A multivariate approach applied to microarray data for identification of genes with cell cycle-coupled transcription

Daniel Johansson†, Petter Lindgren† and Anders Berglund∗

Research group for Chemometrics, Organic Chemistry, Department of Chemistry, Umeå University, Sweden

Received on June 9, 2002; revised on September 13, 2002; accepted on October 6, 2002

ABSTRACT
We have analyzed microarray data using a modeling approach based on the multivariate statistical method partial least squares (PLS) regression to identify genes with periodic fluctuations in expression levels coupled to the cell cycle in the budding yeast, Saccharomyces cerevisiae. PLS has major advantages for analyzing microarray data since it can model data sets with large numbers of variables and with few observations.

A response model was derived describing the expression profile over time expected for periodically transcribed genes, and was used to identify budding yeast transcripts with similar profiles. PLS was then used to interpret the importance of the variables (genes) for the model, yielding a ranking list of how well the genes fitted the generated model. Application of an appropriate cutoff value, calculated from randomized data, allows the identification of genes whose expression appears to be synchronized with cell cycling. Our approach also provides information about the stage in the cell cycle where their transcription peaks.

Three synchronized yeast cell microarray data sets were analyzed, both separately and combined. Cell cycle-coupled periodicity was suggested for 455 of the 6,178 transcripts monitored in the combined data set, at a significance level of 0.5%. Among the candidates, 85% of the known periodic transcripts were included. Analysis of the three data sets separately yielded similar ranking lists, showing that the method is robust.

Contact: anders.berglund@chem.umu.se
Supplementary material: Available at: http://www.chem.umu.se/dep/orgchem/forskning/chemometrics/bioinformatic.stm

INTRODUCTION
In the past decade, DNA and oligonucleotide microarray techniques (Fodor et al., 1993; Lander, 1999; Schena et al., 1995) have been developed that allow gene expression levels to be monitored on a genome-wide scale. Extracting valuable biological insights from such massive amounts of information is not a trivial task, but statistical methods like cluster analysis can be applied to identify internal structure in the data. For example, similarities of expression profiles among different genes can be used to seek genes that are expressed in the same tissue at the same time, i.e. genes that are co-regulated. Although this type of cluster analysis has proven to be useful for reducing the complexity of large data sets and for identifying predominant patterns of changes in expression levels, additional methods for extracting important features about individual genes from these large data sets are needed. Another valuable step when analyzing genome-wide expression is to visualize the results from the statistical analysis in a way that facilitates interpretation of the data (Bassett et al., 1999).

Traditional statistical methods such as ordinary least squares for classification do not work when there are more variables than there are samples. Gene expression data from microarrays characteristically have many measured variables (genes) and only a few observations (experiments). Problems that have been typically associated with analysis of the results from microarray experiments are missing values and noisy experimental data with low reproducibility. When data with similar complications have been analyzed in the field of chemometrics, the multivariate statistical method partial least squares (PLS) (Wold et al., 1984) regression has been valuable, since it can handle data sets with many, noisy, collinear variables, and with missing values. Furthermore, no assumption about the error distribution is made. This suggests that PLS should be well suited for analyzing microarray data.

Our objective was to investigate whether PLS could be conveniently applied to the interpretation of microarray data. To evaluate the utility of this approach, we aimed to identify periodically transcribed genes of the budding yeast Saccharomyces cerevisiae from publicly available data. These data, downloaded from the public domain site:
http://cellcycle-www.stanford.edu have previously been analyzed in a number of different ways, including: visual inspection (Cho et al., 1998), Fourier transformation (Spellman et al., 1998), self-organizing maps (Tamayo et al., 1999), k-means clustering (Tavazoie et al., 1999), single-pulse modeling (Zhao et al., 2001), QT-clustering (Heyer et al., 1999), singular value decomposition (Alter et al., 2000; Holter et al., 2000), correspondence analysis (Fellenberg et al., 2001) and wavelet analysis (Klevecz, 2000).

Our main goals were to develop a method that allowed ranking of genes with respect to the likelihood that their transcription is coupled to the cell cycle, and to identify stage(s) in the cell cycle where peak transcription of coupled genes occurs. Important additional steps included the determination of an appropriate cutoff value, giving a reasonable number of candidate genes, with an acceptably low number of false positives, and development of a method for visualizing the data to facilitate interpretation.

