

# Application of Singular Value Decomposition and Functional Clustering to Analyzing Gene Expression Profiles of Renal Cell Carcinoma

Zhong-Hui Duan  
Department of Computer Science  
University of Akron  
Akron, OH 44325  
duan@uakron.edu

Louis S. Liou, Ting Shi, Joseph A. DiDonato  
Department of Cancer Biology  
Cleveland Clinic Foundation  
Cleveland, OH 44195  
{lioul, shiti, didonaj}@ccf.org

## Abstract

*Microarray gene expression profiles of both renal cell carcinoma (RCC) tissues and a RCC cell line were analyzed using singular value decomposition (SVD) and functional clustering. The SVD projections of the expression profiles revealed significant differences between the profiles of RCC tissues and a RCC cell line. Based on the biological processes, selected genes were annotated and clustered into functional groups. The analysis of each functional group revealed remarkable gene expression alterations in the biological pathways in RCC and provided insights into understanding the molecular mechanism of renal cell carcinogenesis.*

## 1. Introduction

The emergence of DNA microarray technology made it possible to investigate the expression of thousands of genes simultaneously [2, 4]. Recently, it has been widely used to identify and study gene expression patterns for many solid and hematological malignancies. To extract meaningful information from such massive data sets, many clustering and visualization methods have been developed, including hierarchical clustering, K-means, self organizing maps and SVD. In this research, we analyzed the gene expression profiles of both RCC tissues and a RCC cell line using SVD and functional clustering.

## 2. Methods

RCC tissue samples along with the patient-matched normal kidney tissues were obtained from six patients. The RCC tissues were divided into two groups with three samples in each group. The patient-matched normal samples were divided into two groups accordingly. Four pooled

total RNA samples from tissues and a total RNA sample from a metastatic RCC cell line were reverse-transcribed into cDNAs and hybridized to HuFL Genechips containing oligonucleotide probes for 7,129 human genes. The expression levels of the 7,129 genes were preprocessed to eliminate the genes whose signal intensities were not significantly different from the background level. 3,145 genes were selected after the preprocessing.

SVD has been widely used in data compression and visualization [3]. Recently there have been many applications of SVD to analyze microarray gene expression data [1]. The SVD of a real  $m \times n$  ( $m \geq n$ ) matrix  $A$  can be written as:  $A = U\Sigma V^T$ , where  $U = [u_1, \dots, u_n] \in R^{m \times n}$  and  $V = [v_1, \dots, v_n] \in R^{n \times n}$  are orthogonal matrices and  $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_n) \in R^{n \times n}$  is a diagonal matrix and  $\sigma_1 \geq \dots \geq \sigma_r > \sigma_{r+1} = \dots = \sigma_n = 0$ . The vectors  $u_i$  and  $v_i$  are the  $i$ th left and right singular vectors respectively,  $\sigma_i$  are the singular values of  $A$  and  $r$  is the rank of  $A$ . Based on the structure of the decomposition, the SVD expansion can be readily obtained:  $A = \sum_{i=1}^r \sigma_i u_i v_i^T$ . The magnitudes of singular values indicate how close a given matrix  $A$  is to a matrix of lower rank. In gene expression data analysis, each column of  $A$  represents the expression profile of a corresponding sample and each row represents the transcriptional response of a specific gene. The singular values indicate how well a lower dimensional linear projection of the expression data can represent the original data. In this study, the gene expression data were decomposed and projected onto a 2-D subspace spanned by the first two left singular vectors.

To analyze the expression profiles of genes in different biological functional groups, selected genes were annotated for biological process. The ontology is based on the description of the Gene Ontology Consortium. The annotated genes were then categorized into functional groups and analyzed based on the expression levels.

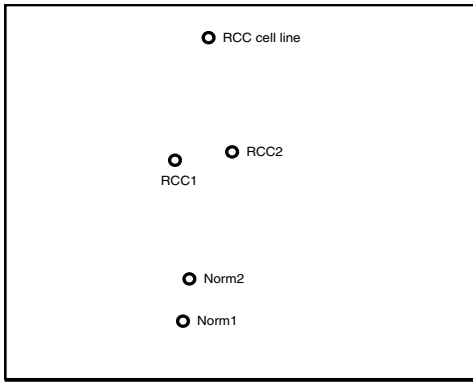


Figure 1. SVD projections.

