cycle data reveals a 110 min period for the cdc15 synchronized experiment, consistent with the cell cycle duration.

Reduction of dimensionality in the sporulation data aids in data visualization; we can immediately see the unimodal quality of the sporulation data (Figure 3). The unimodal distribution of expression in the most informative two dimensions suggests the genes do not fall into well-defined clusters.

In the initial presentation of the data the investigators used clustering techniques to identify several gene classes relevant to sporulation: “metabolic”, “early I”, “early II”, “middle early”, “middle”, “middle late”, and “late” (Chu et al. 1998). For each class a canonical expression profile was calculated from a set of sample genes. These classes are plotted in Figure 4A; each ellipse in the plot represents a class. The location and dimensions of each ellipse was calculated from the average and standard deviation of the sample genes of the class. They are drawn so that approximately 68% (+/− 1SD in both dimensions) of the genes in the class are enclosed; in Figure 4B they are drawn to enclose 95% (+/− 1.96SD) of the genes in the class.

An approximate understanding of a class’s expression dynamic can be obtained quickly by looking at its location in space. For example, genes occupying the lower right quadrant (high PCA1, low PCA2) are up-regulated early but return to background later in sporulation. These genes have expression levels that decrease over time but maintain a high overall expression level relative to the control. Examples of these genes are ZIP1 (synaptonemal complex formation), IME2 (meiosis regulator), and HOP1 (homologous chromosome pairing), classified as “early I” or ”metabolic” genes.

Exploring other quadrants can rapidly identify genes of potential interest. Genes with low overall expression levels that decrease over the course of sporulation can be found in the lower left quadrant. Many genes involved in metabolic or catabolic processes such as ERG6 (ergosterol synthesis), FBP1 (gluconeogenesis), and SAM2 (methionine biosynthesis) are found in this quadrant. Genes in the upper left are initially repressed and return to normal. Many of these genes are involved in protein synthesis. Examples include ISF1 (RNA splicing), BAP3 (valine transporter), and DBP3 (RNA helicase). The early repression may correspond with the cells’ initial cessation of protein synthesis and growth; the renewed expression may function to pack the maturing spores with translation machinery (Chu et al. 1998).

[2] http://genome-www.stanford.edu/cellcycle

Principal components analysis is often used as a preprocessing step to clustering (Everitt 1993). However, our work suggests that clustering genes with certain expression data sets may not be appropriate. In Figure 4A, the genes are not located in clusters - rather they are spread throughout this space. Focusing on the upper right quadrant in Figure 4B, it can be seen that the clusters presented in the original publication have a considerable amount of overlap. For unimodal or other smoothly varying distributions, distinctions drawn by clustering methodologies maybe more confusing than helpful. In particular, these clusters highlight the potential biases used in analyzing clusters using traditional cognitive categories. This observation corroborates the original investigators' finding that the clusters are somewhat arbitrary; many genes were found to have high correlation with multiple cluster representatives (Chu et al. 1998). Perhaps it is more useful to determine the closest neighbors of a gene, rather than to seek well defined clusters.

When we choose the largest principal components, we lose information about experiments that explains the remaining variance in the data (5% of the sporulation variability is not explained by the first three components). However, our analysis identifies the variables that should be used for overall classification of genes, and thereby allows investigators to focus on the other, more subtle, variables whose values may be helpful in understanding the differences in gene expression under different conditions.

Acknowledgments

The authors wish to thank Susan Holmes for many thoughtful discussions, and Raynee Chiang for her assistance in web development and visualization. S.R. is supported by NIH training grant GM-07365; J.M.S. is supported by NIH training grant LM-07033. This work was also supported by NIHLM06244, NSF DBI-9600637 and a grant from the Burroughs-Wellcome Foundation.

References

Basilevsky, A. Statistical Factor Analysis and Related Methods, Theory and Applications. John Wiley & Sons; New York, NY: 1994.

Brazma A, Jonassen I, Vilo J, Ukkonen E. Predicting gene regulatory elements in silico on a genomic scale. Genome Research. 1998; 8:1202–1215. [PubMed: 9847082]

Chee M, Yang R, Hubbel E, Berno A, Huang XC, Stern D, Winkler J, Lockhart DJ, Morris MS, Fodor SPA. Accessing Genetic Information with High-Density DNA Arrays. Science. 1996; 274:610–614. [PubMed: 8849452]

Chen JJW, Wu R, Yang PC, Huang JY, Sher YP, Han MH, Kao WC, Lee PJ, Chiu TF, Chang F, Chu YW, Wu CW, Peck K. Profiling expression patterns and isolating differentially expressed genes by cDNA microarray system with colorimetry detection. Genomics. 1998; 51:313–324. [PubMed: 9721201]

