A Multidimensional Scaling-Based Model for Analysis of Time-Index Biomics Data

Wesner Antoine\(^1,3\) and Jan A. Miernyk\(^1,2,3\)

\(^1\)Department of Biochemistry, 204 Curtis Hall
University of Missouri, Columbia, MO 65211, USA

\(^2\)USDA, Agricultural Research Service
Plant Genetics Research Unit; USA

\(^3\)Interdisciplinary Plant Group, University of Missouri
Columbia, MO 65211, USA
Jan.Miernyk@ars.usda.gov

Abstract

It is critical that the data generated during time-index biomics profiling studies be summarized in a biologically meaningful way; pattern detection techniques and modeling are important tools for this. We describe use of the multidimensional scaling algorithm to detect consensus patterns within clustered time-index data. Retrieved patterns can be used as a reference to describe individual expressions data via a model. The model proposes a profile match index and a level parameter for each individual pattern. A publicly available transcript profiling dataset from developing soybean embryos has been used to illustrate the model. After describing a pattern for each of 11 clusters, the parameters of the model were estimated for each gene. The legitimacy of the consensus pattern for each cluster was assessed relative to a badness of fit value. The profile match index and the individual fit statistics serve to validate assignment of a gene to a cluster, and to isolate outliers. This method allows extraction of meaningful information from any time-index profiling dataset. A description of the gene expression network manifest during physiological change of soybean embryos was developed based on the patterns detected using multidimensional scaling; each distinct pattern was considered as a node.
Keywords: Cluster analysis, data mining, *Glycine max* L., soybean, transcript profile analysis

Introduction

High-throughput biomics profiling methods allow monitoring of the expression of thousand of genes, proteins, or metabolites simultaneously, as they change in response to controllable stimuli [18]. They provide a means to probe the transcriptome/proteome/metabolome for expression patterns indicative of physiologically important events [13,14]. Typically profiling data are analyzed in the context of pathways or networks. The ability to reliably define the time-index changes is critical when reconstructing regulatory networks to reveal the sequence of molecular events that underlie a phenotype. Reconstructing the networks requires tools for discovery, correlation, and pattern interpretation. Techniques typically used in such analyses emphasize a graphics-based information-discovery mode [4,16]. This allows the facile identification of atypical patterns of gene/protein/metabolite expression which can then be targeted for additional detailed analysis.

One of the most commonly used steps in analysis of profiling data is clustering, where the whole data-set is separated into groups that share a similar pattern. Several clustering algorithms have been developed including hierarchical clustering [11], k-means clustering [9], Hidden Markov Model-based clustering [15], and Bayesian-based clustering [21]. All have been productively used to separate clusters based on the correlation of expression amplitude. However, one disadvantage of these methods is the loss of time continuity during the clustering [5]. While summarizing expression patterns in the data, the time-index structure is lost. In the absence of an alternative, it is often assumed that cluster members share a common pattern of time-index expression, and that the mean vector is a valid consensus pattern for the cluster despite the fact that information about the orientation of the vectors between time points is lost.

Herein we describe the use of multidimensional scaling (MDS) to reveal the consensus patterns present within a transcript profiling dataset, although such analyses might just as well be applied to protein or metabolite profiling studies. By using MDS, the underlying structure of time-index data can be described as a geometric representation [3], and pattern similarities can be represented by coordinates in space. We propose a model that represents each member of the cluster as a linear combination of the consensus patterns. The $R^2$ statistic for each gene indicates how much of the variance of an individual gene is explained by the model, estimating indirectly the extent of membership of the gene to the cluster.

Davidson et al. [6] were the first to describe the use of models to analyze expression profiles via MDS. Ding [7,8] described the use of MDS-based
MDS-based cluster analysis

models for analysis of latent change cycles in psychology and education. A previously described application of MDS for analysis of gene expression data is in displaying the relationship between genes with similar patterns of expression as a map [19]. Herein we describe the use of MDS to retrieve inherent time-index expression patterns shared by a set of genes.

