Integrative Analysis of multiple large-scale molecular biological data

Sri Priya Ponnapalli

Literature Survey

Multidimensional Digital Signal Processing

Genomic Signal Processing Laboratory

The University of Texas at Austin, TX 78712

sripriyap@mail.utexas.edu

Spring 2005

ABSTRACT

The Genome Project has revitalized exploration in biological research. DNA microarray technology allows us to probe the genome and monitor gene expression levels on a genomic scale. The high throughput data generated by this technology promise to enhance fundamental understanding of processes on the molecular level and may prove useful in medical diagnosis, treatment and drug design. Analysis of this data requires mathematical tools that are adaptable to large quantities of data, while reducing the complexity of the data to make them comprehensible. This report surveys the different techniques currently used to analyze microarray data.

1. INTRODUCTION

The Genome Project has led to the discovery of thousands of new genes, an exhilarating reminder that much of the natural world remains to be explored at the molecular level. The objective is to discover things we neither knew nor expected, and to perceive relationships and connections among the biological elements, whether previously suspected or not.

DNA Microarray technology [1,2] provides a straightforward and natural vehicle to monitor molecular biological data on a genomic scale. This high-throughput technology has generated large bodies of information that may enhance the fundamental understanding of life on a molecular level and may prove useful in perceiving elementary aspects of growth and development. It also enables us to explore the underlying genetic causes of many human diseases and thereby aid in medical diagnosis, treatment and drug design.

The need for mathematical tools that are adaptable to large quantities of data, while reducing the complexity of the data to make them comprehensible is highlighted. A few important existing frameworks that provide a description for the data, in which the mathematical variables and operations may represent some biological reality have been summarized in this report.

2. BACKGROUND

With only a few exceptions, every cell of the body contains a full set of chromosomes and identical genes [3]. Only a fraction of these genes are turned on however, and it is this subset that is "**expressed**" that confers unique properties to each cell type. "**Gene expression**" is the term used to describe the transcription of the information contained within the **DNA**, the repository of genetic information, into messenger RNA (mRNA) molecules that are then translated into the proteins that perform most of the critical functions of cells. Gene expression is a highly complex and tightly regulated process that allows a cell to respond dynamically both to environmental stimuli and to its own changing needs [4].

Microarrays allow us to analyze the gene expression of thousands of genes simultaneously. A microarray consists of a small membrane or glass slide containing samples of genes arranged in a regular pattern. Therefore, it provides us with a complete blueprint of any organism we want to study. The tight connection between the function of a gene and its expression pattern make the data generated by these microarrays invaluable. Outlined below are some of the important frameworks for gene expression analysis.

3. SINGULAR VALUE DECOMPOSITION

Singular value decomposition (SVD) [5], a linear algebra technique, is a tool that may be used to analyze gene expression [6]. SVD is a type of matrix factorization [7] that may be used for tasks, such as reducing dimensionality and determining the modes of complex dynamical systems, and has properties that are useful for a variety of signal processing problems and applications [8].

If **A** is a real **m x n** matrix, then there exist orthogonal matrices

$$\mathbf{U=[u_{1},......,u_{m}]} \qquad \text{and} \qquad \mathbf{V=[v_{1},......,v_{n}]}$$

such that

$$\mathbf{U^{T}AV = diag(d_{1},d_{2},.....,d_{p})}$$

where **p=min{m,n}** and $d_1 \geq d_2 \geq ..... \geq d_p \geq 0$.

SVD transforms the data from row x column space to reduced "eigenrow" x "eigenarray" space, where the eigenrows (or eigenarrays) are orthonormal superpositions of the rows (or columns). It provides a method to mathematically discover and expose latent relations and correlations within data. When SVD is applied to genome-wide data, the mathematical variables and operations may be assigned biological meaning [6]. In expression data, the rows correspond to genes and the columns correspond to the arrays. Expression data is transformed from the genes x arrays space to reduced "eigengenes" x "eigenarrays" space as indicated in figure 1$^{\Psi}$.

