Application of Bayesian Decomposition for analysing microarray data

T. D. Moloshok¹, R. R. Klevecz², J. D. Grant³, F. J. Manion¹, W. F. Speier IV³ and M. F. Ochs¹,*

¹Bioinformatics Working Group, Fox Chase Cancer Center, Philadelphia, PA 19111, USA, ²Department of Biology, Beckman Research Institute of the City of Hope, Duarte, CA 91010, USA and ³Bioinformatics Facility, Fox Chase Cancer Center, Philadelphia, PA 19111, USA

Received on October 1, 2001; revised on October 31, 2001; accepted on November 7, 2001

ABSTRACT

Motivation: Microarray and gene chip technology provide high throughput tools for measuring gene expression levels in a variety of circumstances, including cellular response to drug treatment, cellular growth and development, tumorigenesis, among many other processes. In order to interpret the large data sets generated in experiments, data analysis techniques that consider biological knowledge during analysis will be extremely useful. We present here results showing the application of such a tool to expression data from yeast cell cycle experiments.

Results: Originally developed for spectroscopic analysis, Bayesian Decomposition (BD) includes two features which make it useful for microarray data analysis: the ability to assign genes to multiple coexpression groups and the ability to encode biological knowledge into the system. Here we demonstrate the ability of the algorithm to provide insight into the yeast cell cycle, including identification of five temporal patterns tied to cell cycle phases as well as the identification of a pattern tied to an ~40 min cell cycle oscillator. The genes are simultaneously assigned to the patterns, including partial assignment to multiple patterns when this is required to explain the expression profile.

Availability: The application is available free to academic users under a material transfer agreement. Go to http://bioinformatics.fccc.edu/ for more details.

Contact: td_moloshok@fccc.edu; RKlevecz@coh.org; jd_grant@fccc.edu; fj_manion@fccc.edu; speier1@prodigy.net; m_ochs@fccc.edu

INTRODUCTION

Recent advances in microarray and gene chip technology (Cho et al., 1998; Spellman et al., 1998; Iyer et al., 1999; Ross et al., 2000) have led to an explosion in the amount of data available for understanding cellular function and pathways, with the potential for revealing the underlying cellular behavior responsible for diseases such as cancer (Young, 2000). Since the amount of data is significantly larger than in standard biological studies, new tools are needed for analysis. While progress has been relatively good in the use of data analysis to identify disease states (Golub et al., 1999; Alizadeh et al., 2000), the overall complexity of the data and underlying biological systems will require the application of more advanced methods from computer science to recover the maximal information for knowledge discovery (Lockhart and Winzeler, 2000; Young, 2000).

Many attempts have been made to apply standard data mining algorithms to discover patterns within gene expression array data. These include the use of standard statistical methods (Claverie, 1999; Alter et al., 2000), self-organizing maps (Tamayo et al., 1999), support vector machines (Brown et al., 2000), clustering (Eisen et al., 1998; Lukashin and Fuchs, 2001), and other methods, reviewed in (Brazma and Vilo, 2000). Progress has also recently been made on new statistical methods which maintain non-Euclidean relationships during reduction of the dimensionality of the data space (Roweis and Saul, 2000; Tenenbaum et al., 2000), which may help in defining relationships in complex data. In general, these methods are not designed with knowledge of the underlying biological system behavior, although there have been some recent advances in clustering which do take into account some experimental knowledge (Heyer et al., 1999). However, all of the current methods still lack an ability to recover fundamental behavior because they force each gene within the expression experiment into a single coexpression group, which overlooks the biological fact that many individual genes are coexpressed in multiple groups in response to different stimuli (Roberts et al., 2000). This fundamental limitation impairs their usefulness as it leads to the loss of all information related

*To whom correspondence should be addressed.
to behavior arising from multiple inputs. Such information is critical for understanding cellular behavior (Bittner et al., 2000). For example, if a set of cell cycle regulated genes included a set of genes which were upregulated in G1, a set upregulated in G2, and a set upregulated in G1 and G2, most current algorithms would require three clusters to model the data. However since there are only two cell cycle phases represented, a method which can assign genes multiply to both G1 and G2 will more accurately reflect the biological basis for the gene expression.