Model and assumptions

We have adopted a MDS-based model for describing time-index profiling data, describing individual patterns as a function of the consensus patterns within clusters. The result is independent from the distribution of the data, however the sum of the error variance is assumed to be equal to the variance of the model. The variance of deviation from the model is equal for all time points (t) within each cluster, and the errors are independent with means 0 and variance $\delta^2$. The clusters are considered independent and all assignments correct. The set of assumptions can be thus summarized:

$$\frac{1}{t} \sum e_{g(t)} = \delta^2(e)$$

$$e_{g(t)} \iid (0, \delta^2)$$

Considering the m-Euclidian multidimensional space defined by m genes and t time points, the distance between two time points i and j is given by:

$$d_{ij} = \left( \sum_{m} (X_{ig} - X_{ij})^2 \right)^{1/2}$$

where $X_{ig}$ is the expression of gene g at time i.

The MDS configurations are free to rotate in Euclidian space. Rigid motions such as rotation, translation, and reflection are admissible transformations to which the configurations can be submitted without changing their shapes [3]. In our context, in order to display the patterns in positive correlation with the data, varimax rotation sometimes needs to be performed [7,8]. A transformation matrix can be derived after rotating the original dataset. The MDS-based model for analyzing time-index gene expression data has the form

$$Y_{g(t)} = I_{gi} \left[ \sum_k W_{gk} * X_{k(t)} + L_g + \xi_{g(t)} \right]$$

Where $Y_{g(t)}$ is the m*s matrix of observed signals for gene g at time t,

$W_{gk}$ is the profile match index for gene g on the dimension k,

$X_{k(t)}$ is the variable parameter estimate (i.e., the scale value) at time point,

$L_g$ is a level parameter estimate for each gene,

$\xi_{g(t)}$ is an error term,

$I_{gi}$ is an indicator function for clustered groups.

We have identified through the MDS procedure prototypical profiles of differential gene expression, and defined in the model three parameters for each gene indicating their relationship with the group patterns. The pattern match indices ($W_{gk}$) represent how well the individual genes match the prototype, indicating the degree of correspondence between the gene profiles and the group profile. A larger match index value indicates close similarity between the individual profile and the prototype group profile. A $R^2$ statistic is calculated for every expression
pattern, and indicates the proportion of variance in the observed data that can be explained by the model. A gene-specific level parameter \( L_g \) is estimated as well, indicating the average signal intensity of the genes over the dimensions.

**Illustrative dataset**

Methods typically used for analysis of gene expression lack a mechanism to detect the occurrence of multiple patterns within clusters. It is assumed that all members of a cluster share the same pattern regardless of the goodness of the clustering solution. We have arbitrarily chosen an illustrative dataset to demonstrate the discriminatory power of the MDS-based method for analysis of time-index data. We show that meaningful structural information in the data is still masked after obtaining a list of differentially expressed genes and clustering them via the k-means algorithm. The proposed model reveals this information as an extensive network of gene expression related to a specific physiological and developmental state. The data are available at the American Society for Plant Biology Internet site (http://www.plantphysiol.org/cgi/content/full/132/1/118/DC1), and have been described by [20].

In order to detect difference in expression patterns between the adaxial and abaxial surfaces of soybean embryos, the original authors conducted transcript profiling using mRNA collected in two or three replicates over 5 time periods; 0, 7, 14, 21, and 28 days [20]. A time-index analysis of transcripts from the adaxial side of the embryo was also conduced. A list of genes differentially expressed at one or more times during an experiment was taken from [20]. The expression patterns were used to separate the genes into 11 clusters, and an average expression pattern was assumed for of each cluster. We have downloaded the list, including the clustering information, and produced scatter plots for the expression of each gene as a function of time and resident cluster, in order to visualize the performance of the clusters. The complete datasets were imported from the Plant Physiology Internet site into SAS, the MDS procedure was implemented to detect the corresponding patterns, and parameters were estimated for the model using the iterative matrix language of SAS.