Cherry JM, Adler C, Ball C, Chervitz SA, Dwight SS, Hester ET, Jia Y, Juvik G, Roe TY, Schroeder M, Weng S, Botstein D. SGD : Saccharomyces Genome Database. Nucleic Acids Research. 1998; 26:73–39. [PubMed: 9399804]

Chu S, DeRisi J, Eisen M, Mulholland J, Botstein D, Brown PO, Herskowitz I. The transcriptional program of sporulation in budding yeast. Science. 1998; 282:699–705. [PubMed: 9784122]

Craig JC, Eberwine JH, Calvin JA, Wlodarczyk B, Bennett GD, Finnell RH. Developmental expression of morphoregulatory genes in the mouse embryo: an analytical approach using a novel technology. Biochem Mol Med. 1997; 60(2):81–91. [PubMed: 9169087]

D'Haeseleer P, Wen X, Fuhrman S, Somogyi R. Linear modeling of mRNA expression levels during CNS development and injury. Pacific Symposium on Biocomputing. 1999; 4:41–52.

Duggan DJ, Bittner M, Chen Y, Meltzer P, Trent JM. Expression profiling using cDNA microarrays. Nature Genetics. 1999; 21:10–14. [PubMed: 9915494]

Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A. 1998; 95:14863–8. [PubMed: 9843981]

Everitt, BS. Cluster Analysis. John Wiley & Sons; New York, NY: 1993.

Everitt, BS.; Dunn, G. Applied Multivariate Data Analysis. Oxford University Press; New York, NY: 1992.

Hilsenbeck SG, Friedrichs WE, Schiff R, O'Connell P, Hansen RK, Osborne CK, Fuqua SAW. Statistical Analysis of Array Expression Data as Applied to the Problem of Tamoxifen Resistance. J Natl Cancer Institute. 1999; 91:453–459.

Lander ES. Array of hope. Nature Genetics. 1999; 21:3–4. [PubMed: 9915492]

Liang S, Fuhrman S, Somogyi R. REVEAL, a general reverse engineering algorithm for inference of genetic network architectures. Pacific Symposium on Biocomputing. 1998; 3:18–29. [PubMed: 9697168]

Michaels GS, Carr DB, Askenazi M, Fuhrman S, Wen X, Somogyi R. Cluster analysis and data visualization of large-scale gene expression data. Pacific Symposium on Biocomputing. 1998; 3:42–53. [PubMed: 9697170]

Pearson K. On Lines and Planes of Closest Fit to Systems of Points in Space. Phil Mag. 1901; 2:559–572.

Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science. 1995; 270:467–470. [PubMed: 7569999]

Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Fucher B. Comprehensive Identification of Cell Cylce-regulated genes of the yeast *saccharomyces cerevisiae* by microarray hybridization. Molecular Biology of the Cell. 1998; 9:3273–3297. [PubMed: 9843569]

Vohradsky J, Li XM, Thompson CJ. Identification of procaryotic developmental stages by statistical analyses of two-dimensional gel patterns. Electrophoresis. 1997; 18(8):1418–28. [PubMed: 9298656]

Wen X, Fuhrman S, Michaels GS, Carr DB, Smith S, Barker JL, Somogyi R. Large-scale temporal gene expression mapping of central nervous system development. Proc Natl Acad Sci. 1998; 95:334–339. [PubMed: 9419376]

Xu L, Ajimura M, Padmore R, Klein C, Kleckner N. NDT80, a meiosis-specific gene required for exit from pachytene in Saccharomyces cerevisiae. Mol Cell Biol. 1995; 15:6572–6581. [PubMed: 8524222]

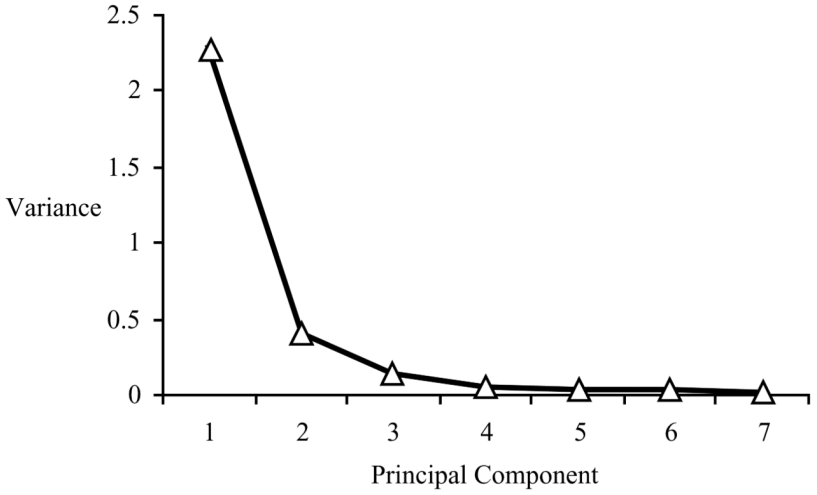

Figure 1.
Plot of eigenvalues of the principal components. Most of the variance in the sporulation data set is contained in the first two principal components.