One key cellular behavior that becomes difficult to identify if responses to multiple stimuli cannot be disentangled is the identification of the set of interacting signaling pathways that have been activated. This is particularly important when the point of a study is the therapeutic response to a drug, either for prevention or treatment. Many identified causes of cancer development are errors in signaling proteins or sets of signaling proteins with the consequent disruption of the normal operation of the signaling pathways (e.g. p53, abl, c-kit). In order to use gene expression measurements to deduce drug response or to interpret studies aimed at identifying key differences between normal and tumorigenic tissues, a technique of identifying the sets of pathways undergoing changes must be used.

Here we present a new method for analysing gene expression data which attempts to identify changes in sets of interacting pathways from gene expression data by encoding into the system behavior that mirrors the physiological behavior more closely than standard analysis methods. The method is an application of the Bayesian Decomposition (BD) algorithm, originally developed for the identification of spectral signatures in cases where there was mixing of components in all data points (Ochs et al., 1999). This algorithm is a matrix factorization method which, unlike Principal Component Analysis (PCA) and other widely used methods, identifies physically meaningful, nonorthogonal basis vectors describing the data. For gene expression time series data, such basis vectors are the time curves associated with a physical process (such as progression through the cell cycle or the activation of a pathway in response to a drug treatment). The algorithm simultaneously identifies the basis vectors and their distributions within the data. The distributions allow the algorithm to ascribe the behavior of each gene to multiple patterns, so that genes that are transcribed in response to multiple stimuli can be identified as belonging to multiple coexpression groups, e.g. activation of multiple pathways at different times in the cell cycle.

**ALGORITHM AND IMPLEMENTATION**

BD is a matrix factorization algorithm which uses prior knowledge in the form of mathematical models of the underlying physical process to identify the basis vectors of the data in their minimal, physically significant form. In our initial work (Ochs et al., 1999), the physical model included basis vectors restricted to sets of positive spectral components with Gaussian or Lorentzian lineshapes. In subsequent work (Ochs et al., 2001), the physical model was the known functional form of magnetization recovery following inversion in magnetic resonance relaxographic imaging (Labadie et al., 1994). For gene expression analysis, the basis vectors are the coexpression response due to individual pathways or linked pathways that respond in a correlated way to stimuli.

The fundamental decomposition performed by BD is the recovery of a distribution matrix (A) and a pattern matrix (P) which combine to form a mock data matrix (M) that reproduces the data matrix (D) within the noise limits, as shown in Figure 1. For gene expression analysis, each row of D represents the expression of a single gene with the columns representing different measurements (different times, different tissues, different individuals, etc.). The matrix M would match the matrix D exactly if A and P were perfect models of the system and if there were no noise, i.e.

\[ D = AP. \]

The distribution matrix A contains rows that describe the amount (amplitude) of each pattern within the corresponding row (gene) in D, with each column being associated with a single pattern. The rows of P are the patterns that show the average behavior of the coexpressed genes across the experiment (e.g. over time, across tissues, across individuals, etc.). For example, a row of P in a cell cycle experiment could contain the time behavior of gene expression representing upregulation related to a single cell cycle phase, so that the rows of A would indicate when a single gene was upregulated, perhaps in multiple phases (i.e. significant values in multiple columns). This allows BD to identify multiple coexpression groups for each gene, which permits the correct identification of a gene as being transcribed in response to multiple different stimuli. This is the key feature of the mathematical model, as it permits the identification of overlapping coexpression groups, which could confuse simpler algorithms that force each gene into a single group.

The mean model for the data is determined by a Markov chain Monte Carlo process using Bayesian estimates of probability (Besag et al., 1995). From Bayes’ equation and (1), the relative probability at each point in the Markov chain is given by

\[ p(A, P|D) = p(D|A, P)p(A, P) \]

where the posterior probability distribution, p(A, P|D), is the probability of the model (A and P matrices at that point) given the data, p(D|A, P) is the likelihood, and p(A, P) is the prior, which gives an indication of
the probability of the model independent of the data (discussed below). The system creates a first model randomly out of the vacuum in accordance with the prior and then attempts changes to the model, determining the suitability of each change by comparison of the resulting \( M \) matrix (the mock data resulting from the model) to the \( D \) matrix (the data) through computation of the change in the likelihood. For a change in \( P \) of \( \delta P \), the change in the likelihood is given by

\[
\Delta L(\delta P) = \frac{1}{2\sigma^2} \text{Tr}(A\delta P^T(A P - D) + (AP - D)^T A\delta P + (A\delta P)^T(A\delta P))
\]