**Results**

**Relationship between datasets**

It is obvious from the scatter plots (Figure 1, columns A and B) that the relationship between the patterns of gene expression is not bijective, despite the mRNA having been isolated from the same source. In some instances, two genes that share an expression pattern in a cluster for one experiment (Column A) have
different patterns in the other experiment (Column B). Examples of this can be found when visually comparing each of the clusters. The difference in expression patterns implies underlying differences in biological function. Considering the relationship between the adaxial and abaxial surfaces of an embryo, the results imply differences in the physiological process ongoing within these tissues during development.
Figure 1. Transcript abundance scatter plots (A and B), and patterns revealed by MDS (C and D) for clusters 1 through 11 [20]. The intensity ratios for transcripts on the adaxial side of soybean cotyledons were determined at 7 and 14 days, 14 and 21 days, and finally 21 and 28 days of incubation on 2,4-D medium. A second dataset is presented where transcript levels on the adaxial and the abaxial sides of the cotyledons were compared at 0, 7, 14, 21, and 28 days after incubation on 2,4-D medium. The average intensity ratio is plotted for the time-index data in column A, and for adaxial versus abaxial expression in column B. The data were analyzed by cluster using the MDS procedure, and the resultant configurations are presented in column C for the time-index data and column D for the adaxial versus abaxial expression data.
The number of patterns per cluster

Scatter plots are useful for visual evaluation of clustering performance, with regard to the assumption that a single distinct pattern is representative. The scatter plots of the clusters presented in Figure 1 indicate that more than one pattern is required to summarize the data present in several of the clusters. Within most clusters, expression of one set of genes is increasing from time $t_i$ to time $t_{i+1}$ while another set is simultaneously decreasing. This scenario is present in the multiple clusters of both datasets (Figure 1). We have determined a one-dimension solution for each of the clusters, and present the retrieved patterns and the badness of fit (BoF) value for each cluster (Figure 1). The BoF values vary from 0.02 to 0.37. A value less than 0.1 is considered good, while a value >0.2 suggests that more dimensions are necessary in order to adequately capture variations in the data [3]. In Figure 1, 10 of the 22 clusters have BoF values of >0.2 indicating that the one pattern-assumption is incorrect.

To further illustrate the utility of the model in refining the cluster assignment, two of eleven clusters were examined in more detail. In Cluster 1, most of the 63 members fit the trend depicted by the MDS procedure with an acceptable $R^2$ value. Considering them as repeated measurements, analysis of variance performed with only Cluster 1 data revealed a significant difference between the time points, and plotting the least squares means revealed a pattern similar to the one retrieved by MDS (data not presented). However, visual examination of the scatter plot reveals that several genes do not follow the generalized increase in expression between days 14 and 21 (Figure 2A). If only a single pattern is considered, then the conclusions will be incorrect for a non-negligible proportion of the entire cluster. After fitting the model and calculating the parameters for each of the genes, 7 of 63 members of Cluster 1 have $R^2$ value of less than 0.65 (Figure 2C). The intensity ratios of these outlying genes are presented as scatter plots in Figure 2. Their profile match index values are near 0 (Table 1), further indicating that their pattern is not well-fit by the consensus pattern.
<table>
<thead>
<tr>
<th>ID</th>
<th>Days of Incubation</th>
<th></th>
<th>W</th>
<th>L</th>
<th>R²</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>7/14 14/21 21/28</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gm-r1070-3763</td>
<td>1.79 1.82 2.05</td>
<td></td>
<td>0.00</td>
<td>1.89</td>
<td>0.00</td>
</tr>
<tr>
<td>Gm-r1070-3966</td>
<td>2.02 2.01 2.08</td>
<td></td>
<td>0.00</td>
<td>2.04</td>
<td>0.03</td>
</tr>
<tr>
<td>Gm-r1070-4091</td>
<td>2.12 2.24 2.00</td>
<td></td>
<td>0.00</td>
<td>2.06</td>
<td>0.33</td>
</tr>
<tr>
<td>Gm-r1070-4769</td>
<td>1.88 1.98 2.14</td>
<td></td>
<td>0.00</td>
<td>1.96</td>
<td>0.09</td>
</tr>
<tr>
<td>Gm-r1070-544</td>
<td>1.67 1.88 1.90</td>
<td></td>
<td>0.01</td>
<td>1.73</td>
<td>0.56</td>
</tr>
<tr>
<td>Gm-r1070-6566</td>
<td>1.58 1.77 2.34</td>
<td></td>
<td>0.00</td>
<td>1.84</td>
<td>0.01</td>
</tr>
<tr>
<td>Gm-r1070-8592</td>
<td>1.56 2.10 2.13</td>
<td></td>
<td>0.03</td>
<td>1.69</td>
<td>0.61</td>
</tr>
</tbody>
</table>