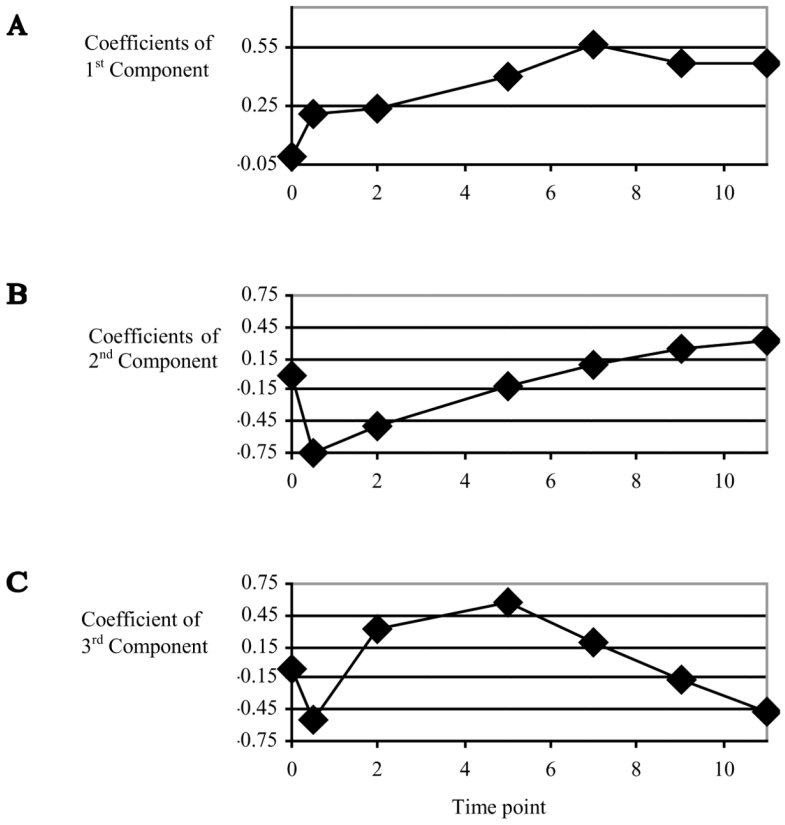

Figure 2.

Plots of the coefficients of the first three principal components. Each coefficient indicates the weight of a particular experiment in the principal component. The first principal component has all positive coefficients, indicating a weighted average. The second principal component has negative values for the early time points and positive values for the latter time points, indicating a measure of change in expression. The third coefficient captures information about the concavity in the expression pattern over time.

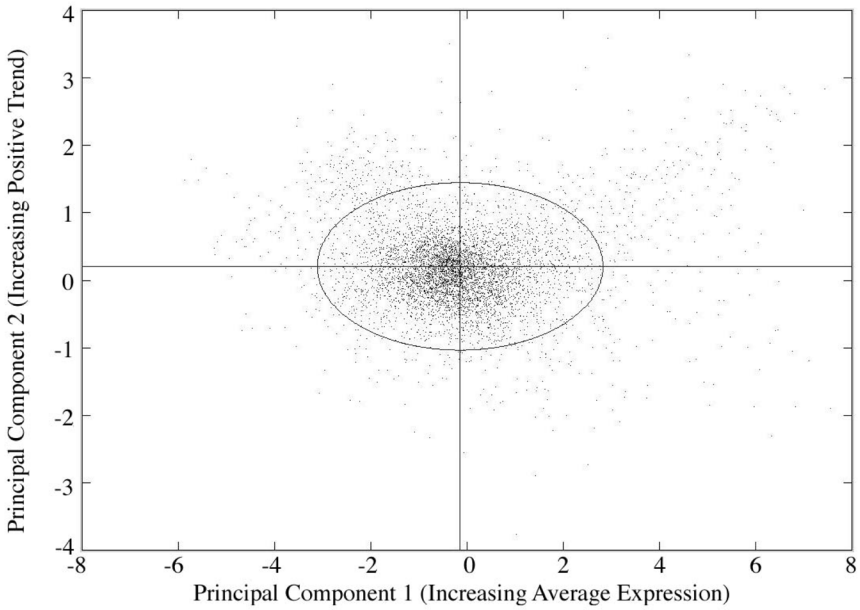

Figure 3.

The rotated and dimensiorially reduced expression data. All yeast genes are plotted on to the first and second principal components. The first principal component is a measure of total average expression, the second is a measure of increasing expression with respect to time. The ellipse at the center contains 95% of the genes.