(3)

where \( \text{Tr} \) indicates the trace, \( ^T \) indicates the transpose, and equal uncertainty \( \sigma \) is assumed here at each point for clarity in writing the equation. There is a corresponding equation for changes in \( A \). The algorithm uses simulated annealing (Kirkpatrick et al., 1983; Geman and Geman, 1984) to minimize the possibility of entrapment in areas of false maxima in likelihood. The simulated annealing process modifies (2) to

\[
p(A, P | D) = p(D | A, P)^S p(A, P)
\]

(4)

where \( S \) is progressively raised from 0 to 1, effectively increasing the power of the data to determine the model during annealing. Sampling occurs after annealing with \( S = 1 \), so that the posterior probability distribution (i.e. (2)) is used.

The algorithm encodes an atomic prior that seeks the minimal structure required to fit the data, therefore minimizing the number of points requiring computation (Sibisi and Skilling, 1997). The prior is enforced within the atomic domain (see Figure 2), where changes to the model are generated. The atomic domain consists of point masses distributed along an infinitely divisible line (in reality two 32 bit ints, one for \( A \) and one for \( P \)). These point masses are then mapped into the model domain (\( A \) and \( P \)), which allows incorporation of additional prior knowledge by the use of convolution functions (\( f_s \) in Figure 2). The minimizing of structure is important as the algorithm has computational complexity of \( O(n \log n) \) in the number of point masses. In the original work with BD (Ochs et al., 1999), the prior knowledge encoded was that a point mass in the \( P \) atomic domain defined a full spectral line along a row of \( P \) in the model domain (i.e. in one pattern). For this work we encode only the positivity of gene expression (i.e. the ratio of the relative amounts of mRNA between experiment and control is a positive value) at this stage, allowing all possible curves. Finally the data and mock data generated from the model are compared by calculation of the change in likelihood given in (3), which permits full utilization of knowledge of the uncertainties associated with each data point. The change in this likelihood is used to determine the probability of accepting the randomly generated change to the model.

For this work there is a direct mapping from the atomic domain into the model domain, so that the convolution functions (\( f_s \) in Figure 2) take the atom flux and place it fully into a single matrix element. The prior here includes only the requirement of positivity and additivity, which still effectively reduces the search space by \( 2^N \) where \( N \) is the number of dimensions, since only the positive hyperquadrant is allowed for the model. However, the mathematical structure allows for multiple coexpression of genes, since each gene may be assigned behavior arising from different patterns (e.g. expression in response to drug treatment and in response to stress). The simple mapping here makes this process similar to nonnegative matrix factorization (Lee and Seung, 1999). However, BD is more flexible in that it allows complex correlations to be mapped and is limited to positivity only in the atomic domain.

BD requires an estimate of the number of dimensions required to fit the data as an input parameter (i.e. the number of columns in \( A \) and rows in \( P \)). This can be provided by estimation of the number of dimensions needed to fit the data through statistical analysis (such
as PCA). In addition, BD is usually run using a number of different estimates. For many analyses specific features appear indicating that the number of patterns is excessive. In spectroscopic studies and time domain modelling, these are often the emergence of patterns which appear to be unrelated to spectral features or to natural time behaviors. In data sets which have no likely correlated structures between points, the estimation of the dimensionality is more problematic.

A primary feature of (1) is that it is mathematically degenerate, permitting multiple solutions. BD searches through the possible solutions for those which are most probable and samples the likely solutions. In general, if the number of elements in \( D \) is significantly larger than the number of elements in \( A \) and \( P \) combined or if there is a good mathematical model of the data, the sampler will visit multiple representations of the same solution, yielding a mean and standard deviation for each element of the matrix. In cases where there are multiple, significantly differing solutions, the sampler can move between these. This will generally yield cases where there are significant uncertainties associated with the points, however BD saves snapshots of individual solutions which can be examined to verify that multiple solutions exist. In addition, by repeating the analysis using different random seeds in the Markov process, different Markov chains are generated and the results can be compared. This reduces the probability that the results represent one solution out of many which might fit the data equally well.