**Table 1.** Intensity ratios and parameter estimates for the outlying group (genes with R² values < 0.65) in Cluster 1. The intensity ratios for transcripts on the adaxial side of soybean cotyledons were determined at 7 and 14 days, 14 and 21 days, and finally 21 and 28 days of incubation on 2,4-D medium. The expression data were analyzed using the MDS procedure, and the output fitted to the model. Parameter estimates are the profile match indices (W), and level (L) and R² values.

The outlying group (Figure 2) highlights the discriminatory power of the model in locating genes within the cluster that have atypical patterns of expression. The specific expression patterns of the outlying genes are marginalized in a cluster that uses the mean vector to represent the pattern. While the goal is to summarize the data, important information must not be ignored or discarded. Considering that even a single gene could be responsible for a physiological response, an ideal approach would not omit any subset of genes. We have analyzed these genes, as they are expressed in the two datasets, side by side, in order to make full use of the relationship between datasets, tissues, physiological processes, and direct functions. We have chosen to interpret the members of the outlying group in Clusters 1 and 10. The genes of Cluster 10 have similar pattern in the location dataset, but different pattern in the time-index dataset. They represent the opposite of the scenario described for the outlying groups of cluster 1. Together, this illustrates how the two datasets complement one another, and how two specific genes with similar patterns in the time-index dataset can lead to different interpretations when considered in the context of the other dataset.
Figure 2. Pattern detection and modeling of Cluster 1. The average intensity ratios for the time-index measurements from Cluster 1 are plotted in panel A. The intensity ratios present comparison between transcript levels on the adaxial side of the cotyledon after 7 and 14 days, 14 and 21 days, and 21 and 28 days of incubation on 2,4-D medium. The MDS configuration representing the consensus pattern for Cluster 1 is presented in panel B. Panel C is the scatter plot of the genes that poorly fit the model, with $R^2$ values of <0.65. Panel D is the expression of the genes presented in panel C when comparing the adaxial to the abaxial side at 0, 7, 14, 21, and 28 days after incubation on 2,4-D medium.

Considering the seven gene expression patterns that are outliers from Cluster 1, the intensity ratios of their expression maintain the same increasing trend (>1.5) throughout the experiment, which yields a flat pattern (Figure 2). It is tempting to conclude that these genes are involved in embryogenesis since their transcripts accumulate on the adaxial side during incubation on the 2,4-D medium. For a better estimate of the involvement of these genes, it is
appropriate to examine the dataset comparing the two sides of the cotyledon. The genes can be grouped into three classes, assuming that the differences between \( \alpha \), \( \beta \), and \( \gamma \) are statistically significant. Class A, with gene Gm-r1070-544 (from an uncharacterized EST clone), was more abundantly expressed on the abaxial side at day 0, then from day 7 on, both sides have similar transcript abundance for the duration of the incubation. Class B includes genes Gm-r1070-3966 (histone H4), Gm-r1070-8592 (40S-ribosomal protein) and Gm-r1070-4091 (acyl carrier protein), transcripts of which are more abundantly represented on the adaxial side at 21 days after incubation. The upward trend on the adaxial side was observed after day 14 in the time-index dataset, suggesting an upward trend on the abaxial side as well and that transcripts begin to decline differentially at 21 days on the abaxial side. These genes are involved in growth and maintenance, and their pattern of expression matches their function considering the physiological processes ongoing on both sides of the cotyledons. Class C groups the genes Gm-r1070-3763 (high mobility group protein HMG1), Gm-r1070-6566 (high mobility group protein HMG1) and Gm-r1070-4769 (protein phosphatase 2A inhibitor). Their transcripts are more abundant on the adaxial side as early as day 14. The upward trend was not observed on the abaxial side, suggesting that these genes are involved in embryo-specific processes.