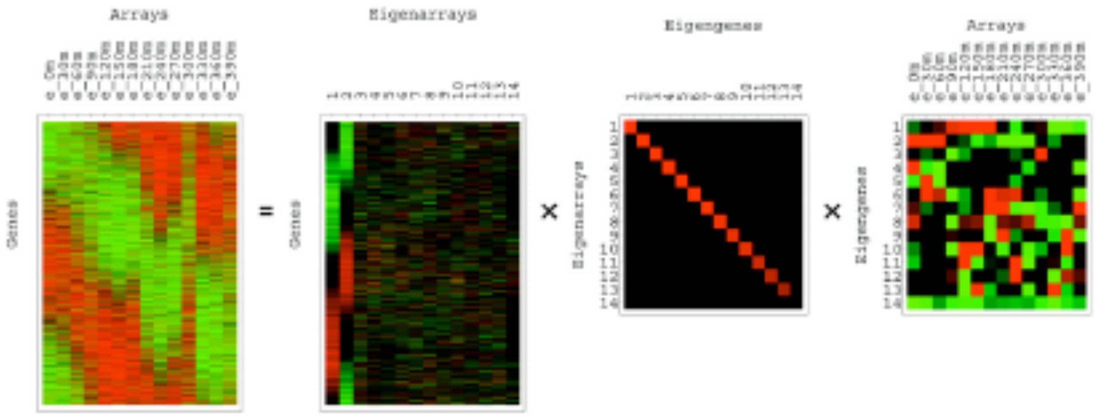

Figure 1$^{\Psi}$. Singular Value Decomposition

4. GENERALIZED SINGULAR VALUE DEOMPOSITION

Comparative analysis of two or more data sets promises to enable comprehension of the universality as well as the specialization of molecular biological mechanisms. To perform this comparison, we need mathematical tools that can distinguish the similar from the dissimilar among two or more large-scale datasets. Generalized singular value decomposition (GSVD) [5,7] is a tool that provides a comparative mathematical framework for two genome scale datasets. It allows comparative reconstruction and classification of the rows and columns of both datasets [9].

If $\mathbf{A_1}$ and $\mathbf{A_2}$ are two real matrices of dimensions $\mathbf{m_1 \ x \ n_1}$ and $\mathbf{m_2 \ x \ n_2}$, then there exist orthogonal matrices

$$\mathbf{U_1=[u_{1,}......,u_{m1}]} \qquad \mathbf{V=[v_{1,}......,v_n]}$$

And

$$\mathbf{U_2=[u_{1,}......,u_{m2}]} \qquad \mathbf{V=[v_{1,}......,v_n]}$$

such that

$$\mathbf{U_1{}^T A_1 V = diag1(d_1,d_2,.....,d_{p1})} \quad \text{and} \quad \mathbf{U_2{}^T A_2 V = diag1(d_1,d_2,.....,d_{p1})}$$

where $\mathbf{p1=min\{m_1,n_1\}}$, $\mathbf{p2=min\{m_2,n_2\}}$ and $\mathbf{n= n_{1=} n_2}$.

The one-to-one correspondence between the columns of the two matrices is at the foundation of the GSVD comparative framework and should be mapped out carefully. GSVD is the simultaneous linear transformation of the two expression data sets from $\mathbf{m_1}$ rows x $\mathbf{n}$ columns and $\mathbf{m_2}$ rows x $\mathbf{n}$ columns space to two reduced rowlets x columnlets spaces. In expression data, this would correspond to transformation two genes x arrays spaces to two reduced “genelets” x “arraylets” space as indicated in figure 2[Ψ].

[Ψ] In figures 1 and 2, Red indicates an increase in the expression of a gene and green indicates a decrease in the expression of a gene.

