Mapping Physiological States from Microarray Expression Measurements

Gr. Stephanopoulos*, D. Hwang, W. A. Schmitt, J. Misra and Geo. Stephanopoulos

Department of Chemical Engineering, Massachusetts Institute of Technology

Room 56-469, Cambridge MA 02139

Correspondence should be addressed to
Gregory Stephanopoulos
Massachusetts Institute of Technology, Room 56-469
77 Massachusetts Avenue
Cambridge, MA 02139, USA
e-mail: gregstep@mit.edu
Tel: (617) 253-4583; Fax: (617) 253-3122

Keywords: ***Microarrays, cell physiology, expression phenotype, discriminant analysis, physiological state***

ABSTRACT

MOTIVATION: The increasing use of DNA microarrays to compare different physiological conditions requires methods for elucidating the effect of multiple genes in concert to bring about a particular state.

RESULTS: We propose a method for the mapping of the physiological state of cells and tissues from multidimensional physiological data such as those obtained from gene expression measurements with DNA microarrays. The method uses Fisher Discriminant Analysis to create a linear projection of gene expression measurements where the projected measurements from different types of classes are maximally separated. As an improvement over other typical classification methods, this method provides insight into discriminating characteristics in the expression measurements in terms of the contribution of individual genes to the discrimination of distinct physiological states. This projection method offers visualization of classification results in a reduced discrimination space to help understand discrimination characteristics within and between classes. Examples from four different cases demonstrate the ability of the method to produce well-separated groups in the projection space and to identify important genes and their interactions in discriminations. The method can be augmented to also include data from the proteomic and metabolic phenotypes and is useful in disease diagnosis, drug screening and bioprocessing applications.

CONTACT: gregstep@mit.edu

INTRODUCTION

Cell function is the integrated outcome of numerous cellular processes and it is therefore difficult to categorize the physiological condition of a cell accurately without a host of information about these processes. In some simplistic considerations, growth is used as an all-encompassing physiological descriptor. Growth (and growth rate) can usually be supplemented by an array of extracellular variables in describing cell function and physiology, such as respiration rate or rate of glucose consumption or by derivative quantities such as the rates of glycolysis, TCA cycle activity, pentose phosphate pathway flux, etc.(Vallino & Stephanopoulos, 1993; Vangulik & Heijnen, 1995; Stephanopoulos, 1999). Cell function, as described by the above variables, is the expression of a particular cellular state that can be quantified by a variety of methods probing the trancriptional, proteomic and metabolic state of a cell. Since all cellular processes originate at the transcription level, it can be argued that transcriptional profiling provides a broad, albeit incomplete, descriptor of the cellular physiological state. Consequently, gene transcription measurements by various types of microarrays (Schena *et al.*, 1995; Lockhart *et al.*, 1996) contain information that should be useful, in principle, in defining the physiological state of a cell. However, although no one doubts the value of information residing in microarray expression data, it is yet not clear how these data can be used in a comprehensive definition of the physiological state of a cell.

Given gene expression data obtained from different physiological states, classification techniques from multivariate statistics can be used as diagnostic tools and also to generate insights into the underlying physiological states. Several tools have been introduced for analysis of gene expression data. Complex, non-linear classifier such as Supportive Vector Machines (SVM; Furey, T.S. *et al*., 2000) have been used to classify samples for diagnostic purposes, but due to the complexity of these classifiers, they can not easily provide insight to the relationships between individual genes and different physiological states. Furthermore, these techniques do not lend themselves to easily interpreted visualization methods. Linear projection methods such as singular value decomposition (SVD; Alter et al., 2000; Holter, N.S. *et al*., 2000)/principle components analysis (PCA; Dillon & Goldstein, 1984; Wen, X.L. *et al*., 1998) have been used to visualize data structure *{sets??}*, but can not optimially recognize differences based on defined biological classes. Similarly, unsupervised clusterings (Dillon & Goldstein, 1984; Eisen *et al*., 1998; Tamayo *et al*., 1999; Golub *et al*., 1999) such as Self-Organizing Maps (SOM) and hierarchical clustering have been used to find patterns in biological data, but they are also sub-optimal in recognizing the differences based on the defined biological classes. {Thus, */we find that* } all these methods are not suitable for use as classification tools {*to*/that should} improve our understanding of physiological differences in genomic basis and adequately assign new samples to previously defined classes.

In view of these limitations, we are in need of a technique which, a) is robust in classification, b) explicitly connects individual genes to discriminating among classes and thus provides insights into underlying physiology in terms of these genes and their interactions, and c) facilitates a visualization of the discrimination of samples among classes in the case many classes exist. No previous work has provided a comprehensive discussion on applications of their classification tools to high-throughput biological data on this basis, as current work has been limited to the consideration of only two classes at one time (Golub *et al*., 1999), to unsupervised learning from microarray expression data (Tamayo *et al*., 1999) and to multiple class discrimination on lower-dimensional transcript data (Spanakis & BroutyBoye, 1997).