Cluster 10 was analyzed using the location (adaxial versus abaxial) dataset as the main reference, being represented by a unique pattern within the cluster. The \( R^2 \) and profile match indices (Table 2) for the time-index experiment identify genes that do not follow the consensus patterns. Alternative patterns in the dataset suggest different interpretations concerning gene product function during embryo development. At time zero, higher transcript levels were recorded on the adaxial side for the members of this cluster. Subsequently similar transcript levels were seen on both sides of the cotyledons, with the exception of gene Gm-b10BB-2 (Phenylalanine ammonia-lyase) where transcript levels increased at 14 and 21 days. The pattern indicates equal transcript levels on both sides, but does not indicate the direction of variation. They could be up-regulated, down-regulated, or remain constant, information that is provided by the adaxial time-index dataset. In fact the dominant pattern in this dataset indicates accumulation of transcripts over time. This suggests that the genes are involved in activities associated with both callus and embryo development. Again, the pattern of Gm-b10BB-2 expression is exceptional. This transcript was more abundant on the adaxial side at 14 and 21 days, while declining on the adaxial side from 14 to 28 days after incubation on enriched medium. A downward trend was also observed for Gm-r1070-8532 (glutathione S-transferase) at 28 days.
From patterns to networks

If a clustering solution were ideal, then a single dimension would represent the pattern of the data, and fit the model with a good $R^2$ value. This would allow pattern interpretation in the context of pathways and networks. However, in real-life non-ideal clustering solutions, it is still possible to select a dominant pattern, or data with good $R^2$ values, in order to construct a network. For the adaxial versus abaxial comparison, it is possible to arrange the patterns according to the differences in timing of expression. A summary of molecular events occurring on the adaxial side of the cotyledon can be generated by including the clusters (from Figure 1) whose members have accumulated chronologically after incubation of soybean embryos on 2,4-D medium. Regulatory networks can be constructed based upon time-index, gene function, or both. A time-index based network was constructed based upon the clustered data from Figure 1 (Figure 3). As an example, a transcription factor from Cluster 6 could be involved in expression of genes in Clusters 1, 2, 3, 4, 5, 8 or 11 but not a gene in Clusters 7, 9, or 10.

<table>
<thead>
<tr>
<th>ID</th>
<th>Days of Incubation</th>
<th>W</th>
<th>L</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gm-bl0BB-2</td>
<td>1.85 0.90 0.92</td>
<td>-0.15</td>
<td>1.63</td>
<td>0.66</td>
</tr>
<tr>
<td>Gm-r1070-21</td>
<td>1.64 2.44 2.20</td>
<td>0.13</td>
<td>1.73</td>
<td>0.88</td>
</tr>
<tr>
<td>Gm-r1070-3714</td>
<td>1.75 1.96 1.66</td>
<td>0.04</td>
<td>1.67</td>
<td>0.60</td>
</tr>
<tr>
<td>Gm-r1070-408</td>
<td>0.94 3.45 2.14</td>
<td>0.43</td>
<td>1.01</td>
<td>0.99</td>
</tr>
<tr>
<td>Gm-r1070-6358</td>
<td>1.77 1.47 1.95</td>
<td>-0.05</td>
<td>1.88</td>
<td>0.49</td>
</tr>
<tr>
<td>Gm-r1070-7794</td>
<td>2.07 1.93 2.98</td>
<td>-0.04</td>
<td>2.45</td>
<td>0.05</td>
</tr>
<tr>
<td>Gm-r1070-7890</td>
<td>1.82 2.09 2.34</td>
<td>0.04</td>
<td>1.89</td>
<td>0.56</td>
</tr>
<tr>
<td>Gm-r1070-8455</td>
<td>1.31 2.95 2.28</td>
<td>0.28</td>
<td>1.43</td>
<td>0.05</td>
</tr>
<tr>
<td>Gm-r1070-8532</td>
<td>1.85 2.05 0.83</td>
<td>0.06</td>
<td>1.41</td>
<td>0.07</td>
</tr>
<tr>
<td>Gm-r1070-872</td>
<td>3.33 1.23 3.29</td>
<td>-0.38</td>
<td>3.64</td>
<td>0.85</td>
</tr>
<tr>
<td>Gm-r1070-9181</td>
<td>0.59 4.15 1.87</td>
<td>0.62</td>
<td>0.53</td>
<td>0.99</td>
</tr>
</tbody>
</table>