**METHODS**

To test the ability of BD to recover biologically significant information from microarray data, we analysed the \( cdc28 \)-mutant yeast cell cycle data (Cho et al., 1998) as summarized in the Stanford cell cycle data set (Spellman et al., 1998), containing 763 genes with cell cycle periodicity. Since the data presented in the Stanford data set are \( \log_2 \) measurements of experimental to control expression, the data were transformed to a linear scale, giving ratios that are always positive. BD requires a calculation of the likelihood, so that uncertainty (or noise) estimates for each point in the data were needed. Since it appears that the noise level is often tied more closely to the gene expression level, rather than being an absolute level across all genes, we used a noise estimate of 25% of signal. In addition, the data set had data missing at many time points. For these points we set the expression ratio to 1.0, but set the noise estimate to 100, so that BD would treat the models as unconstrained by these data points.

We performed 11 separate runs of BD with different random seeds, which is equivalent to different random starting points for the Markov chain. Standard statistics were calculated on the results, yielding an estimate of the uncertainty for each point in the pattern and for the distribution of the patterns for each gene in the data set. The results of these runs were then compared to the results reported elsewhere. In addition, specific sets within the data were analysed, including genes encoding known kinases which often are activated at multiple points in the cell cycle. These data were interpreted in light of information within the Saccharomyces Genome Database (SGD; Cherry et al., 1997), the Curagen Pathcalling Database (CPD; Uetz et al., 2000), and the Proteome Yeast Database (YPD; Costanzo et al., 2000, 2001).

The patterns of time expression behavior from BD were further analysed by wavelet decomposition (Daubechies, 1992) at all permitted scales as described previously (Klevecz, 2000). In the six gene expression patterns, the periodicities that are apparent in the profiles were confirmed by the wavelet scaling. The decompositions were done using the Daubechies daublet \(-6\) transform.

**RESULTS**

Figure 3 shows the average gene expression patterns together with error bars derived from the 11 repeated runs of BD. The identification of the time curves with the cell cycle phases is clear, as is the existence of a single solution across the multiple runs. Specifically the solid orange curve corresponds with high expression in the M/G1 phase.
Fig. 3. The patterns determined by BD for the yeast cell cycle data. There are five patterns tied to the cell cycle and one additional oscillatory pattern. In order of peaks, the cell cycle patterns are tied to M/G1 (solid orange), G1 (G1A, solid black), S/G2 (dashed orange), M (dotted orange), and G1 (G1B, dashed black). The oscillator pattern (osc) is shown in red. The two G1 patterns differ primarily in the presence of expression during the first cell cycle upon release from cell cycle arrest.

Fig. 5. Modelling of expression by BD in complex cases. In each case, the mock data created from the model by BD is compared to the original data, with the mock data in orange and the original data in red. The genes shown are (a) ISR1; (b) RAD53; (c) HSL7; (d) CDC5; (e) DBF20; (f) GIN4.

transition, the solid black curve with high expression in the G1 phase (including the first pass through the cell cycle, G1A group), the dashed orange curve with high expression in S and G2 phases, the dotted orange curve with high expression in the M phase, and the dashed black curve with high expression in the G1 phase (but not during the
Table 1. All genes with more than 70% of their behavior explained by the oscillator pattern which are also identified within the Proteome YPD database

