DIMENSIONALITY REDUCTION FOR SAGE-BASED GENE IDENTIFICATION

Junior Barrera, Roberto M. Cesar Jr*, David C. Martins Jr†,
Paulo J. S. Silva
Department of Computer Science - Institute of Mathematics and Statistics - University of Sao Paulo - USP
Rua do Matao, 1010, Sao Paulo, SP

Helena Brentani, Elisson Osorio, Sandro J de Souza
Ludwig Institute of Cancer Research
Rua Prof. Antonio Prudente, 109 4th floor, Sao Paulo, SP, Brazil
Computational Biology Lab

Abstract

This abstract describes an ongoing research on dimensionality reduction methods [2, 4, 3] applied to SAGE data. The molecular pathways underlying brain cancers progression are poorly understood, making the development of novel diagnostic and therapeutic strategies difficult. Gene expression patterns are crucial for maintaining and altering phenotypes of cells. Recent technological advances have resulted in several widely used methods for large-scale study of gene expression, including comprehensive open systems, such as SAGE (Serial Analysis of Gene Expression). SAGE is a method to efficiently count large numbers of mRNA transcripts by sequencing short tags, usually 10 bp in length. SAGE Genie uses a new analytical method of reliably matching SAGE tags to known genes. SAGE can evaluate the expression patterns of tens of thousands of genes in a quantitative manner. Using SAGE Genie (http://cgap.nci.nih.gov/SAGE) we selected 22 brain libraries and the best tag for each full length represented in that library. SAGE profiles of 16 brain tumor libraries were compared with SAGE profiles of 6 normal brain libraries to identify differentially expressed genes. We constructed a matrix of known genes and their expression ratio in tumors/ normal SAGE libraries and tried to group genes with correlated expression profiles across tumor types. We used 4 different types of brain tumors and 4 different regions of normal brain.

The data has been normalized through the application of the normal transform [1] in order to allow the analysis of genes with low expression profiles. We have used the concept of strong genes sets introduced recently by Seungchang et. al. [3] to select differentially expressed genes. A strong gene set is a small group of genes that can resist to large errors in the gene expression measurement. Finally, we are using feature selection algorithms, like sparse support vector machines, to reduce the processing time and make it manageable by regular desktop computers.

*Supported by FAPESP and CNPQ - Brazil
†Supported by grant 02/04611-0 from FAPESP - Brazil

References

[1] L. da F. Costa and R. M. Cesar Jr. *Shape Analysis and Recognition: Theory and Practice.* CRC Press, 2001.

[2] R. O. Duda, P. E. Hart, and D. Stork. *Pattern Classification.* John Wiley & Sons, 2000.

[3] S. Kim, E. R. Dougherty, J. Barrera, Y. Chen, M. Bittner, and J. M. Trent. Strong feature set from small samples. *Journal of Computational Biology*, 2002. Accepted.

[4] S. Theodoridis and K. Koutroumbas. *Pattern Recognition.* Academic Press, 1999.