Table 2. Intensity ratios and parameter estimates for Cluster 10. The intensity ratios for transcripts on the adaxial side of soybean cotyledons were determined at 7 and 14 days, 14 and 21 days, and finally 21 and 28 days of incubation on 2,4-D medium. The expression data were analyzed using the MDS procedure, and the output was used to fit the model. Parameter estimates are the profile match indices (W), the level (L), and $R^2$. 
Discussion

Relationship between datasets

The experiments, from which the data used in this study were collected, were designed to reveal the molecular events that occur during embryo development in soybean seeds. Because most of the events occurs after 14 days on enriched medium, data were collected at 0, 7, 14, 21, and 28 days in order to study variation of gene expression on the adaxial side of the cotyledon and compare gene expression variation with the abaxial side. The two experiments are designed in such a way that clear statements about gene expression relative to
MDS-based cluster analysis

callus development can be derived when interpreting the patterns from the two datasets side-by-side. A set of genes might share common pattern in one dataset and different patterns in the other dataset, indicating that they are differentially involved during the two physiological processes. Instead of pulling the data together and clustering them, it is best to cluster the gene list based on one dataset, and to use the clustering information to group the other dataset. We have interpreted the patterns out-group of Cluster 1 and Cluster 10 for the two datasets side-by-side in this paper to show how this method can provide more insight into the molecular events ongoing during soybean embryo development.

Clustering data

Considering the number of genes involved in microarray experiments, there is a need to summarize the data by grouping them into classes of genes that display certain similarities. For example, the signaling pathways leading to expression of effector genes and phenotypic control are time-indexed. Transcription factors upstream of a signaling pathway need to be expressed first, and delay in expression could compromise the ability of the pathway to provide the appropriate response at the appropriate time. Therefore, clustering gene expression based on the timing of expression is appropriate to allow meaningful inference about the underlying physiological process. An ideal solution yields clusters with low BoF for a one dimension solution and high R² for the members after fitting the model. The MDS procedure helps refine poorly clustered datasets and find the patterns that best represent the clusters. In Figure 1, the 1-dimensional solution is not suitable for the Clusters 6A, 7A, and 8A as indicated by the BoF values. Ideally, for at least one dataset, the members of the cluster would share the same time-related pattern of expression. This is assumed to be true by most biologists performing cluster analysis. However, scatter plots of the data over time show that this is often not the case. While one consensus pattern can represent the data in Cluster 5A, this is not true for Clusters 6A, 7A, and 8A (Figure 1). At least two patterns are required to represent the activity of the members of these clusters. A BoF statistic of more than 0.10 suggests that additional dimensions are required to capture all the patterns there are in the data. Cluster analysis and MDS complement each other since the cluster procedure is substantially blind [12], and the dimensions retrieved by the MDS procedure are influenced by the distribution of the data in space [3].