<table>
<thead>
<tr>
<th>Gene name</th>
<th>YPD</th>
<th>Proteome role</th>
<th>Proteome function</th>
</tr>
</thead>
<tbody>
<tr>
<td>YOL0T6C</td>
<td>CMK2</td>
<td>Mating response [E]</td>
<td>Protein kinase [P]; transferase [P]</td>
</tr>
<tr>
<td>YEL060C</td>
<td>PRB1</td>
<td>Protein degradation [E]</td>
<td>Protein folding [E]; protein modification [E]; protein degradation [E]</td>
</tr>
<tr>
<td>YOL040W</td>
<td>UTR2</td>
<td>Cell wall maintenance [E, P]</td>
<td>Serine rich protein [P]; hydrolase [P]</td>
</tr>
<tr>
<td>YOL060W</td>
<td>ARG1</td>
<td>Amino-acid metabolism [E]; other metabolism [E]</td>
<td>Ligase [E]</td>
</tr>
<tr>
<td>YOR116C</td>
<td>YOR115C</td>
<td>Vesicular transport [E]</td>
<td>Unknown</td>
</tr>
<tr>
<td>YOR024C-A</td>
<td>PMP1</td>
<td>Protein synthesis [E]</td>
<td>tRNA synthetase [E]; ligase [E]; RNA-binding protein [E]</td>
</tr>
<tr>
<td>YOL196W</td>
<td>GIN1</td>
<td>Amino-acid metabolism [E]; protein synthesis [P]; other [E]</td>
<td>Activator [E]</td>
</tr>
<tr>
<td>YMR202W</td>
<td>ERG2</td>
<td>Lipid, fatty-acid and sterol metabolism [E]; vesicular transport [E]</td>
<td>Isomerase [E]</td>
</tr>
<tr>
<td>YML183C</td>
<td>PHO84</td>
<td>Phosphat metabolism [E]; small molecule transport [E, P]</td>
<td>Major facilitator superfamily [P]; transporter [E]; active transporter, secondary [P]</td>
</tr>
<tr>
<td>YCL043C</td>
<td>PDI1</td>
<td>Protein folding [E]; protein modification [E]; protein degradation [E]</td>
<td>Oxidoreductase [E]; isomerase [E]; chaperones [E]</td>
</tr>
<tr>
<td>YOR323C</td>
<td>PRO2</td>
<td>Amino-acid metabolism [E]</td>
<td>Oxidoreductase [E]</td>
</tr>
<tr>
<td>YOR283W</td>
<td>YOR283W</td>
<td>Carbohydrate metabolism [P]</td>
<td>Hydrolase [P]</td>
</tr>
<tr>
<td>YMR246W</td>
<td>FAA4</td>
<td>Lipid, fatty-acid and sterol metabolism [E]; vesicular transport [E]</td>
<td>Ligase [E]</td>
</tr>
<tr>
<td>YGR124W</td>
<td>ASN2</td>
<td>Amino-acid metabolism [E]</td>
<td>Ligase [P]</td>
</tr>
<tr>
<td>YML058W</td>
<td>YML058W</td>
<td>Nucleotide metabolism [E]; DNA repair [E]; Chromatin/chromosome structure [E]</td>
<td>Regulatory subunit [P]</td>
</tr>
<tr>
<td>YIL078W</td>
<td>SEC28</td>
<td>Vesicular transport [P]</td>
<td>Vesicle coat protein [P]</td>
</tr>
<tr>
<td>YJR004C</td>
<td>SAG1</td>
<td>Mating response [E]; cell adhesion [E]</td>
<td>Adhesive/agglutinin [E]</td>
</tr>
<tr>
<td>YHR208W</td>
<td>BAT1</td>
<td>Amino-acid metabolism [P]</td>
<td>Transferase [E]</td>
</tr>
<tr>
<td>YOA077W</td>
<td>SED1</td>
<td>Cell structure [E]</td>
<td>Unknown</td>
</tr>
<tr>
<td>YLR467W</td>
<td>YLR467W</td>
<td>Chromatin/chromosome structure [P]</td>
<td>Hydrolase [P]; helicase [E]; ATPase [P]; DNA-binding protein [E]</td>
</tr>
<tr>
<td>YLR342W</td>
<td>GLS1</td>
<td>Carbohydrate metabolism [E]; cell wall maintenance [E]</td>
<td>Transferase [E]</td>
</tr>
<tr>
<td>YCR002C</td>
<td>CDC10</td>
<td>Cell polarity [E, P]; cytokinesis [E]</td>
<td>Hydrolase [P]; GTP-binding protein/GTPase [P]; structural protein [E]</td>
</tr>
<tr>
<td>YPL323W</td>
<td>SSO1</td>
<td>Vesicular transport [E]; membrane fusion [E]</td>
<td>Docking protein [E]</td>
</tr>
<tr>
<td>YOR332W</td>
<td>VMA4</td>
<td>Small molecule transport [E, P]</td>
<td>Hydrolase [P]; ATPase [P]; Transporter [P]; active transporter, primary [P]</td>
</tr>
</tbody>
</table>

For each gene its gene name from the SGD is given, together with its YPD identification, the role from the YPD, and the function from the YPD. In the role and function columns a [P] indicates a predicted function and an [E] indicates an experimentally verified function.

first pass through the cell cycle after arrest, G1B group. In addition, there is a sixth curve shown in red which has a roughly constant level of expression but with a small oscillator-like behavior superimposed (osc). Wavelet decomposition of the oscillator data finds a periodicity of ~40 min. The ~40 min cycle is consistent with published reports on metabolic oscillations in S. cerevisiae (Satroutdinov et al., 1992). It is also in agreement with a genome-wide wavelet analysis of these data sets in which both ~40- and ~80-min periodicities were uncovered in the expression profiles of the majority of the genes (Klevecz, 2000; Klevecz and Dowse, 2000). A selection of the genes which are most strongly associated with this pattern are shown in Table 1. Most appear to be metabolic genes which would be expected to be involved with routine cellular function.