METHODS

Multivariate statistical methods

Multivariate data analyses can extract information from large data sets containing observations related to a wide range of variables. Principal component analysis (PCA) (Joliffe, 1986) and PLS are two multivariate projection methods that can handle problems associated with most microarray data such as missing values, the presence of more variables than observations, and noise.

PCA is the oldest and best known of the multivariate projection techniques. The central idea of PCA is to reduce the dimensionality of a data set, X, while retaining as much as possible of the variation present in the data. The reduction is accomplished by introducing a new set of variables, the principal components, which are linear combinations of the original variables and uncorrelated to each other. The principal components can be determined using the NIPALS algorithm (Wold, 1966) or by singular value decomposition (SVD) (Joliffe, 1986). PLS is a multivariate regression method that relates the data matrix X to a y-response that can be either single (y) or multiple (Y). The PLS theory and methods discussed in this report concern single y-responses. As in PCA, principal components are constructed to reduce the dimensions of X. In order to obtain the principal components, PLS maximizes the covariance between the response variable y and a linear combination of the original variables t = Xw, where t is the score vector, X is the data matrix and w is the weight vector, for a more in-depth description of PLS please see (Burnham et al., 1999; Höskuldsson, 1996; Martens and Naes, 1989) and references therein. Traditionally, PLS is used to generate a model that predicts y from X. Here, however, an interpretative PLS approach is used. The response is constructed to represent cyclic behavior with the same periodicity as the cell cycle. Thus, the variables/genes that contribute significantly to the models have expression patterns that appear to be coupled to the cell cycle (here described, for convenience, as ‘cyclically expressed’). These genes include those that regulate cell cycling as well as those that are regulated by it.

Choice of response

Visual inspection of the expression profiles of genes known to be cyclically expressed showed that they resemble sine curves, and there is evidence that sine functions accurately model cell cycling patterns (Alter et al., 2000). Since the genes are being expressed in different stages of the cell cycle the sine curves will have different phases. If we assume that all such cyclically expressed genes are characterized by purely periodic expression data, say, of the form \( \sin(\omega t + \phi) \), with a constant frequency \( \omega \) but different phases, \( \phi \), then only two contributing models, \( \sin(\omega t) \) and \( \cos(\omega t) \), are needed to model all cyclic gene expression profiles.

Interpretation of variable importance

The parameter used for interpreting the importance of variables is the weight vector of the first principal component, \( w_1 \). This vector is calculated to maximize the estimated covariance of \( Xw_1 \) and y. Genes that have a large \( w \)-value, negative or positive, will covary strongly with y. Therefore they are important for the model and, in our case, genes whose expression is coupled to the cell cycle. For a single vector, y, it was recently suggested by Trygg (Trygg and Wold, 2001), that \( w_1 \) should contain more useful interpretational information than the more commonly used regression coefficients, \( \beta \), since the regression coefficients are affected by information in X that is not related to y while the first \( w \) vector does not.

Ranking the genes and cut-off value

For a single y-response, two models are needed, one for the sine curve (y1) and one for the cosine curve (y2). A cyclically expressed gene with a similar expression profile to the sine curve will score highly in the first model, and poorly in the second. To be able to detect such genes from all phases in the cell cycle the weight vector of both models has to be considered. This is accomplished by plotting the weight vector, \( w_1 \), for the two models against each other. Each gene will be represented by one point in the plane, where the length d from the origin of the coordinates to the point is a measure of how strong a candidate the gene is for being coupled to the cell cycle (Fig. 1). Genes with cyclic expression are positioned far away from the center, while genes with non-cyclic profiles are positioned close to the center, since they will not fit
Microarray data for identification of genes

Fig. 1. Ranking and cell cycle phase assignment. Genes far away from the center are likely to be coupled to the cell cycle, and genes with similar expression profiles will be clustered together. Genes A and B are ranked according to the distances $d_A$ and $d_B$, respectively, and the cell cycle phase assignments are determined by the angles $\theta_A$ and $\theta_B$, respectively.

Our model. The angle between the coordinate axis and the vector indicates the cell cycle phase assignment.

A ranking list was obtained according to the weights of the genes, in which genes with a cyclic expression profile are highly ranked and others have low ranking. A cut-off value giving an acceptably low proportion of genes that are not cyclically expressed as false positives, while correctly identifying as many as possible of the genuine cyclically transcribed genes, was sought.