We propose the use of projections defined by Fisher Discriminant Analysis (FDA) to overcome the limitations inherent in the techniques described above in mapping cellular physiological states from gene expression measurements. FDA classifies samples based on linear composites of genes. The sheer magnitude of the expression phenotype necessitates that either a small number of genes or composites of gene expression measurements be used in the definition of the physiological state. Concentrating on only a few genes implies that, despite its apparent complexity, cellular function is still determined by a small number of genes, a corollary that is not supported by the accumulating evidence of microarray data (Golub *et al*., 1999). The use of composites of gene expression measurements in lieu of the expression measurements themselves, however, allows for consideration of all genes of interest by projecting the expression phenotype into a lower dimensional space where the various physiological states can potentially be identified.

METHODS

Fisher Discriminant Analysis (FDA) has the following objectives:

1) find the discriminant axes/linear composites of predictor variables (genes), with the property of maximizing the ratio of between group to within group variability of projections onto these axes, subject to the constraint that discriminant scores (point projections onto the axes) are uncorrelated with the scores on any previously obtained axis,

2) determine how many discriminant axes are statistically significant (the dimensionality of the discrimination space) and whether the group centroids are different in a statistically significant way,

3) successfully assign new samples to one of multiple groups based on a classifier developed using the discriminant scores.

4) determine which of the predictor variables (genes) contributes most to discriminating among the groups.

Generation of discriminant axes/linear composites

Given c classes, the discriminant axes (or discriminant weights) are generated from the eigenvalue decomposition (ED) of the ratio of between group variance ($\boldsymbol{B}$) to within group variance ($\boldsymbol{W}$), $\boldsymbol{W}^{-1}\boldsymbol{B}$:

$$\boldsymbol{W}^{-1}\boldsymbol{B}\boldsymbol{L} = \boldsymbol{L}\boldsymbol{\Lambda}$$

$$\text{with } \boldsymbol{W} = \sum_{k=1}^{c}\left(\boldsymbol{X}_k - \boldsymbol{1}\bar{\boldsymbol{x}}_k\right)^T\left(\boldsymbol{X}_k - \boldsymbol{1}\bar{\boldsymbol{x}}_k\right) \text{ and } \boldsymbol{B} = \boldsymbol{T} - \boldsymbol{W} = \left(\boldsymbol{X} - \boldsymbol{1}\bar{\boldsymbol{x}}\right)^T\left(\boldsymbol{X} - \boldsymbol{1}\bar{\boldsymbol{x}}\right) - \boldsymbol{W}$$

where $\boldsymbol{X}_k$ is the data matrix composed of only samples in class k in which predictor variables (in this case, g genes) are in columns and expression measurements are in rows and $\bar{\boldsymbol{x}}_k$ is the group mean ($1\times g$) for class k. The eigenvector matrix ($\boldsymbol{L}$) defines the discriminant axes, called discriminant weights, and the eigenvalue matrix (Λ) represents the discriminant powers of each corresponding axis. The resulting axes meet the constraint that the scores are uncorrelated one another, but the axes are not orthogonal, because $\boldsymbol{W}^{-1}\boldsymbol{B}$ is not positive definite. When the axes are described in the linear composite forms, the linear composites are called the discriminant functions (DFs) and the discriminant score ($\boldsymbol{y}$), projection of a sample, is calculated for the actual $\boldsymbol{x}$ using the DFs (for the discriminant axes j with g genes, $y_j = \boldsymbol{x}\boldsymbol{L}_j = \sum_{i=1}^{g} x_i L_{ij}$).

However, we have found that sometimes, the singular value decomposition (SVD) of $\boldsymbol{W}^{-1}\boldsymbol{B}$ would produce better discriminant axes than the eigenvalue decomposition (ED) of $\boldsymbol{W}^{-1}\boldsymbol{B}$, and thus the axes would more effectively capture the between group variance. For those cases, the SVD was applied to find the axes and to calculate the discriminant scores:

$$\boldsymbol{W}^{-1}\boldsymbol{B} = \boldsymbol{U}\Lambda\boldsymbol{L}^{T}$$

where $\boldsymbol{U}$ is the left singular vector, $\boldsymbol{L}$ is the matrix of discriminant axes, and Λ is the singular values representing the discriminant powers along the corresponding axes.

Determination of dimensionality of discrimination space

The number of statistically significant DFs and thus the dimensionality of discrimination space can be determined from in total (min[g,c]) possible discriminant functions, using Bartlett's statistic V which assesses the significance of each factor:

$$V_j = \left[(n-1) - \frac{1}{2}(g+c)\right]\ln(1+\lambda_j) \sim \chi^2_{g+c-2j}$$

where λ_j is eigenvalue j from the eigenvalue decomposition above and n, g and c are the number of samples, predictor variables (genes) and classes, respectively. The statistic V approximately follows chi-square distribution with g+c-2j degrees of freedom. If the value of V is greater than a critical point of significance from the χ^2 distribution with the appropriate degrees of freedom, the DF (eigenvector j) is concluded to be significant. In practice, however, the application of this statistic V to expression measurements is limited, because the small number of samples results in negative statistic V values (g >> n in the above equation).