Figure 4 demonstrates the ability of the algorithm to identify the time behavior of expression (i.e. patterns). In each case, the time behavior determined by BD is shown with a thick solid line. Superimposed on this time curve are the original data (thin lines) and average of the original data (thick dashed line) for the genes which have at least 70% (80% for the G1A and the oscillator patterns) of the time behavior explained by the single pattern shown. These genes would typically be identified as a ‘cluster’ in a clustering algorithm, as they essentially belong to only a
single pattern. The identification of the specific genes for the cell cycle clusters, a–e in Figure 4, are given in Table 2.

Figure 5 demonstrates the capabilities of the BD algorithm to handle cases reflecting complex regulation in response to multiple input signals. Since BD simultaneously identifies patterns (e.g., time behaviors of expression) and specifies how these must be combined to recover a good model of expression for each gene, genes which are multiply coregulated can be properly assigned to multiple patterns (e.g., genes upregulated in G1 and G2 are assigned partially to G1 and partially to G2). This means that rather than creating additional clusters which contain genes which are multiply coregulated, a physiologically meaningful number of patterns can be combined to explain the expression behavior of such genes. In Figure 5 six examples of the expression behavior of kinases are given. The transcriptional response of many kinases is known to be multiply regulated (see, for example, kinases in the SGD; Cherry et al., 1997), so that they are upregulated in multiple parts of the cell cycle. For each gene, the time series data is shown in red and the model data reconstructed by BD is shown in orange. BD has successfully matched the behavior of each gene expression time series using only six patterns by combining different amounts of each of the patterns.

For those genes which are turned on multiple times in the cell cycle, BD is able to correctly assign these genes to multiple cell cycle patterns. For example, in Figure 5a, the expression behavior of ISR1 is shown. ISR1 is clearly upregulated in S, G2, and G1, although its function is not known. BD assigns this gene 56% to the G2/S pattern (Figure 4c) and 30% to the G1 phases (Figure 4a and b). On the other hand, RAD53 (Figure 5b), a G1 checkpoint arrest protein, is clearly strongly tied to the G1 phase of the cell cycle, and BD assigns RAD53 95% to the G1 phases. For each of the expression patterns shown, the match from BD is excellent. As data quality improves, this ability to construct a model that has physiologically meaningful bases should increase in power.

**DISCUSSION**

BD has shown its usefulness in the fields of image analysis (Ochs et al., 2001) and spectroscopic analysis (Ochs et al., 1999). In the analysis presented here, the ability of BD to disentangle gene expression patterns that arise when genes are coregulated at different points within the cell cycle is demonstrated. For genes that are strongly tied to a single aspect of the cell cycle, the results agree in many ways with the recent work on the use of simulated annealing to identify the optimal number of clusters (Lukashin and Fuchs, 2001). In that work, five clusters from the 20 identified genes were linked strongly to the
Table 2. All genes with more than 70% of their behavior explained by a single pattern. Identifications are standard names from the SGD