By calculating a PLS model for a data matrix of non-cyclic expression profiles and comparing the results with those obtained by the original model, the distribution of genes with non-cyclic expression patterns in our ranking list could be estimated. The data set of non-cyclic genes was simulated by randomizing the order of the observations for each gene in the original data set. In this way, we not only destroyed the cyclic expression profile, but also retained the original variance for the genes. To avoid skews when comparing the results the weight vector is not normalized to unit length as it generally is, since normalization would increase the weights for the simulated genes.

A ranking list for the simulated non-cyclic genes was obtained, and a suitable cut-off value was set. Choosing a cut-off value of level $\alpha$ means that the scores of $(1 - \alpha) \times 100$ percent of the random genes are below the selected value. Applying this value to the ranking list of the original genes will give an approximate frequency of genes that are coupled to the cell cycle (Fig. 2).

The importance of the significance level of the regression can be illustrated by considering the last candidate gene in the ranking list. If the expression profile of this gene is not cyclically expressed the probability of scoring at least this high, purely by random chance, is $(100\% - \alpha)$ percent. That is, the $p$-value for the specific gene is $\alpha$. Conversely, the probability that the gene’s expression is genuinely related to the cell cycle is approximately equal to $(1 - \alpha)$. The smaller the value of $\alpha$, the fewer false positives there will be among the candidate genes, but also the more false negatives there will be among the genes below the threshold. The usual settings for the significance level, $\alpha$, are 0.05, 0.01 or 0.005. The genes above the selected threshold will all have different $p$-values, but the probability that all genes with scores above the threshold have cell-cycle-coupled expression is very small. However, this is not important, since we are prepared to accept a number of false positives in order to detect genes that are genuinely coupled to the cell cycle.
Origin and nature of the data.

Three data sets derived from yeast cultures that were synchronized with three different strategies were used for the analysis, since each of the strategies introduces a characteristic disturbance to the system. For instance, alpha-factor has regulatory effects on genes involved in the mating process. Factors such as choice of carbon source and temperature, as well as the genetic background of the different strains, will also affect the transcriptional profile obtained in such experiments (for details, see Spellman et al. (1998)). The data were accessed from the public domain site: http://cellcycle-www.stanford.edu.

Two of the data sets, ‘alpha’ and ‘cdc15’, were generated in experiments described by Spellman et al. (1998). Mating factor alpha was used to synchronize the cells in the alpha experiment, and samples for RNA analysis were taken every seven min for 140 min following synchronization. Values for samples from time 0 to the 119 min time point (18 observations) were included in the gene expression matrix we accessed. In the second experiment a cdc15 temperature-sensitive strain was used to establish synchrony in the culture by halting the cell cycle in the G2/M-phase. Samples were taken every 10 min for 300 min after releasing the cells. The gene expression matrix used in this study lacked observations for the 0, 20, 40, 60, 260, 280 and 300 min time points and therefore contained 24 sets of observations. Spellman and co-workers used the Stanford approach (Schena et al., 1995) for microarray analysis, and amplified yeast ORFs (open reading frames) to print microarray chips for the alpha and cdc15 experiments.

The third data set was generated by Cho et al. (1998), using a yeast strain with a temperature sensitive mutation in the cdc28 gene to establish synchrony at START (M/G1). Samples were taken every 10 min for RNA measurements. The last samples were collected 160 min after transferring the culture to the permissive temperature, yielding observations from 17 time-points for the cdc28 data set, which was generated using oligonucleotide arrays (Fodor et al., 1993). The cdc28 data were rescaled by Spellman et al. (1998) to mimic the test/control ratios obtained in the alpha and cdc15 experiments.

Model evaluation

In this study we have used two different criteria to evaluate our models. The standard way is to look at how many of the 104 known cell cycle regulated genes that are found among the candidates detected by the method. This yields a good estimate of the false negative rate, assuming that the known cell cycle regulated genes are ranked high. One problem with this assumption is that there are an unknown number of genes that also are cyclic but have not yet been experimentally verified as such. There could be an unknown number of genes that are also cyclic and might score lower than the last known gene. Another issue is that due to experimental errors, some of the known genes do not show a cyclic expression profile even when inspected visually. Setting the cutoff value by a known gene with a low score caused by experimental errors would increase the number of false positives dramatically. A better method for evaluating the results is to consider the number of genes that are suggested to be cyclic in all three data sets when these are analyzed separately. This method is only dependent on the data sets and not on the expression profiles of the known cyclic genes. It also addresses the large problem of getting reliable results from microarray analysis since a robust method would find the same genes independently of data set. The drawback is that there will be some genes that will differ since different techniques for inhibition have been used.