Also, the statistic V can be used to determine whether the centroids of the discriminant scores for classes are statistically different. Because DFs are uncorrelated, the V_js can be added to give V. The resulting V can be used to assess the statistical significance of full discriminant functions ($V = \sum_{j=1}^{r} V_j \sim \chi^2_{g(c-1)}$, where r is the dimensionality). Assessing the statistical significance of the full DFs is equivalent to determining whether the centroids of the scores are statistically different, because the centroids should be different in full discriminant space, if the full discriminant functions are statistically significant. Multivariate Analysis of Variance (MANOVA; Johnson & Wichern, 1992; Dillon & Goldstein, 1984; SAS Guide, 1989) can also be used to analyze the scores in discriminant space, but MANOVA is used to detect whether there are subgroups in each group whose centroids in discrimination space are statistically different. This subgroup detection using MANOVA can help identify new subgroups comprising a known physiological subtype.

Development of a classifier using the group centroids.

A FDA classifier can be built on the first *r* DFs as follows:

Allocate a new sample, $\boldsymbol{x}$ ($1\times g$) to class *k* if

$$d_k(\boldsymbol{x}) = \sum_{j=1}^{r}\left(\hat{y}_j - \bar{y}_{kj}\right)^2 = (\boldsymbol{x}-\bar{\boldsymbol{x}}_k)\boldsymbol{L}\left[(\boldsymbol{x}-\bar{\boldsymbol{x}}_k)\boldsymbol{L}\right]^T \le d_i(\boldsymbol{x})\sum_{j=1}^{r}\left(\hat{y}_j - \bar{y}_{ij}\right)^2 \text{ for all } i \ne k$$

where $\boldsymbol{L}$ is the *r* discriminant weights obtained from the eigenvalue decomposition of $\boldsymbol{W}^{-1}\boldsymbol{B}$ above, $\bar{\boldsymbol{y}}_k$ is the centroid of discriminant scores in class *k*, and $\hat{\boldsymbol{y}}_j$ is a projection of the new sample into the *j*-th discriminant weights (the *j*-th axis).

Contributions of predictor variables to discrimination among classes

For most of cases where the data have a significant correlation among predictor variables (genes), contributions of predictor variables to discrimination among classes are determined using discriminant loadings ($\boldsymbol{L}^*$), rather than using the discriminant weights ($\boldsymbol{L}$):

$$\boldsymbol{L}^* = \boldsymbol{R}\boldsymbol{D}^{1/2}\boldsymbol{L}$$

where $\boldsymbol{R}$ is the correlation matrix of the data matrix and $\boldsymbol{D}^{1/2}$ is the diagonal matrix of standard deviations for predictor variables (genes) (Dillon & Goldstein, 1984). However, for data with a small correlation among the predictor variables, the discriminant weights can be directly used to determine the contributions of the variables. Sometimes, interpreting the significance of the predictor variables for classification can be improved through the rotation of the discriminant weights/loadings under adequate criteria such as varimax and promax (Johnson & Wichern, 1992). The absolute value of the rotated discriminant weights/loadings will typically be closer to one or zero, more clearly identifying genes as important to discrimination or not, respectively.

RESULTS

For a data matrix $\boldsymbol{X}$ with *g* columns of gene expression data measured in *n* rows of samples that can be *a priori* classified in *c* classes (or groups), FDA defines a linear projection to a lower dimensional space such that the mean differences among the *c* classes are maximized. This is analogous to the use of Principal Component Analysis (PCA; Wen *et al*., 1998) that maximizes, instead, the variability in the data (i.e. the covariance matrix). As such, the two methods yield different outcomes. However, FDA is known to be a better classification tool than PCA, because FDA it emphasizes differences among pre-defined classes, while PCA merely provides visualization of overall data variance (Zhao & Maclean , 2000).

The FDA method is described schematically in Figure 1 depicting y_1 and y_2 in a

2-D discrimination space where the scores (projections) of the expression data of three genes shows that the separation of the three sample classes is maximized. Each discriminant axis (or DF) in the FDA space is defined as a linear composite of the actual gene expression measurements and is obtained by eigenvalue decomposition of $\boldsymbol{W}^{-1}\boldsymbol{B}$. The statistic V shows that 1) the two DFs (*e.g.* dimensionality of discrimination space = 2) are statistically significant and 2) the centroids of the scores are statistically different: 1) V_1= 86.8 is larger than the critical value V_c= 9.49 derived from the χ^2 distribution with 4 df. at the 0.05 significance level and V_2=41.7 is also larger than V_c=5.99 and 2) $V=V_1+V_2$=96.3 is larger than V_c=12.6. The coefficient multiplying each gene expression measurement (discriminant weights) provides a measure of that gene's importance in defining the projection (an improved measure, discriminant loadings, is discussed in the second example [Figure 3] for a strongly correlated data set).