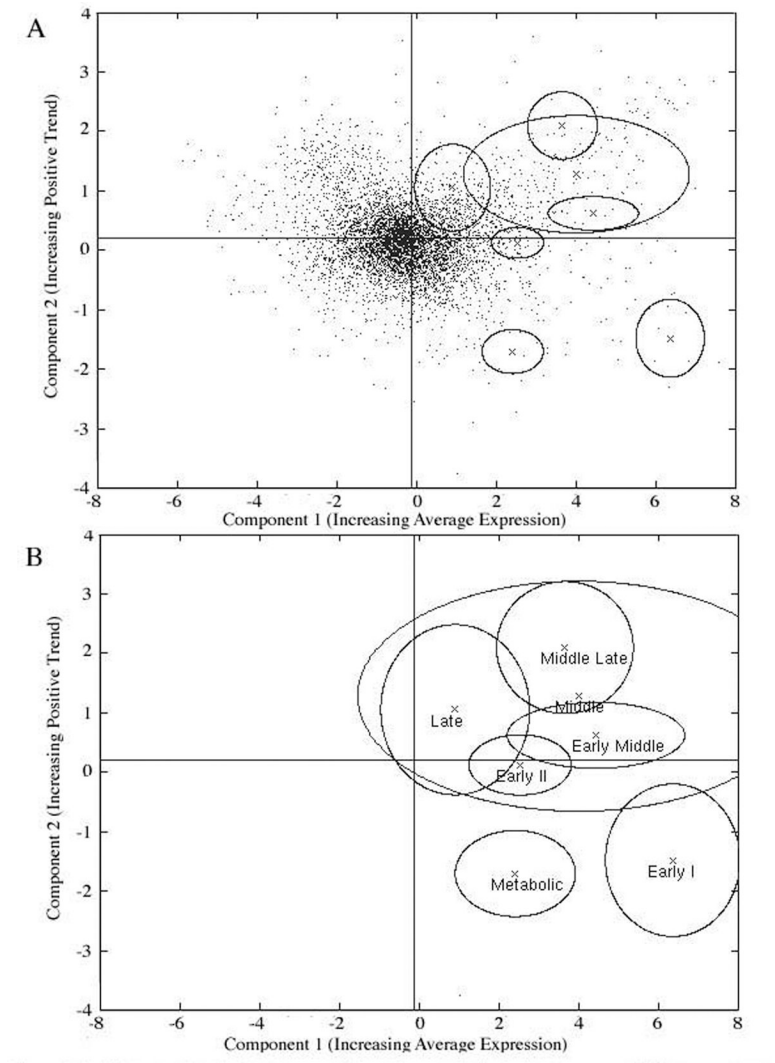

Figure 4.

A. All genes plotted with respect to first and second principal components. Ellipses represent clusters identified in the original publication of the sporulation data. Ellipses are drawn to include 68% of the genes in the cluster. B. Ellipses are labelled using labels reported by the original investigators (Chu et al. 1998) and drawn to include 95% of genes in the cluster.

Table 1

Summary of the experimental data collected by Chu and his colleagues (1998). The table contains average relative expression ratios after application of a natural log transform.

Time point	T=0	T=.5	T=2	T=5	T=7	T=9	T=11
Median	−0.122	−0.182	−0.104	−0.166	−0.095	−0.104	−0.131
Mean	−0.119	−0.214	−0.096	−0.119	−0.007	−0.032	−0.025
Variance	0.029	0.369	0.269	0.428	0.737	0.552	0.596

Table 2

Results of PCA on the sporulation time series data. The values in the columns are coefficients of the principal components that are related to each of the experimental time points. The eigenvalues express the variance of a principal component over all genes. Principal component 1 and 2 contain over 90% of the total variance in the data.

	Principal Components						
Projection On condition	**1**	**2**	**3**	**4**	**5**	**6**	**7**
T = 0	−0.0072	−0.0116	−0.0631	−0.2166	0.0764	−0.7433	0.625
T = .5	0.2076	−0.7524	−0.5373	0.2606	0.1545	−0.0683	−0.0756
T = 2	0.2358	−0.4925	0.3296	−0.5935	−0.453	0.1713	0.0803
T = 5	0.3975	−0.1156	0.5612	−0.002	0.5919	−0.2532	−0.3151
T = 7	0.554	0.0862	0.1869	0.4959	−0.1112	0.2889	0.5559
T = 9	0.4671	0.2517	−0.153	0.1169	−0.5413	−0.4488	−0.4324
T = 11	0.4671	0.3273	−0.4748	−0.5229	0.3307	0.254	0.044
Eigenvalue	2.2928	0.401	0.1322	0.0594	0.0406	0.0288	0.025
% variance	76.9 %	13.5 %	4.4 %	2.0 %	1.4 %	1.0 %	0.8 %