

Comparative Analysis of Gene Expression and DNA Copy Number Data for Pancreatic and Breast Cancers using an Orthogonal Decomposition

John A. Berger^{1,*}, Sampsa Hautaniemi², and Sanjit K. Mitra¹

¹Department of Electrical and Computer Engineering
University of California, Santa Barbara, CA 93106-9560, USA

²Tampere International Center for Signal Processing
Tampere University of Technology; P.O. Box 553, FIN-33101 Tampere, FINLAND

Abstract

The causes of over-expression for many diseases are typically unknown, but current studies show that copy number aberrations may be strong candidates for driving gene over-expression. We present the use of the generalized singular value decomposition (GSVD) for simultaneously identifying relevant influences common to only copy numbers, gene expression, or both measurements in conjunction. These groups are reported and gene ontology (GO) annotations are used as a functional assessment of the groupings accompanied by probabilistic significance obtained by combinatorics. We illustrate this method for two independently published studies of pancreatic cancer and breast cancer, where public gene expression and DNA copy number data is provided and measured across numerous tumor cell lines.

1. Introduction

The use of gene expression and copy number data is combined for analyzing two separate cancer studies under the GSVD framework. We locate specific inference patterns and processes common to both genome-wide input measurements across 10k unique clones. Our results suggest that several genes are characterized as having strong influences by both copy number variation and differential gene expression and lists of about 100 genes were compared to the cancer literature for both types of cancers. Preliminary analysis of gene influence is assessed by using GO annotations for each reported group of genes. Accordingly, these genes are likely to be closely involved in cancer development and progression, and could be promising targets for therapeutic intervention.

* This work was supported in part by a University of California MICRO grant with matching support from Philips Research Laboratories and in part by Microsoft Corp.

2. Results

Two publicly available data sets were separately utilized in this study of the effectiveness of the GSVD. Fourteen breast cancer cell lines (BT-20, BT-474, HCC-1428, Hs578t, MCF7, MDA-361, MDA-436, MDA-453, MDA-468, SKBR-3, T47D, UACC-812, ZR-75-1, and ZR-75-30) were selected for analysis of copy numbers and gene expression profiles using cDNA microarrays. Each breast cancer slide contained 13,824 original cDNA clones before preprocessing. Details on the microarray protocols, image processing, cell line labels, and the original data are given in Hyman *et al.* [4]. In addition, thirteen pancreatic cancer cell lines (AsPC-1, BxPC-3, Capan-1, Capan-2, CFPAC-1, HPAC, HPAF-II, HS 700T, Hs 766T, MIA PaCa-2, PANC-1, SU.86.86, and SW 1990) were selected for analysis of copy numbers and expression profiles. Each pancreatic cancer slide contained 12,232 transcripts before preprocessing. Further information is provided by Mahlamäki *et al.* [5].

Multi-input data was carefully preprocessed as follows. If data from either cancer study was missing from one or more samples in either gene expression or copy number data, that transcript was removed across all samples. No missing data in this study was estimated or inferred. Data was further processed by DEARRAY software where unreliable ratios were discarded. Consequently, over half of the transcripts from both cancer experiments were eventually discarded across all samples.

2.1. Generalized Singular Value Decomposition

The GSVD is the simultaneous linear transformation of two data sets $\mathbf{R}_a \in \mathbb{R}^{N_a \times m}$ with $N_a \geq m$ and $\mathbf{R}_b \in \mathbb{R}^{N_b \times m}$ to the reduced $m \times m$ space [3],

$$\begin{aligned} \mathbf{U}^T \mathbf{R}_a \mathbf{X} &= \mathbf{C} = \text{diag}(c_1, \dots, c_m) \quad c_i \geq 0, \\ \mathbf{V}^T \mathbf{R}_b \mathbf{X} &= \mathbf{S} = \text{diag}(s_1, \dots, s_m) \quad s_i \geq 0. \end{aligned} \quad (1)$$

This framework had originally been proposed for comparing the expression profiles of two completely different genomes by Alter *et al.* [1]. For comparative cDNA and CGH analysis, we are concerned with generalizing the SVD to two matrices with the same number of rows (genes) and columns (samples), $\mathbf{R}_a, \mathbf{R}_b \in \mathbb{R}^{N \times m}$, i.e. $N_a = N_b = N$. Hence, for each transcript we have expression measurements in \mathbf{R}_a and copy number measurements in \mathbf{R}_b across m cell lines. Correspondingly in Eq. (1), the matrices \mathbf{U} and \mathbf{V} are real, orthogonal $N \times m$ matrices and $\mathbf{X} \in \mathbb{R}^{m \times m}$ is a shared, invertible matrix.