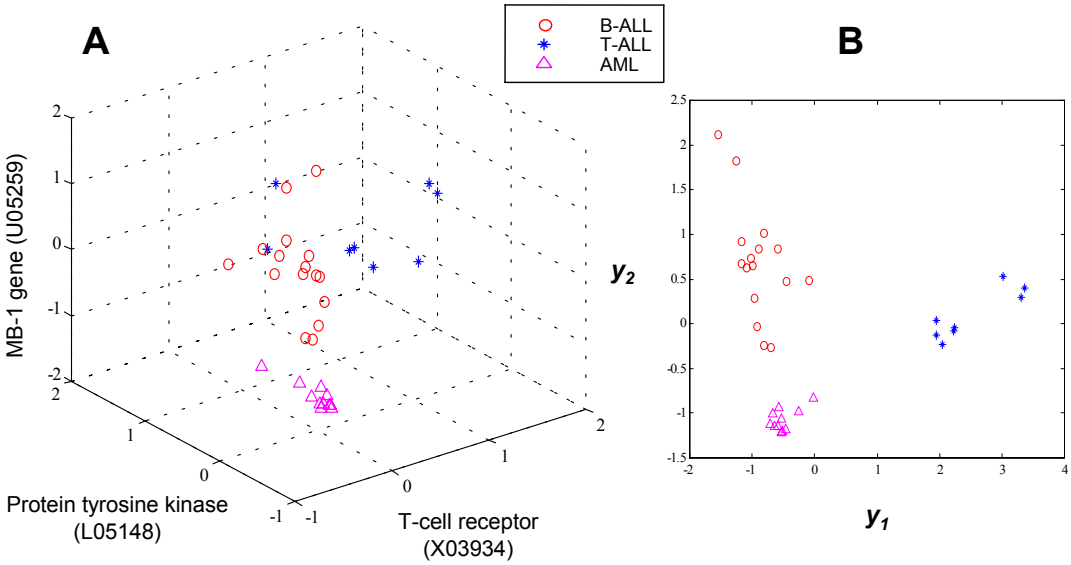

Figure 1. An illustration of FDA in three dimensions. Three classes of physiological states are shown corresponding to three different types of leukemia (B-ALL (red circles), T-ALL (blue asterisks), and AML (pink triangles)). Each point represents a different tissue sample plotted in the 3-D space shown in (A) that is defined by the expression levels of the three genes. The class separation in the original gene expression space is not complete. FDA projection onto the discrimination space allows complete class separation as shown in (B) with the same symbols representing the scores. The first DF differentiates the B-ALL and AML from T-ALL and the second DF B-ALL from AML. Application to 35 samples (8 T-ALL, 16 B-ALL and 11 AML) gives y_1=-0.8220g_{X03934}-0.2268g_{L05148}+0.5224g_{U0529} and y_2=-0.2916g_{X03934}-0.3922g_{L05148}-0.8724g_{U0529}. Expression data shown from (*http//waldo.wi.mit.edu/MPR/*).

We applied FDA projections to four examples of gene expression phenotypes generated in our laboratory and also published in the literature. In the first example, cultures of the photosynthetic bacterium *Synechocystis* sp. PCC 6803 were cultivated through an initial period of 48 hours of growth under light followed by 24 hours of darkness. The cultures were then cycled between light and dark conditions for 100 minutes each (Figure 2). The expression levels of 88 genes associated with harvesting of light energy and central carbon metabolism were measured at 23 time points (29 total samples, including duplicates) using DNA microarrays (Gill *et al.*, 2001). Total signal to noise ratio of the microarray fluorescence was determined to be c.a. 4.0 indicating that background noise minimally interfered with the fluorescence of hybridized spots.

Reproducibility of expression measurements, evaluated from microarray to microarray measurements, as well as from intra-microarray triplicate spots, was 45% suggesting that a 90% difference in fluorescence is reproducible within 95% confidence level.

Of the 88 total genes considered, 27 discriminatory genes were identified based on their Wilks' lambda measure (Dillon & Goldstein, 1984) with a stringent 99% confidence level. Figure 2 shows the projection of the expression phenotype of the 27 *Synechocystis* discriminatory genes to the FDA-defined discrimination space. The statistic V was used to determine that all three possible DFs are statistically significant (V_1=80.0597, V_2=62.4680, and V_3=22.4723) and the centroids of the scores are statistically different (V= 165). Thus, three DFs in the projection space effectively distinguish the four phenotypic classes of growth under the light and dark conditions shown in Figure 2. The 2-D diagrams of the three DFs (Figure 2c) help us see how the individual DFs discriminate the four classes. DF1 distinguishes group 2 from the other groups while DF2 separates groups 1 and 3. Hence, the second discriminant weights provide information on the identity of the genes supporting the differences in the cellular processes occurring under light and dark conditions.