The MDS algorithm is sometimes used to represent similarity in gene-expression data on a 2-dimensional plot [19]. This type of representation displays the real distance between expression levels, and is desirable for classification purposes. Herein, the observed dissimilarities between time points for every gene in the cluster are considered instead, in order to capture the time-index patterns in the data.
Finding the right pattern

The sequence defined by the vectors of intensity of gene expression between time points is called a pattern, and is characterized by its scale, orientation, and direction [19]. The direction is from the time axis, and the scale is determined during the MDS analysis. The orientation of the vector (intensity ratio) between two time points becomes critical, in order to explain variation in the data. The problem of uniformity of vector orientation is achieved during clustering. If the pattern of expression of a specific gene varies randomly within the cluster over the duration of the experiment, a high individual $R^2$ statistic will result. This would flag that particular pattern for additional analysis. Even when the model explains the data well, it is important to compare how individual genes differ along the prototypic profiles using the profile match index and $R^2$ values. Information about each individual gene is important because their specific variation might be correlated with specific function in a network context.

It is assumed that genes within clusters share common features, and are likely to have similar patterns of expression, although this is not always true [2]. The degree of similarity in patterns between the genes can be assessed by observing the scatter plots by for each cluster. Preferably, the $R^2$ statistics for each cluster can be obtained showing how well an individual pattern fits the model. Since every gene is expressed as a linear combination of the dimensions, the better the cluster solution, the better the model will fit the data. In Figure 2, the BoF was low enough to accept the one dimension solution of the MDS procedure. The outliers did not impose enough stress on the MDS configuration to influence the BoF to a large extent. However the model provides an approach greedy enough to evaluate the suitability every gene assigned to the cluster.

Parameters of the model

The model allows comparison of two members of the same cluster, using the profile match index which depicts how much the individual pattern resembles the pattern of the class. More useful for data-mining, the individual $R^2$ statistics indicate how well an individual pattern corresponds to the consensus pattern of the cluster. As demonstrated by our analysis of Cluster 1, the $R^2$ statistic helps identify unique patterns within the cluster that might otherwise be obscured.

Database management and data-mining

Data mining is often directed towards detecting and describing patterns, trends, and relations in biomics profiling data. Fayyad et al. [10] defined data mining as “a mechanized process of identifying or discovering useful structure in the data.” Therefore data mining is a distinct activity larger than graphics-based data exploration. Adrienko and Adrienko [1] distinguish three stages in data mining.
process; the initial exploration, model building or pattern identification, and deployment or application of the model for general prediction. Most of the techniques used to mine profiling data are graphics-based. However, additional knowledge about the data can be gained using computational data mining techniques [1]. Herein we present and illustrate a model that describes the data with respect to the common inherent patterns within clusters. The model calls for the scale values of gene expression over time that describes the inherent pattern for a particular cluster through a MDS algorithm. Gene-specific parameters describe the variation in expression with respect to consensus patterns that define the cluster. The integration and implementation of this model will add an important component to any package dedicated to comprehensive data mining of time-index biomics profiling data.

Conclusions

We have used a publicly-available dataset to illustrate the benefits of the MDS-based model for analysis of time-index data. The data chosen for illustrative purposes are from transcript profile analysis, but the method is equally applicable to profiling time-index studies of proteins, metabolites, metals, etc. The MDS procedure applied to the data comprising each cluster allows retrieval of the structure inherent in the data and graphic display of such structure. This visual representation of the common variation in gene expression over time within cluster facilitates summarizing the overall expression in the datasets and elaborating networks of gene expression. The model explains each gene expression profile as a linear combination of the MDS configuration, and assesses the degree of membership of each gene to the cluster based on their pattern of expression over time. The model allows refinement of the clustering solution by flagging outliers, and allows the method to be as greedy as necessary in analyzing the data. The proposed method allows retrieval of meaningful information that would otherwise be obscured in larger datasets, and presents a reliable way to exhaustively mine these datasets.

We have developed a basis for extracting the most information from a profiling dataset by using methods that allow inference about biological phenomenon not directly measured. Because of the relationship in the datasets, we were able to probe activities important during callus formation from an experiment originally designed to characterize soybean embryo development. This subtle information was buried within the larger dataset, and would not be accessible without efficient statistical procedures to define optimum grouping solutions and to detect structures in the data.

Acknowledgement. M.L. Johnston was instrumental in preparing the Figures.
References


MDS-based cluster analysis


Received: November, 2008