Integrative Analysis of multiple large-scale molecular biological data

Project Report

Multidimensional Digital Signal Processing

Sri Priya Ponnapalli

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. The high throughput data generated by DNA microarray technology promise to enhance fundamental understanding of processes on the molecular level. 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. Recent work showed that singular value decomposition (SVD) provides such a framework for genome-wide expression data [1]. Alter *et al.* [2] showed that generalized SVD (GSVD) provides a framework that can distinguish the similar from the dissimilar between two genome-scale data sets. This report describes a mathematical framework that allows the comparative analysis of multiple large- scale data sets. This framework is illustrated by relating three datasets containing attributes of the NCI60 human cancer cell lines.

I. 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 [3,4] 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. Singular value decomposition is one such tool in which the mathematical variables and operations represent some biological reality [1]. Generalized singular value decomposition offers a comparative framework for two genome-scale data sets.

Comparative analysis of two or more data sets promises to enable comprehension of the universality as well as the specialization of molecular biological mechanisms. This project addresses the issue of building a consolidated method to enable the integrative analysis of three or more data sets and discover relations between them. To illustrate this method, three datasets from the National Cancer Institute's Developmental Therapeutics Program's (DTP) studies on 60 human cancer cell lines, the NCI60, are used. These cell lines have been extensively used as experimental models of neoplastic disease to screen potential anticancer drugs. Each of the datasets contains an attribute of the NCI60:

the gene expression profiles of the NCI60 [5], the sensitivities of these cell lines to more than 70,000 different chemical compounds including all common chemotherapeutics [6], and the proteomic profiling of the NCI60 using reverse phase Lysate micro arrays [7].

II. Singular Value Decomposition (SVD)

Singular value decomposition (SVD) [8], a linear algebra technique, is a tool that may be used to analyze gene expression [1]. SVD is a type of matrix factorization [9] 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 $\mathbf{U= [u_{1,......,} u_m]}$ and $\mathbf{V=[v_{1,......,}v_n]}$ such that $\mathbf{U^TAV = diag\ (d_1, d2,\ldots.., d_p)}$ where $\mathbf{p=min\{m, n\}}$ and $\boldsymbol{d_1 \geq d_2 \geq \ldots.. \geq d_p \geq 0}$.

When SVD is applied to genome-wide data, the mathematical variables and operations may be assigned biological meaning [1]. SVD transforms expression data from the genes x arrays space to reduced "eigengenes" x "eigenarrays" space.

Pattern Inference: The de-correlated eigengenes (and eigenarrays) represent independent biological processes (and corresponding biological states). The process represented by an eigengene is inferred from its expression pattern across all the arrays, which is usually biologically interpretable.

III. Generalized Singular Value Decomposition (GSVD)

Generalized singular value decomposition (GSVD) [8,9] 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 [2].

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: $\mathbf{U_1= [u_{1,......,}u_{m1}],V= [v_{1,......,}v_n]}$ and $\mathbf{U_2= [u_{1,......,}u_{m2}],\ V= [v_{1,......,}v_n]}$ such that,

$$\mathbf{A_1= U_1\ diag1(d_1,d_2,\ldots..,d_{p1})\ V^{-1}} \text{ and } \mathbf{A_2= U_2\ diag2(d_1,d_2,\ldots..,d_{p2})\ V^{-1}} \qquad \mathbf{(\ A\ )}$$

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 **n** columns and $\mathbf{m_2}$ rows x **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.

The anti-symmetric angular distance between the datasets $\boldsymbol{\theta}$**=arctan (diag1/diag2) –** Π **/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.

Pattern Inference: These genelets represent regulatory programs and biological processes that contribute to the overall expression signal in each data set. An angular distance of 0 indicates a genelet of equal significance in both data sets and that of +/- Π /4 indicate no significance in the second data set relative to the first.

IV. Proposed Centroid Method

A novel method for extending the principle of GSVD to multiple datasets where a one-to-one correspondence between the columns is conserved is proposed. The method also seeks to establish a framework for handling and correlating multiple data sets. For ease of understanding, consider the case of relating three datasets. The aim is to best approximate the decomposition:

$$\begin{aligned} &\mathbf{D_1 \sim U_1\ diag1(d_1,d_2,\ldots..,d_{p1})\ V^{-1}} \\ &\mathbf{D_2 \sim U_2\ diag2(d_1,d_2,\ldots..,d_{p2})\ V^{-1}} \qquad \mathbf{(B)} \\ &\mathbf{D_3 \sim U_3\ diag3(d_1,d_2,\ldots..,d_{p3})\ V^{-1}} \end{aligned}$$

IV (a) Geometric Interpretation

If a dataset is thought of as a line, the GSVD as in **(A)** of two datasets gives us the matrix $\mathbf{V^{-1}}$ that is common to both datasets and may be thought of as the point of intersection of the lines represented by the two datasets. Extending this analogy to three data sets D_1, $D_{2,}$ and $D_{3,}$ each of these data sets represents one line. Any two non-parallel lines intersect at a point. Three non-parallel lines form a triangle. In the ideal case, the three lines are concurrent and all three vertices of the triangle converge to the point of concurrency, making the approximation **(B)** exact.