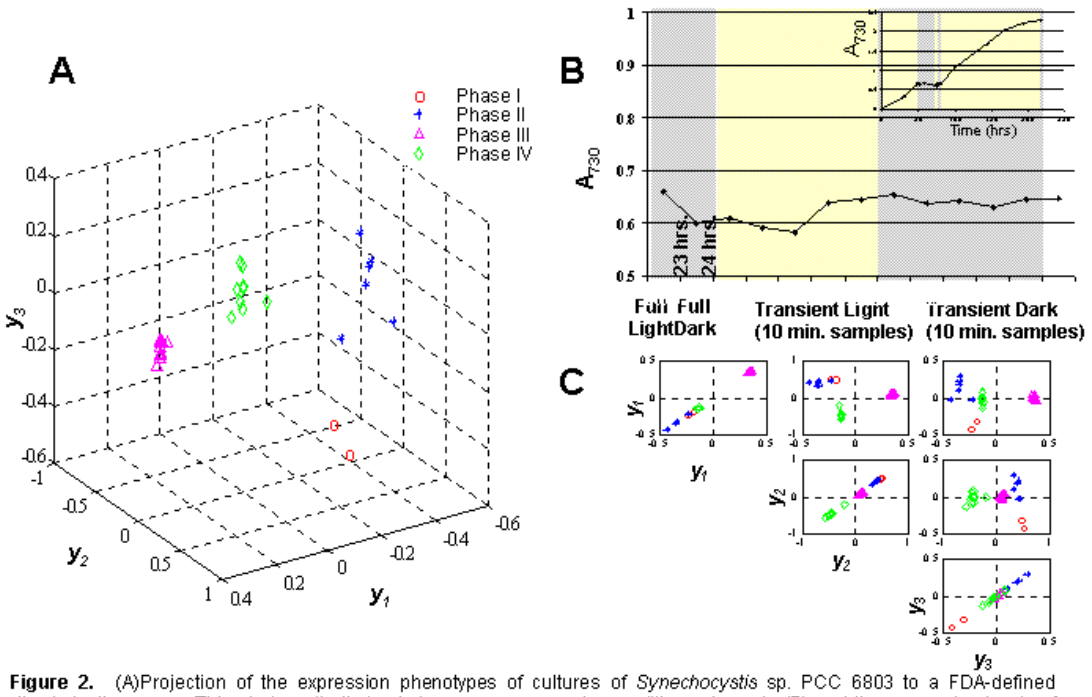

Figure 2. (A)Projection of the expression phenotypes of cultures of *Synechocystis* sp. PCC 6803 to a FDA-defined discrimination space. This photosynthetic bacterium was grown under conditions shown in (B) and the expression levels of 88 genes were measured by a DNA microarray at 29 time points spanning the entire course of the experiment. Of the 88 genes, 27 were identified as most discriminating of the four classes defined by the four different light conditions and their expression levels were projected to the FDA-defined space. It can be seen that the four phenotypic classes are clearly identified in the 3-dimensional FDA projection space. (C) The first DF shows the largest discrimination power separating all the groups, discriminating clearly Phase III from the others. The second DF separates Phase IV from Phases I and II, while the third DF is necessary to separate Phase I from II.

As another example, this procedure was applied to the expression phenotypes measured in samples from patients with acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML) (Golub *et al*., 1999). Additionally, the ALL samples were further subdivided into B-lineage ALL and T-lineage ALL (B-ALL and T-ALL, respectively). To reduce the number of genes considered, the Wilks' lambda measure

was used again to uncover those genes that provide significant discrimination among the three classes by exceeding the critical value from F distribution of Wilks' Lambda at 0.01 significance level. 1226 genes met this criterion of which 50 genes were selected for use in the FDA projection based on the error rate obtained from leave-one-out cross validation (SAS Guide, 1989). The statistic determined that the three leukemia subtypes (groups) are separable by two DFs (Figure 3A), each of which is statistically significant (V_1=193.0641, V_2=124.9875, and V= 318.0516). Discriminant loadings in Figure 3B shows how those 50 genes individually behave and interact to separate the three disease classes: 1) the genes were clustered into three groups, 2) each gene has its specific contribution to a particular class except two genes between AML and T-ALL specific genes, and 3) all genes in each gene group have common expression patterns, which might be considered as co-regulation patterns. The specificity of a group of genes to a particular sample class implies that the expressions of the group of genes are highly elevated only in the sample class and down-regulated in the other classes.

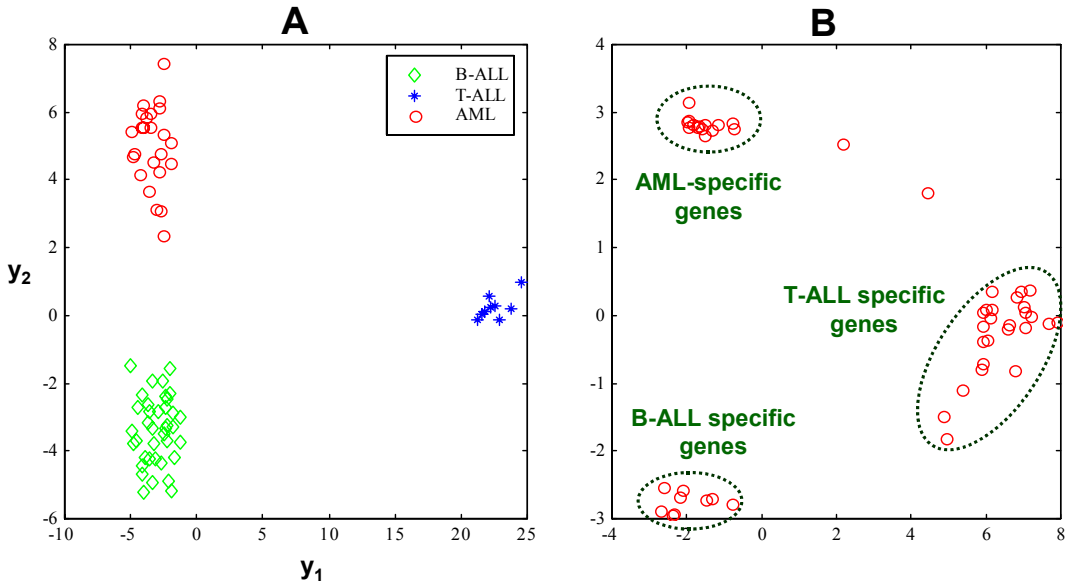

Figure 3. FDA projection of expression data obtained from patients with B-ALL, T-ALL, and AML. Projection of 50 discriminatory genes shown in (A) allows FDA to clearly separate the three classes of leukemia expression phenotype in a 2-D discrimination space. The first DF distinguishes the B-ALL group from T-ALL and AML. The second DF separates T-ALL group from AML to complete the group separation. The contributions of individual genes to the discrimination and their interactions are shown in (B) where the genes are clustered to three groups, and shows group-specific regulation patterns, except two genes between AML-specific gene group and T-ALL specific gene group.