| G1A       | YNL289W | YER070W | YNL102W | YDL003W | YGR221C | YPL153C | YBR089W | YCL022C | YOL007C | YPL267W | YML027W | YLR103C | YKL113C | YDR097C | YLR313C | YKR013W | YJL074C | YBL055C | YOR074C | YAR008W | YBR088C | YMR078C | YFR027 |
|-----------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| G1B       | YGR152C | YLR235C | YML060W | YIL026W | YBR070C | YFL008W | YNL273W | YHR154W | YNL312W | YOL090W | YHR153C | YML027W | YMR102W | YOL017W | YCL061C | YAR007C | YCL060C | YLL022C | YIL187C | YLR338W | YMR076C | YMR078C |
| S/G2      | YLR067C | YOR263C | YHR114W | YIL309W | YGR188C | YPL241C | YJL181W | YPR054W | YER095W | YLR234W | YNL173C | YMR102W | YDL018C | YCL061C | YAR007C | YCL060C | YL096C  | YLR326W | YPR256C | YGR109C | YCR065W |
| M         | YPL021W | YMR003W | YBL063W | YPL265W | YGR089C | YKL178C | YPL061W | YJR010W | YDR380W | YNL173C | YYL040W | YDL179W | YOR138C | YOR025W | YAR007C | YER111C | YER022C | YMR096C | YGR065W | YDR507C | YNL300W |
| M/G1      | YPR157W | YPR156C | YGL116W | YLR190W | YOR073W | YDR451C | YOR314W | YOL142W | YHR315W | YMR315W | YBR138C | YGR143W | YIL051W | YOR025W | YOL017W | YLR103C | YMR076C | YMR076C | YMR076C | YMR076C | YMR076C |

cell cycle, similar to the five patterns (a–e) of Figure 4. Because BD can mix these patterns together to explain the expression behavior for each gene, it does not require additional patterns in order to model the data. Instead BD provides physiologically meaningful expression patterns which together can describe all expression behavior. In this way, BD eliminates the need for additional clusters not required by the underlying biology, but instead needed only because of the inability of a clustering algorithm to partition the expression behavior of a single gene into multiple groups.

One interesting issue with the five cell cycle clusters identified both here and with the simulated annealing clustering algorithm (Lukashin and Fuchs, 2001) is the lack of a peak in one pattern during the first cell cycle phase. The first peak in the G1A cluster occurs at 20 min and the second at 100 min, while the peak in the G1B cluster occurs at 90–100 min with no corresponding peak at 10–20 min. Reviewing the genes in the G1B cluster (Table 2), it is seen that many genes encode proteins involved in amino acid metabolism and carbohydrate metabolism. The genes in the G1A cluster on the other hand are primarily involved with DNA synthesis and repair, mitosis, cell cycle control, and Pol II transcription. Although speculative, one reasonable interpretation of the absence of a peak at 10–20 min is that the necessary building blocks (amino acids and carbohydrates) had been processed prior to arrest, therefore eliminating the need for the metabolic proteins and for the transcription of the corresponding genes upon exit from arrest. This could be a result of negative regulation of metabolic processes by metabolic products (Rao and Arkin, 2001).

The oscillator pattern identified by BD is of particular interest (Figure 4f). This pattern appears to be tied to routine metabolic processes within the cell as shown by Table 1. It may be that this pattern has not drawn attention previously because in clustering algorithms it appears as one (or several) of many clusters, few of which can be reasonably interpreted due to the problem of multiple coregulation. BD has the advantage of being able to assign multiple coregulation appropriately to multiple patterns, allowing the data to be explained with fewer, interpretable patterns. This allows biologically relevant clusters to be fully examined.

The ability to add additional prior knowledge through a convolution function (f in Figure 2) in BD is a powerful tool which has not yet been exploited in expression analysis. We are presently exploring the addition of
new distribution functions in two ways. First, we are adding models of time behavior of mRNA species during transcription in order to model expression time series data more accurately. This is done by mapping a single atom in the P atomic domain into a full curve (sampled at the times in the experimental data set) in the model domain (P matrix). This involves modelling the possible delays between signal initiation and transcriptional initiation, the rise time of mRNA following initiation of transcription, and the half lives of mRNA species. These models may be too complex in higher organisms with complex transcriptional control, however simplifications may be adequate to provide insight. For simpler organisms such as yeast and bacteria, such modelling is likely to be adequate. Second, we are introducing correlations in the distributions by using known, reliable coregulation (i.e. known common promoter elements) to determine which genes must have their expression levels correlated. Since BD is free to use multiple patterns to explain the expression profile of a single gene and to add additional genes to a coregulated group, this should not introduce excessive bias. Still this method must be used carefully, as it creates a strong prior in favor of the linking of certain genes. With these and future improvements, BD should continue to provide a good tool for gene expression analysis in the future.

ACKNOWLEDGEMENTS

This work was supported by the National Institutes of Health, National Cancer Institute (Comprehensive Cancer Center Core grant CA06927 to R.Young) and the Pew Foundation. This work was also supported in part by a grant from the Beckman Foundation. We would like to thank the referees for helpful suggestions.

REFERENCES