Three points of intersection corresponding to applying GSVD on each of the 3 pairs of data sets are obtained. Let A, B, and C be the $\mathbf{V^{-1}}$ matrices obtained as a result of applying GSVD on $\{D_1, D_2\}$, $\{D_2, D_3\}$ and $\{D_3, D_1\}$ respectively. The aim of the method is to replace A, B and C with a matrix X that optimizes the approximation **(B)**.

For X to best approximate A, B and C simultaneously, the Frobenius distance between the matrices must be minimized. The Frobenius distance between two matrices X and Y is defined as $\| \mathbf{X} - \mathbf{Y} \|_{\mathbf{F}} = \sum_{i,j} | \mathbf{X_{ij}} - \mathbf{Y_{ij}} |^{2}$.

The matrix X that minimizes $\| \mathbf{X} - \mathbf{A} \|_{\mathbf{F}} + \| \mathbf{X} - \mathbf{B} \|_{\mathbf{F}} + \| \mathbf{X} - \mathbf{C} \|_{\mathbf{F}}$ must be estimated. This is nothing but minimizing the sum:

$$\sum_{i,j} | \mathbf{X_{ij}} - \mathbf{A_{ij}} |^{2} + \sum_{i,j} | \mathbf{X_{ij}} - \mathbf{B_{ij}} |^{2} + \sum_{i,j} | \mathbf{X_{ij}} - \mathbf{C_{ij}} |^{2}$$

As all the terms in this summation are non-negative, minimizing the above equation is equivalent to minimizing $| \mathbf{X} - \mathbf{A}|^{2} + | \mathbf{X} - \mathbf{B}|^{2} + | \mathbf{X} - \mathbf{C}|^{2}$ i.e. each element in the diagonal matrix is minimized and this minimizes the Frobenius distance. To this effect, the equation is differentiated and set to 0. Note that differentiating the equation twice leads to the constant value 6(>0), which indicates that the value obtained at X, subject to the constraint, is a minimum value [10].

$\Rightarrow \quad \mathbf{2 \mid X - A| + 2 \mid X - B| + 2 \mid X - C| = 0}$

$\Rightarrow \quad \mathbf{6\,X = 2\,(A + B + C)}$

$\Rightarrow \quad \mathbf{X = (A + B + C)/3}$

The matrix X is the centroid of the triangle formed by A, B and C. When the matrix $\mathbf{V}^{-1}$ is equal to X, the centroid of the triangle, the approximation **(B)** is optimized. The simplicity of the solution lies in the ease with which this method can be extended to any number of working data sets.

The described method has been implemented using MATLAB v7.0.1. The data sets are first filtered to improve the quality of the data. Rows in the data sets with more than 20% missing entries are discarded. The remaining missing values are then replaced with the average value of that row. The GSVD of every two pairs of data sets is computed to find their "intersection" i.e. the $\mathbf{V}^{-1}$ matrix. The centroid of the three $\mathbf{V}^{-1}$ matrices is then computed. Each data set is then multiplied with the inverse and the values in the diagonal matrices are calculated by normalizing the column vectors of the resulting matrices.

V. Results

The centroid method was applied to the three NCI60 data sets. Cancer cell lines corresponding to Melanoma, Colon, Ovarian, Leukemia, and breast tumor were used while cell lines corresponding to prostrate and lung cancer were discarded, owing to noisy data and inadequate number of samples. The RNA expression dataset records the expression of 4636 genes across 32 tumor samples. The protein dataset had 52 antigens across 32 tumors and the drug dataset records the reaction of the 32 tumors to 1400 drugs. The Centroid matrix containing patterns shared by all three data sets is shown in figure 1.