Figure 4A shows the FDA application to gene expression data of 10 samples from 5 normal and 5 malignant oral epithelium tissues. Wilks' lambda identified the 171 most discriminating genes from a total of 7070 genes and the error rate selected the 35 genes out of the 171 genes for use of FDA. Only one DF can be generated from the ED. However, the statistic V can not determine the statistical significance of any DF, because the number of samples is so small that the resulting V value becomes negative. The projection of the 35 gene expressions separate the normal samples from the malignant samples in the 1-D discrimination space, characterizing oral epithelium

malignancy. The discriminant loadings in Figure 4B show that two groups of genes distinguished by their signs in the loadings are identified and are differently regulated (anti-regulated) to produce the two different physiological states, normal and malignancy. One group of genes with the positive coefficients, down-regulated in cancer tissues, includes NmU, aldehyde dehydrogenase 9 and 10, Her3, etc., while the other group with negative coefficients, up-regulated in cancer tissues, includes Ferritin, Urokinase plasminogen activator, Gro2 oncogene, etc (for detailed list of genes, see captions in Figure 4).

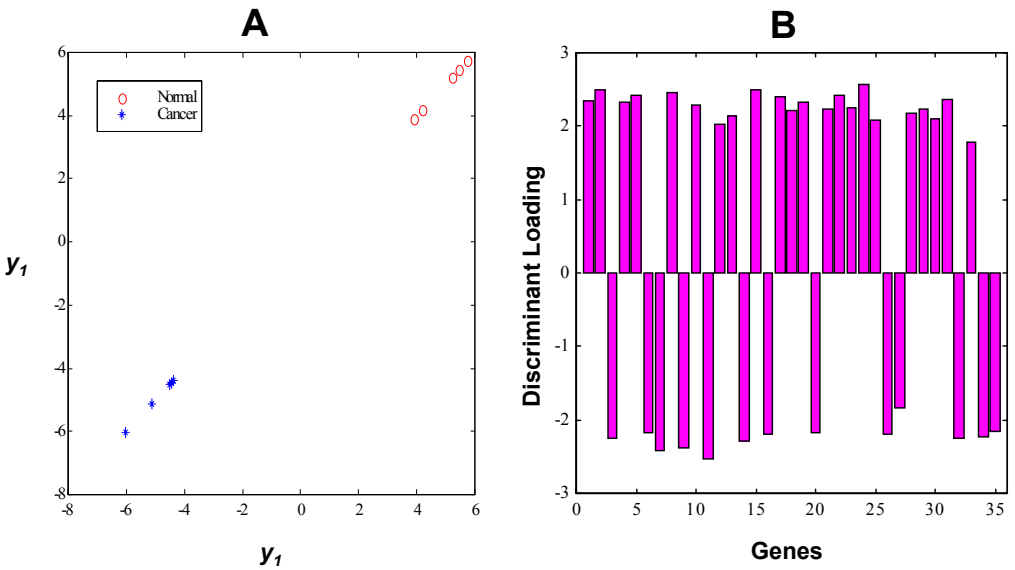

Figure 4. FDA projection of the expression phenotypes comprising 7070 genes measured in samples obtained from healthy individuals (5 samples) and patients with oral epithelium cancer (5 samples). (A) 35 discriminatory genes out of 7070 total genes allow FDA to clearly separate the two groups in one dimensional discrimination line. (B) Discriminant loading shows how 35 genes behave for separation in (A): positively co-regulated group includes NmU, aldehyde dehydrogenase 9 and 10, Her3, KIAA0089, diazepam binding inhibitor, monoamine oxidase B, crystallin alpha B, carboxylesterase 2, Wilm tumor-related protein, Zinc finger protein 273, MHC class I polypeptide related sequence A, Hpx-42, Lysophospholipase like, placental protein (PP11), cytochrome c oxidase subunit Vb, Cytochrome P4502C9 subfamily IIC, TF 20, FUT6, TYRO3, Keratin 4, and HLF. The negatively co-regulated group includes Ferritin, Urokinase plasminogen activator, Gro2 oncogene, 5T4 oncofetal trophoblast glycoprotein, HSP 90, Cathepsin L, Runt-related TF, Phospholipase A2, FAT tumor suppressor, macropain subunit zeta, CD38, TAL1 (SCL) and G protein-coupled receptor (AZ3B). These two groups are anti-correlated with respect to each other.