### 3. Results and Discussion

The expression matrix of the five samples based on the 3,145 selected genes was decomposed using SVD. The resulting singular values {0.588, 0.176, 0.128, 0.0656, 0.0428} form a spectrum. We see clearly from the magnitude of the values that the first two singular vectors account for more than 76% of the total variance in the expression data. The projections of the five expression profiles onto the first two singular vectors are displayed in Figure 1. We are satisfied to see that the gene expression profiles of the two normal tissue samples were clustered together. The difference between the two normal profiles reflects the variations among different patients. The gene expression profiles of the two RCC tissue samples were clustered into a distinct group. More notably, we can clearly see that the profile of the RCC cell line is well separated from the tissue groups, indicating that the gene expression profile of the RCC cell line is significantly different from the profiles of either normal or RCC tissue samples.

1,340 selected genes for RCC tissues were annotated for biological process. The annotated genes were associated with 72 functional groups. 16% of the total 1,340 genes are up-regulated by at least 2-fold while only 9% of them are down-regulated. The majority (75%) of the genes are not differentially expressed. In many functional groups, such as cell adhesion, cell motility, proliferation, dominant majorities of the differentially expressed genes are up-regulated, which suggests these categories are in the up-regulated pathways and play significant roles in carcinogenesis. On the other hand, only very few categories appear to be in the down-regulated pathways. More notably, significant numbers of genes in metabolism and transport are down-regulated, although some important genes in the two groups such as manganese superoxide dismutase is up-regulated and its overexpression at protein level is also observed (T. Shi, et al., in preparation). These interesting gene

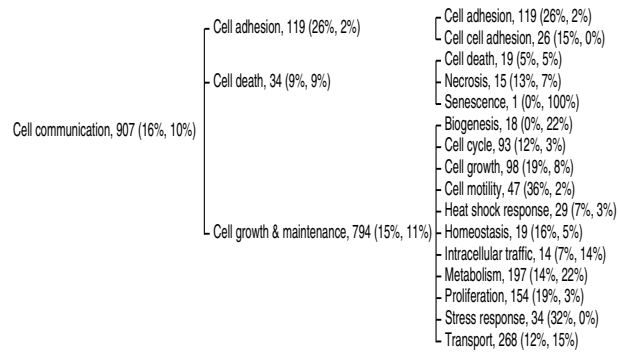


Figure 2. Cell communication ontology tree.

expression patterns in RCC tissue samples suggest important gene regulation pathways and also reveal remarkable variations in the gene expression levels even within a functional group such as metabolism. The gene ontology tree that describes the gene expression patterns in the functional group cell communication is shown in Figure 2. The first integer following the name of each functional group represents the number of genes associated with the group. The first percent number stands for the percentage of genes in the group that are at least two-fold up-regulated in average. The second number is the percentage of down-regulated genes. Gene expression patterns in the RCC cell line were also identified and analyzed (data not shown). The results show that much higher percentage of genes in the RCC cell line are differentially expressed than that in the RCC tissue samples. This result further confirms that the gene expression profile in RCC cell line is significantly different from that in RCC tissue samples as shown in Figure 1.

This study clearly demonstrates the usefulness of SVD and functional clustering for the analysis of large microarray gene expression data sets.

### References

- [1] O. Alter, P. O. Brown, and D. Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci.*, 97(18):10101–10106, August 2000.
- [2] J. DeRisi, L. Penland, P. O. Brown, M. L. Bittner, P. S. Meltzer, M. Ray, Y. Chen, Y. A. Su, and J. M. Trent. Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nature Genetics*, 14(4):457–460, December 1996.
- [3] G. H. Golub and C. F. van Loan. *Matrix computations*, 3rd edition. Johns Hopkins University Press, Baltimore and London, 1996.
- [4] S. Ramaswamy, K. N. Ross, E. S. Lander, and T. Golub. A molecular signature of metastasis in primary solid tumors. *Nature Genetics*, 33(1):49–54, January 2003.