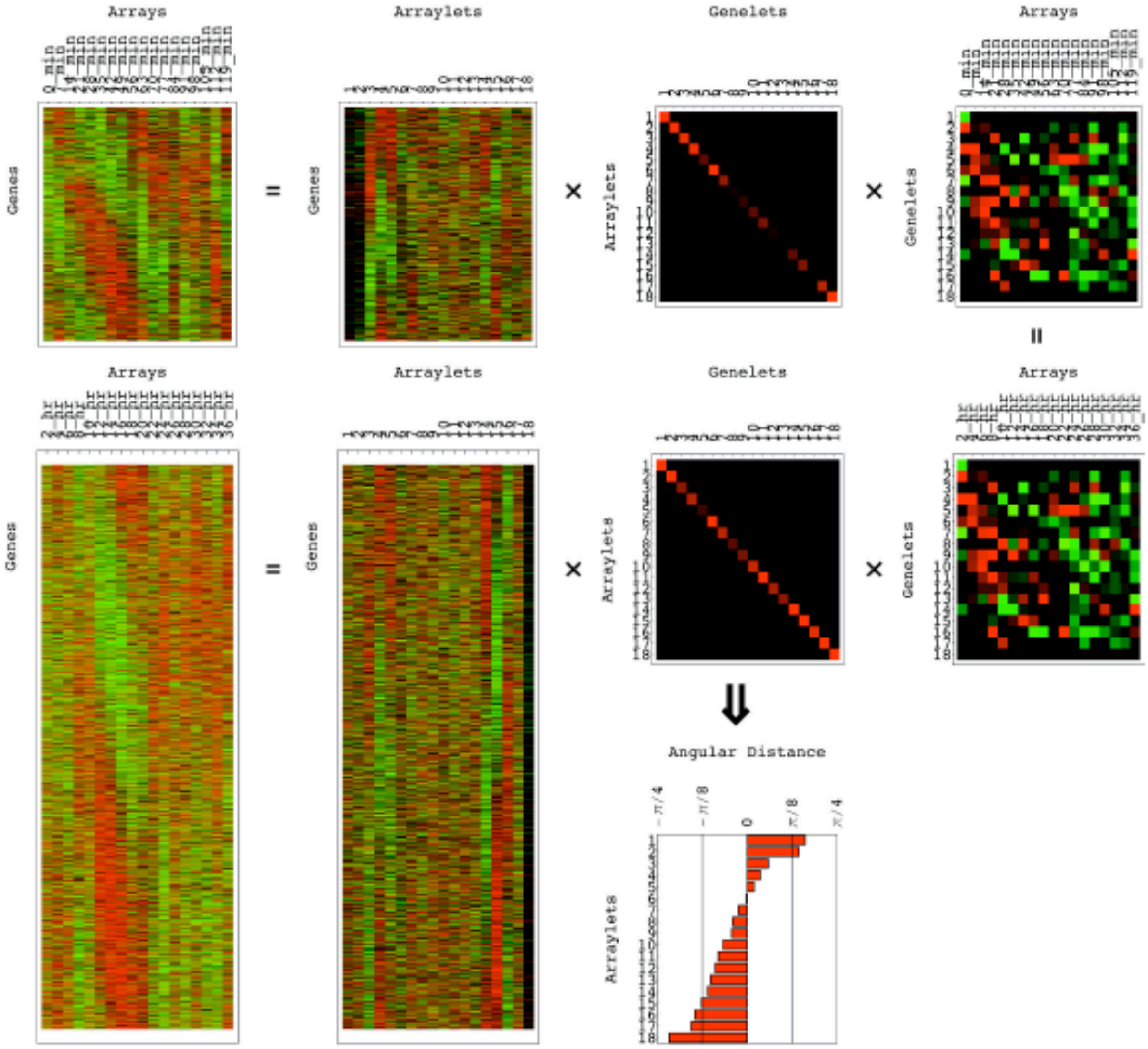

Figure 2$^{\Psi}$. Generalized Singular Value Decomposition

The anti-symmetric angular distance between the datasets

$$\boldsymbol{\theta = \arctan(diag1/diag2) - \Pi/4},$$

indicates the relative significance of genelets of the first dataset relative to those in the second in terms of the ratio of the expression information captured by the genelets in the first dataset relative to that in the second.

5. CLUSTERING ANALYSIS

In order to infer useful biological information and determine the relationships between individual genes, a system of "clustering" is available that can group similarly expressed genes into sub-groups, and, therefore, categorize genes appropriately. Using clustering analysis, one can identify sets of genes that are coordinately regulated. Information about a gene's function can be deduced by identifying genes that share its expression pattern. Genes that belong to the same cluster may be involved in common cellular processes. Clustering may be divided into two broad categories, Hierarchical Clustering and Partitional Clustering.

Hierarchical clustering [10] can be divisive or agglomerative producing nested clusters and the results are usually visualised by treelike dendrograms. Different similarity metrics such as Euclidean distance, Correlation and dot product may be used.