The compendium of gene expression data recently published by Rosetta Informatics (Hughes *et al*., 2000) was also analyzed. In this study, groups of related single-gene deletion mutants of yeast were identified on the basis of the presumed function of the deleted gene and also by applying clustering algorithms. Using the Euclidean distance as a metric of similarity between a drug-treated wild type sample and the mean of a group of deletion mutants, the drug compounds could be categorized as "most similar" to a set of related deletion mutants. 1756 genes met Wilks' lambda criterion of which 200 genes were selected for the FDA projection based on the error rate obtained from leave-one-out cross validation. By projecting the expression phenotype of four such groups onto the 3-D space defined by FDA, four distinct physiological states are identified describing genetic disruptions in mitochondrial activity, cell wall synthesis, ergosterol synthesis, and protein synthesis (Figure 5). For

this example, the statistic V can not determine whether the DFs are statistically significant, because of the small number of samples as in the third example. A classifier to discriminate the four classes in the deletion mutants was built using the gene expression measurements in the samples of the deletion mutants. The projections of Figure 5 show how the action of a drug causes a nearly equivalent physiological state as a disruption through genetic deletion, based on assignments of the expression measurements in drug treated tissues to four deletion mutant groups.

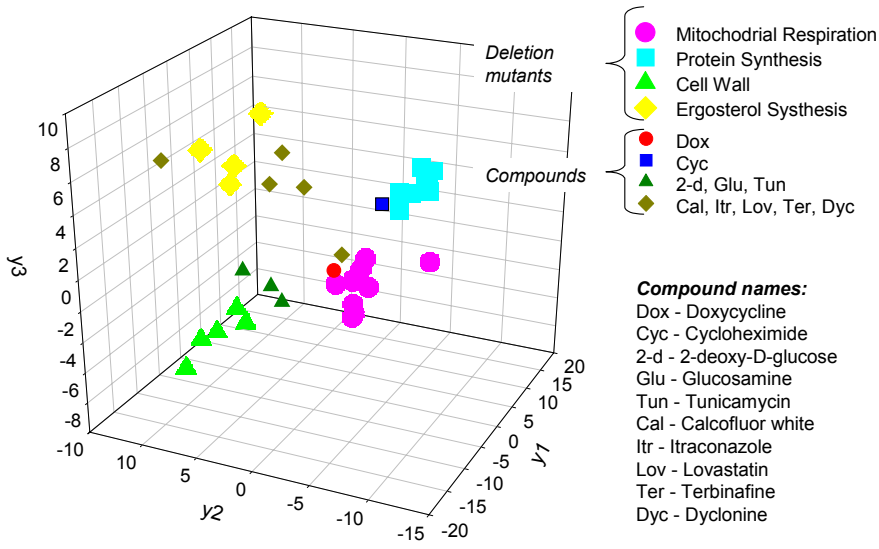

Figure 5. FDA projection of 27 yeast deletion mutant expression phenotype experiments grouped by the functionality of the eliminated gene. Four groups of related mutants have been distinguished using three DFs by projecting the expression levels of 200 of the most discriminating genes. The expression phenotypes obtained from the application of 10 chemical compounds to the wild-type yeast cultures are also projected into the FDA space defined by the mutants. The proximity in FDA space of these projections to those of the expression phenotype of the deletion mutant groups helps characterize the action of the compound on cell physiology. Note that one compound experiment (Cal) which appears incorrectly classified is actually in the center of the 3-D diagram, and not clearly associated with any of the groups shown. The classification suggested by the proximity of the projected phenotypes to the deletion mutants groups agrees with classification provided by Huges *et al*. (2000).

DISCUSSION

This study shows that for the four examples, the FDA properly achieves the four objectives mentioned earlier by integrating the information content of the large volumes of data in the expression phenotype: 1) generation of linear composites for the discrimination space, 2) determination of the dimensionality of the discrimination space and separability of the group centroids in the space, 3) development of a classifier using the significant DFs, 4) analysis of contribution of individual genes to the discrimination patterns observed.

A classifier and the important predictor variables characterizing the discrimination obtained from the application of FDA can be used to achieve various goals in medical and biotechnological areas. The FDA classifier can correctly diagnose

a disease subtype present in a sample and efficiently screen candidate drugs as shown in the last example (Figure 5). The important variables determined through FDA can help find a way to make a desired change in the physiological state that restores the expression phenotype to that of a normal tissue. In addition, the group specific genes shown in Figure 3B and Figure B4 can be used as markers in disease diagnosis. In the biotechnological area, the important variables can help control bioreactor to establish a desirable pattern of gene expression that corresponds to high productivity based on the classifier and the important variables.

The FDA is an optimal classification procedure in the sense of producing the smallest misclassification error rates under the following assumptions: 1) multivariate normality of the *g* pedictor variables, and 2) equal covariance matrices in each of the *c* groups. For a data meeting these two assumptions, the FDA produces the same classification rule and thus the same error rate with maximum likelihood classification. Although the multivariate normality is not an essential factor in the FDA-based classification, the equality of covariance matrices is a key factor strongly affecting the classification performance. On the other hand, the equality of covariance matrices is highly influenced by the number of predictor variables. Thus, in most of gene expression measurements where covariance matrices are different to a moderate extent, the number of the predictor variables (genes) should be determined carefully to minimize the effect of inequality of the covariance matrix on classification performance.