RESULTS AND DISCUSSION

GENERATING THE y-RESPONSE

A prerequisite for establishing a good PLS model is to choose an appropriate y-response, carrying information that helps address the questions posed in a valid way. As mentioned above, two y-response vectors with the forms \( \sin(\omega t) \) and \( \cos(\omega t) \) were generated for each data set. The frequencies (\( \omega \)) used for the models (58, 115 and 85 min) for the alpha, cdc15 and cdc28 data sets, respectively, were suggested by Zhao et al. (2001). The vectors were generated from numerical values of one sine and one cosine function at the observed time points for each of the respective data sets.

PLS ANALYSIS OF THE ALPHA DATA SET

The use of mating factor alpha to obtain a synchronized cell culture affects genes involved in mating responses. Inevitably, this is reflected in the mRNA expression profiles in alpha factor-induced synchronous cultures, particularly for the first couple of observations after the release. A visual inspection of the mRNA profiles indicated that a significant number of genes showed clearly elevated mRNA levels for the first two observations (after 0 and 7 min). However, disregarding these two observations did not significantly affect the quality of the model.

The evaluations of the different models were based on the proportion of the 104 genes known to have cell cycle-coupled expression detected using our cut-off strategy. Data were missing for 103 genes in the alpha data set, one of which was previously shown to be cyclically expressed. Using our cut-off strategy, with a significance level of 0.5%, 178 genes scored above the threshold, including 56 of the previously recorded cyclically expressed genes.
PLS ANALYSIS OF THE CDC15 DATA SET

The cdc15 data set was extended over approximately two and a half cell cycles and sampling covered a 300-min period. The synchronization of the cells in all three analyzed cultures decayed over time, flattening the expression profiles towards the end of the experiment. However, this decay of synchronicity was most severe for the cdc15 data set (Zhao et al., 2001). PLS models with y-response vectors that decayed either linearly or exponentially were also calculated, but they did not yield significantly different results than the initial response vectors. There were missing data for a total of 505 genes, 12 of which were known. Using the same significance level as for the alpha data set, 227 genes appeared to be possibly coupled to the cell cycle and, again, 56 of the previously known genes were among them.

PLS ANALYSIS OF THE CDC28 DATA SET

As mentioned above, the Affymetrix oligonucleotide approach was used by Cho et al. (1998) to generate the cdc28 data set and the data were rescaled by Spellman et al. (1998) to mimic the test/control ratios obtained from the other experiments. As previously reported (Heyer et al., 1999), there are concerns about the reliability of the data gathered at the 90-min time point in the cdc28 data set. In an attempt to improve the PLS model for the cdc28 data set, data related to this point were removed, but no significant difference in the results was observed. Visual inspection of the cdc28 data has suggested 416 genes may be cyclically expressed (Cho et al., 1998). The PLS analysis identified 151 candidate genes, 96 of which were reported by Cho to show periodic behavior in this experiment. However, only 32 of the 104 known cyclically transcribed genes were among the 151 genes scoring higher than the determined threshold. This indicates that the quality of the cdc28 microarray data, is not as good as those from the alpha and cdc15 experiments. This is further discussed below.

Comparing results from the different analyses

Ideally, the same genes would be identified as cyclically expressed in all three data sets. To evaluate the stability of the models and quality of the data, the results from the separate analyses were compared. Forty-two genes (11% of the 367 candidates) were identified as cyclic in all three of the experiments. This can be compared to e.g. 6.5% of 1088 candidates as suggested by Zhao et al. (2001). There are 149 (40%, Zhao et al 23%) genes detected in more than one data set (Fig. 3). In total 367 genes matched the requirements for candidate genes in at least one data set. Among these, 75 of the 104 known genes can be found. Using less restrictive cut-off strategies, Zhao et al. (2001) suggested an SPM-model in which 71 genes were identified as cyclically expressed in all three data sets. If the same numbers of genes as used by Zhao et al. are selected for each data set using our PLS model, 125 genes appear in all three cases.