Partitional clustering divides data into a (pre-) chosen number of classes. K-means [11] that belongs to this class is typically used to cluster gene expression data. The drawback of this method is that choosing the number of clusters is difficult. Partitional clustering may be further divided into hard clustering and soft clustering. Hard clustering assigns a gene to exactly one cluster whereas soft clustering can assign a gene to several clusters.

6. PRINCIPAL COMPONENTS ANALYSIS (PCA)

PCA [11] is a statistical technique that has found application in identifying patterns in data of high dimensions. It enables us to express data in such a way, as to highlight the similarities and differences. PCA reduces the dimensionality of the data and so serves as a powerful tool in higher dimensions, where the luxury of graphical representation is not available. This makes it suitable for use on microarray data.

PCA is also based on eigen-decomposition. It is the linear projection of data onto the principal components defined by the eigen vectors of the covariance matrix. PCA is the statistical equivalent of SVD. When applied to gene expression data, it generates the **diagona**l matrix containing the eigenvalues and the **V** matrix consisting of the eigengenes. It does not however generate the **U** matrix that represents the eigenarrays.

Table 1. A Qualitative Comparison of Existent Microarray Data Analysis Techniques

Analysis Technique	Degree of Computational Complexity	Approximation of Biology	Visualization of Results	Experimental artifacts	Robustness to noise
SVD [Alter et al, 2000] [6]	Low	Very Good	Very Good	Can be detected and filtered out	Very Good
GSVD [Alter et al, 2003] [9]	Low	Very Good	Very Good	Can be detected and filtered out	Very Good
PCA [Nishimura et al, 2003] [12]	Low	Good	Good	Can be detected and filtered out	Good
Hard Clustering	High	Good	Good	Performance degrades markedly	Bad
Soft Clustering	High	Very Good	Good	Performance degrades	Good
Hierarchical Clustering [Eisen et al, 1998] [10]	Very High	Good	Very Good	Performance degrades	Reasonable
Gene Shaving [Hastie et al, 2000] [13]	High	Very Good	Good	Performance degrades	Reasonable
Plaid Models [Lazzeroni et al, 2002] [14]	Very High	Very Good	Good	Performance degrades	Good
Probabilistic Relational Models (PRM) [Segal et al, 2001] [15]	Very High	Very Good	Good	Performance Degrades	Good

7 CONCLUSION

This survey evaluates the various techniques currently employed to analyze microarray data and draws our attention to a deficiency in these methods in that they cannot be extended to collectively analyze more than 2 datasets. This project aims at addressing this issue by building a consolidated method to enable the integrative analysis of three or more data sets and discover relations between them. Both SVD and GSVD methods will be utilized in doing so. The National Cancer Institute's Developmental Therapeutics Program (DTP) has carried out intensive studies of 60 cancer cell lines, the NCI60, derived from a variety of human tissues and organs. These cell lines have been extensively used as experimental models of neoplastic disease to screen potential anticancer drugs. Among the chief objectives is to relate three data sets containing attributes of the NCI60: the gene expression profiles of the NCI60 [16], the sensitivities of these cell lines to more than 70,000 different chemical compounds including all common chemotherapeutics [17], and the proteomic profiling of the NCI60 using reverse phase lysate micro arrays [18].

As most molecular markers and targets are proteins, analyzing the protein levels for abnormalities may give answers to fundamental functional and pharmacological questions. This project shall attempt to address a question that has intrigued researchers for years: how similar are expression and protein profiles at the RNA and protein levels. Also, it shall aid in identifying candidate diagnostic markers for distinguishing between various kinds of tumors and in differential diagnosis e.g. determine what type of chemotherapy to give.

References:

[1] S. P. Fodor, R. P. Rava, X. C. Huang, A. C. Pease, C. P. Holmes & C. L. Adams, " Multiplexed Biochemical Assays with Biological Chips", *Nature (London),* vol. 364, pp. 555-556, August 1993.

[2] M. Schena, D. Shalon, R. W. Davis & P. O. Brown, "Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray", *Science,* vol.270, Issue 5235, pp.467-470, October 1995.

[3] A Science Primer, National Center for Biotechnology information, " Microarrays: Chipping Away at the Mysteries of Science and Medicine," March 2004, http://www.ncbi.nlm.nih.gov/About/primer/microarrays.html.