Accordingly, an initial set of discriminatory genes should be screened based on Wilks' lambda and then a subset of those screened genes used for the FDA are further selected based on the error rate obtained from cross-validations. In general, too small number of genes does not provide the FDA classifier with complete discriminatory information, while too many genes contaminate the classifier with non-discriminatory information from the genes with overlapping expressions across the classes. The initially screened genes determined by considering Wilks' lambda is an upper bound as a set of discriminatory genes, because the initial set of genes includes many genes with small discriminatory characteristics whose F values (transformed from Wilks' lambda) are only slightly larger than the critical value. Thus, those genes with small discriminatory characteristics were eliminated based on error rate for use in FDA.

We note that other classification tools have been applied to high-throughput expression analysis before ranging from simple two-dimensional classsifiers (Tamayo *et al*., 1999) to sophisticated neural networks (Furey *et al*., 2000). The performance of such classifiers is variable and depends on the type and quality of data analyzed. Although a direct comparison of such methods is beyond the scope of this article, FDA outperforms other projection methods like SVD/PCA in separating sample classes in the reduced dimensional space. Additionally, it yielded very satisfactory classification results and direct visual evaluation in all cases tested. In general, we would suggest that conclusions about sample diagnosis and importance of individual genes in defining cell physiology be based on a consensus from a multiplicity of methods. In this sense, FDA belongs in the portfolio of techniques to be used for this purpose.

It should be noted that the FDA method requires that *a priori* classification of samples be provided. Although this was rather straightforward for the cases presented, we note that, in general, this is not a trivial matter. For example, samples may be classified as malignant without any note as to the type of specific cancer involved, or, in production systems, a state of low productivity may reflect more than one-expression phenotypes. Although such heterogeneous samples will generally produce less well-defined states in their FDA projections, one can take further steps to identify possible subdivisions within a particular physiological class using MANOVA in the discrimination space.

Although the expression phenotype is an ample measure of the cellular state, it is by no means a complete one. As the techniques for probing the proteomic and metabolic state of the cell improve, a more comprehensive picture of cellular state will emerge. This data can nevertheless be handled by the same projection approach described herein. In this way, the use of projections to describe physiological state is as flexible as the data available and will find further applications as the type and volume of data accumulate in the future.

ACKNOWLEDGEMENTS

This work was supported by the Engineering Research Program of the Office of Basic Energy Science at the Dept. of Energy, Grant No. DE-FG02-94ER-14487 and DE-FG02-99ER-15015. Additional support was provided by NIH grant number 1-RO1-DK58533-01. We also acknowledge the expression data from oral epithelium samples provided by the group of Dr. D. Wang, Harvard School of Dental Medicine.

REFERENCES

Alter, O., Brown, P.O. & Botstein, D. Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 10101-10106 (2000).

Dillon, W.R. & Goldstein, M. Multivariate Analysis. (Wiley, New York; 1984).

Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 14863-14868 (1998).

Furey, T.S. *et al*. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* **16**, 906-914. (2000).

Gill, R.T., Katsoulakis, E., Schmitt, W., Taroncher-Oldenburg, G. & Stephanopoulos, G. Dynamic transcriptional profiling of the light to dark

transition in Synechocystis sp. PCC6803. (submitted, 2001).

Golub, T.R. *et al*. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* **286**, 531-537 (1999).

Holter, N.S. *et al*. Fundamental patterns underlying gene expression profiles: Simplicity from complexity. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 8409-8414 (2000).

Hughes, T.R. *et al*. Functional discovery via a compendium of expression profiles. *Cell* **102**, 109-126 (2000).

Johnson, R.A. & Wichern, D.W. Applied Multivariate Statistical Analysis. (Prentice Hall, Englewood Cliffs, New Jersey; 1992).

Lockhart, D.J. *et al*. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology* **14**, 1675-1680 (1996).

SAS/STAT User's Guide, Edn. Fourth Edition. (SAS Institute Inc., Cray, N.C.; 1989).

Schena, M., Shalon, D., Davis, R.W. & Brown, P.O. Quantitative Monitoring Of Gene-Expression Patterns With a Complementary-Dna Microarray. *Science* **270**, 467-470 (1995).

Spanakis, E. & BroutyBoye, D. Discrimination of fibroblast subtypes by multivariate analysis of gene expression. *Int. J. Cancer* **71**, 402-409 (1997).

Stephanopoulos, G. Metabolic fluxes and metabolic engineering. *Metab Eng* **1**, 1-11 (1999).

Tamayo, P. *et al*. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 2907-2912 (1999).

Vallino, J.J. & Stephanopoulos, G. Metabolic Flux Distributions In Corynebacterium-Glutamicum During Growth and Lysine Overproduction. *Biotechnology and Bioengineering* **41**, 633-646 (1993).

Vangulik, W.M. & Heijnen, J.J. A Metabolic Network Stoichiometry Analysis Of Microbial-Growth and Product Formation. *Biotechnology and Bioengineering* **48**, 681-698 (1995).

Wen, X.L. *et al*. Large-scale temporal gene expression mapping of central nervous system development. *Proceedings Of the National Academy Of Sciences Of the United States Of America* **95**, 334-339 (1998).

Zhao, G. & Maclean, A. L. A comparison of canonical discriminant analysis and principal component analysis for spectral transformation. *Photogramm Eng. Rem.* **S 66**, 841-847 (2000).