The two best overlapping data sets were alpha and cdc15, for which almost 40% of the genes identified as cyclically expressed were detected in both data sets. In contrast, the overlaps between the cdc28-ranking list with the alpha and cdc15 lists accounts (in both cases) for only approximately 20% of such genes, indicating that the cdc28 results differ most from the others. This might be due to problems associated with either the oligonucleotide microarray experiments, or the conversion of the oligonucleotide data to resemble the test/control ratios of the other experiments.

COMBINING ALL THREE DATA SETS IN ONE PLS ANALYSIS

To use the information from all three data sets in one PLS model, the y-responses must be synchronized. The cell cultures in the different experiments were not all in the same phase. This means that the phase-determining angle will be not be the same for any specific gene in the different data sets. To make sure that the scores from the different experiments are synergistic rather than antagonistic the y-responses must be shifted to synchrony. The shifts required were calculated using the phase angles from the 42 genes that were classified as cyclically expressed in all three data sets. Average differences between the angles for the genes in the different data sets were calculated and, accordingly, the y-response vectors for the alpha set were shifted 46.3 degrees to the right, whereas
the vectors for cdc15 were shifted 5.5 degrees to the left. One benefit of analyzing all data in one model is that the number of observations is greatly increased. This reduces the risk of getting false positives among the candidate genes. A PLS analysis using the synchronized y-response vectors yielded a ranking list in which 455 genes (including 88 of the previously known cell cycle-coupled genes) scored better than the threshold value with a significance level of 0.5%. Spellman et al chose a cut-off strategy based on the appearance of the 95th known gene and suggested 799 genes to be cyclic whereas, using our approach, the 95th gene appears at position 753. Visualization of the results for the 455 candidates (Fig. 4) shows that the phase determination for the previously known genes are consistent with the phases reported for the respective genes. It is worth noting that the new candidate found within the histone cluster (S-phase) is in fact also a histone (HHO1) that has not previously been characterized as being coupled to the cell cycle. The significance level of 0.5% suggests that ca. thirty-one false positive genes are likely to be included among the 455 candidates. As an extra test, nine genes reported to be cyclically expressed (since publication of the list of 104 known genes cited above) were used to evaluate the model (Doolin et al., 2001; Ho et al., 1999; Rodriguez-Pena et al., 2000). Seven of these were included among the 455 candidates.

A Distance-to-Model on X-block (DmodX) value is a sum of the residuals for one observation to the PLS model, which gives a measure of how well the observation fits the model in the X-space. A high DmodX value indicates that there may have been discrepancies or errors in the experiment that generated the observation. A tendency for the first couple of observations for each data set to yield slightly elevated DmodX values is shown in the plot (Fig. 5). The concerns about the 90-min observation of the cdc28 data set were also confirmed. Furthermore, the observations from the alpha data set give the lowest residuals, suggesting that these data are the most reliable. Attempts to optimize the model by removing outlying observations, trying different cell cycle frequencies, changing missing value tolerance levels or decaying the y-response vectors did not significantly improve the quality of the model.

**CONCLUSIONS**

Our experiments on three microarray gene expression data sets show that PLS works well for interpreting the importance of variables in coordinated gene expression data. It is very important to have a good model, since interpretation of a model that does not contain valuable information is useless. The model must also be stable.
in the sense that small changes in the data sets or the parameters must not substantially distort the results. The $y$-response we chose generates a model that fulfills all these requirements. We have also shown that the number of genes classified as cyclically expressed in more than one data set using our model is larger than when other methods have been used. In addition, our model allows us to identify the phase in which the respective genes are most strongly expressed. Refinements, such as decay of the $y$-response, adjustment of the frequencies and missing value tolerance levels did not improve the model further, indicating that it is robust. We believe that the data sets are insufficiently precise to identify cyclically expressed genes with certainty, and we do not expect that much more information could be extracted from them.

To obtain a cut-off value, a new non-cyclic data set was generated by randomizing the order of the observations for each gene. A PLS model was calculated for the new data set and the results were compared to the calculations for the original data set to estimate the number of false positives. The transcription patterns of 104 known cyclically expressed genes were visually inspected, and evidence for periodicity in at least one data set could be observed for 95 of them. A few of these genes were considered borderline cases. Among our 455 candidate genes, 88 had been previously reported to be periodically transcribed, suggesting that our cut-off is reasonable.

ACKNOWLEDGMENTS

We are grateful for support from the Knut and Alice Wallenberg Foundation and the Swedish Research Council.

REFERENCES