[4] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts and P. Walter, *Molecular Biology of the Cell,* 4th ed. New York, NY: Garland Science, 2002.

[5] G. Golub and C. Loan, *Matrix Computations,* 3rd ed. Baltimore, MD: Johns Hopkins Univ. Press, 1996.

[6] O. Alter, P. O. Brown and D. Botstein, "D. Botstein, "Singular Value Decomposition For Genome-Wide Expression Data Processing and Modeling," *Proceedings of the National Academy of Sciences, vol.* 97, no.18, pp. 10101–10106, August 2000.

[7] L. N. Trefethen and D. Bau, *Numerical Linear Algebra.* Philadelphia, PA: Siam, 1997

[8] S. Lee and M. H. Hayes," Properties of the Singular Value Decomposition for Efficient Data Clustering", *IEEE Signal Processing Letters,* vol.11, no.11, November 2004.

[9] O. Alter, P. O. Brown and D. Botstein*,* "Generalized singular value decomposition for comparative analysis of genome-scale expression datasets of two different organisms", *Proceedings of the National Academy of Science USA 2003,* vol.100, no.6, pp.3351-3356, March 2003.

[10] M. B. Eisen, P. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns", *Proceedings of the National Academy of Sciences*, vol. 95, no. 25, pp.14863-14868, December 1998.

[11] A. Gasch and M. B. Eisen, "Exploring the conditional co-regulation of yeast gene expression through fuzzy k-means clustering", *Genome Biology*, vol. 3, no. 7, pp.0059.1–0059.22, March 2002.

[12] K. Nishimura, S. Ishikawa, K. Hirota, M. Hirose, K. Abe, S. Tsutsumi and H. Aburatani, "A PCA based method of Gene Expression Visual Analysis", *Genome Informatics,* vol. 14, pp. 346-347, 2003.

[13] T. Hastie, R. Tibshirani, R. Levy, L. Staudt, M. B. Eisen, A. Alizadeh, W. C. Chan, D. Botsstein and P. O. Brown, " 'Gene Shaving' as a method for identifying distinct sets of genes with similar expression patterns", *Genome Biology,* vol. 1, no. 2, pp. 0003.1-0003.21, August 2000.

[14] L. Lazzeroni and A. Owen, "Plaid models for Gene Expression data", *Statistica Sinca,* vol. 12, pp. 61-86 , 2002.

[15] D. Koller and A. Pfeffer, ``Probabilistic frame-based systems", *in Proceedings of the Fifteenth National Conference on Artificial Intelligence*, 1998, pp. 580 – 587.

[16] D. T. Ross, U. Scherf, M. B. Eisen, C. M. Perou, C. Rees, P. Spellman, V. Iyer, S. S. Jeffrey, M. V. de Rijn, M. Waltham, A. Pergamenschikov, J. C. F. Lee, D. Lashkari, D. Shalon, T. G. Myers, J. N. Weinstein, D. Botstein and P. O. Brown, "Systematic variation in gene expression patterns in human cancer cell lines", *Nature Genetics*, vol.24, pp.227-35, March 2000.

[17] U. Scherf, D. T. Ross, M. Waltham, L. H. Smith, J. K. Lee, L. Tanabe, K. W. Kohn, W. C. Reinhold, T. G. Myers, D. T. Andrews, D. A. Scudiero, M. B. Eisen, E. A. Sausville, Y. Pommier, D. Botstein, P. O. Brown, and J. N. Weinstein, "A Gene Expression Database for the Molecular Pharmacology of Cancer ", *Nature Genetics*, vol. 24, no.3, pp.236-24, March 2000.

[18] S. Nishizuka, L. Charboneau, L. Young, S. Major, W. C. Reinhold, M. Waltham, H. K. Mehr, K. J. Bussey, J. K. Lee, V Espina, P. J. Munson, E. Petricoin III, L. A. Liotta, and J. N. Weinstein, "Proteomic Profiling of the NCI-60 Cancer Cell Lines Using New High-density Reverse-phase Lysate Micro arrays", Proceedings *of the National Academy of Science USA 2003,* vol.100, no.24, pp.14229-14234, November 2003.