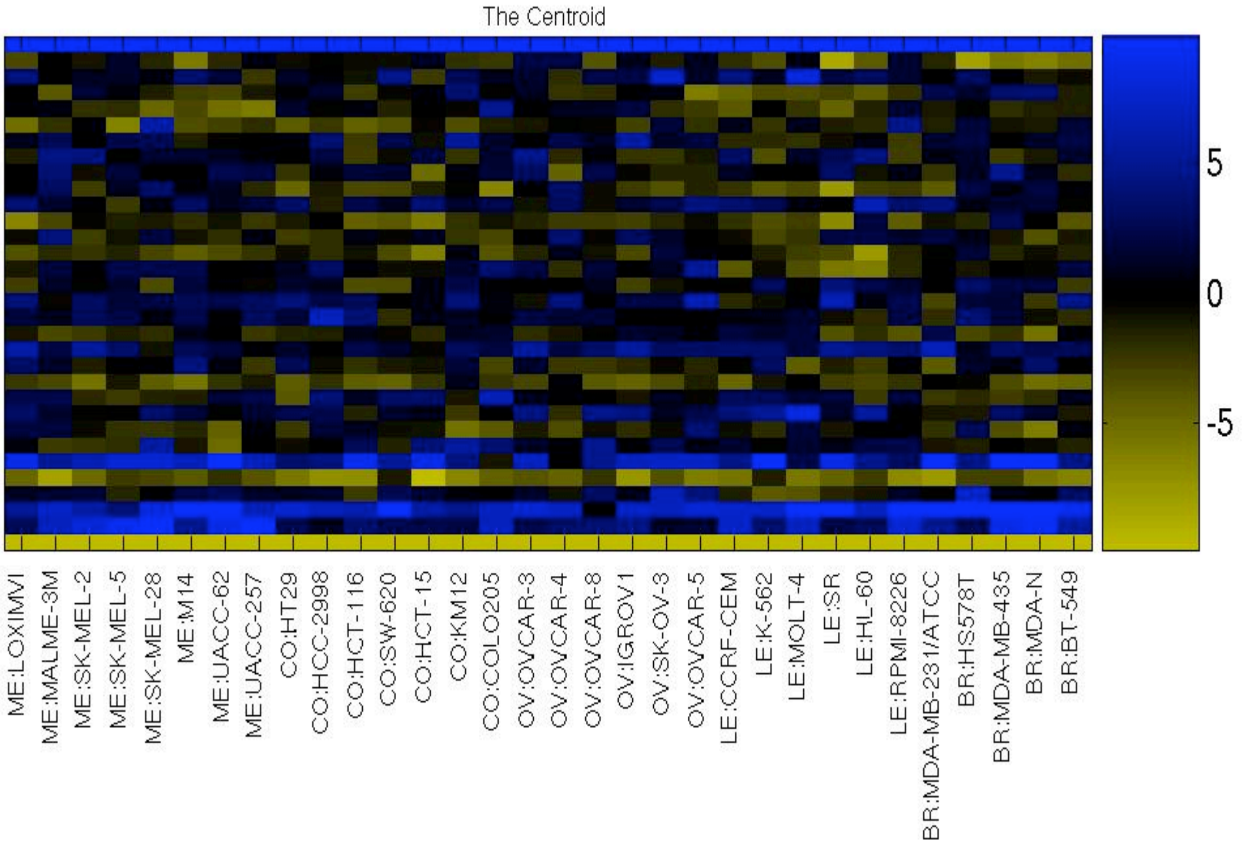

Figure. 1. Raster plot of the Centroid containing patterns shared by the RNA expression, Protein and Drug data sets with over expression (blue), no change in expression (black), and under expression (yellow).

The expression levels of the first two patterns of this Centroid are plotted in figure 2. Some interesting observations can be made from these patterns. For instance, the breast cancer and ovarian cancer lines are clustered together in the first pattern. It is known that the same hormone estrogen plays an important role in both these organs, which appear to be behaving in a similar fashion as indicated by their RNA expression and protein profiles. This suggests that similar drugs may be used to treat these cancers. Analyzing all patterns of the Centroid will enable us to draw similar inferences. An objective measure of calculating the relative significance of each of the data sets (analogous to the angular distance of the GSVD method) in the centroid is under development.

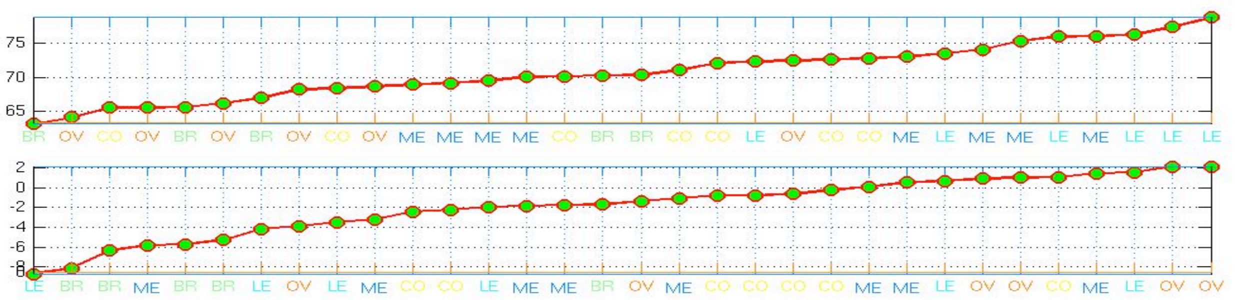

Figure. 2. Expression levels of the first two patterns of the Centroid.

VI. Comparison with existing frameworks

It is often very useful to collectively analyze more then two datasets. When compared to the existing frameworks of SVD and GSVD, the proposed centroid method overcomes the limitation of being restricted to a maximum of two data sets at a time. It may be very easily extended to allow an integrative analysis of any number of data sets as long as a one-to-one correspondence is conserved along the columns. Unlike SVD and GSVD however, it is not an exact decomposition. It is however the best possible approximation as shown.

VII. Conclusion

This report puts forth a new mathematical framework for the comparative analysis of multiple datasets. Using this method, patterns shared by three data sets containing attributes of the NCI60 were discovered. Future work includes attempting to address the intriguing question as to how similar are expression and protein profiles at the RNA and protein levels. Also, the discovered patterns will be further studied to identify candidate diagnostic markers for distinguishing between various kinds of tumors and in differential diagnosis.

References:

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

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

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

[10] K. Anurag (private communication), 2005.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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

[16] 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.