We assume that the m generalized singular value pairs (c_i, s_i) of \mathbf{C} and \mathbf{S} are ordered such that

$$c_1 \geq c_2 \geq \dots \geq c_m, \quad s_1 \leq s_2 \leq \dots \leq s_m. \quad (2)$$

The ratios $\sigma_i = c_i/s_i$, for $i = 1, \dots, m$, are called the generalized singular values. It follows that $\sigma_i = (1 + \tan \theta_i)/(1 - \tan \theta_i)$ and in some situations, we need to only use one angle θ_i to represent either a generalized singular value pair or a generalized singular value. An angular distance of 0 indicates that genes may be of equal significance in both data sets, with $c_i = s_i$. An angular distance of $\pm\pi/4$ indicates no significance in the second data set relative to the first, with $c_i \gg s_i$, or in the first relative to the second $c_i \ll s_i$. The angular distances are arranged in decreasing order of significance in the first data set relative to the second such that $\pi/4 \leq \theta_1 \leq \dots \leq \theta_n \leq -\pi/4$.

For pancreatic cancer, we examined the projections of data that were highly significant to gene expression, θ_1 , highly significant to copy numbers, θ_{13} , and relevant to both, θ_4 . Expression and copy number values were sorted and the top $n_t = 50$ and bottom $n_b = 50$ genes were selected for gene ontology assessment for functional analysis. In addition, for the breast cancer data we separately performed the same analysis using the respective projections of this data.

2.2. Gene Ontologies

The n interesting genes were assessed their relevance by GO annotations as follows. For a given GO category F , a gene is either in the category or not in the category. Suppose that K out of the N reference genes and k out of the n interesting genes are in category F . Ultimately, we want to find out what is the probability of these k genes selected from n in F happening by chance. The probability that a certain category occurs k times just by chance in the list of selected genes is appropriately modelled by a hypergeometric distribution with parameters (N, K, n) :

$$P(k|N, K, n) = \frac{\binom{K}{k} \binom{N-K}{n-k}}{\binom{N}{n}}. \quad (3)$$

Based on this, the p -value of having k genes or fewer in F can be calculated by summing the probabilities of a random list of K genes having k genes of category F :

$$p = \sum_{i=k}^n \frac{\binom{K}{i} \binom{N-K}{n-i}}{\binom{N}{n}}. \quad (4)$$

See [2] for further details. GO annotations were studied using GoMiner software by Zeeberg *et al.* [6].

3. Conclusions

For the pancreatic cancer data, our analysis shows that the genes common to both gene over-expression and copy number amplification had high relevance to cell growth and maintenance. Transcripts that were highly influential to both gene under-expression and copy number deletion had high relevance to defense response processes. From the lists of genes selected under this criteria, we report very low p -values for a few ontology categories, therefore the probabilistic significance of these associations by annotations is high. Similar analysis was performed for the breast cancer data set and we found transcripts that were jointly influential to both gene over-expression and copy number amplification. Further analysis is currently underway to verify our lists of relevant genes in the context of pancreatic and breast cancer development and progression. Ultimately, we show that the GSVD framework is highly useful for analyzing genome-scale, multi-input data.

References

- [1] O. Alter, P. O. Brown, and D. Botstein. Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. *Proc. Natl. Acad. Sci. USA*, 100:3351–3356, Mar. 2003.
- [2] S. Draghici, P. Khatri, R. P. Martins, G. C. Ostermeier, and S. A. Krawetz. Global functional profiling of gene expression. *Genomics*, 81:98–104, Feb. 2003.
- [3] G. H. Golub and C. F. Van Loan. *Matrix Computations*. The Johns Hopkins Univ. Press, Baltimore, MD, 3rd edition, 1996.
- [4] E. Hyman, P. Kauraniemi, S. Hautaniemi, M. Wolf, S. Mousses, E. Rozenblum, M. Ringnér, G. Sauter, O. Monni, A. Elkahoulou, O.-P. Kallioniemi, and A. Kallioniemi. Impact of DNA amplification on gene expression patterns in breast cancer. *Canc. Res.*, 62:6240–6245, Nov. 2002.
- [5] E. Mahlamäki, P. Kauraniemi, O. Monni, M. Wolf, S. Hautaniemi, and A. Kallioniemi. High-resolution genomic and expression profiling reveals 105 putative amplification target genes in pancreatic cancer. *Neoplasia*, 2004. In Print.
- [6] B. R. Zeeberg, W. Feng, G. Wang, M. D. Wang, A. T. Fojo, M. Sunshine, S. Narasimhan, D. W. Kane, W. C. Reinhold, S. Lababidi, K. J. Bussey, J. Riss, J. C. Barrett, and J. N. Weinstein. GoMiner: A resource for biological interpretation of genomic and proteomic data. *Gen. Bio.*, 4(4):R28, Mar. 2003.