A PCA Based Method of Gene Expression Visual Analysis

Kunihiro Nishimura[1]
kuni@cyber.rcast.u-tokyo.ac.jp

Koji Abe[1]
abe@cyber.rcast.u-tokyo.ac.jp

Shumpei Ishikawa[2]
shumpei@genome.rcast.u-tokyo.ac.jp

Shuichi Tsutsumi[2]
shuichi@genome.rcast.u-tokyo.ac.jp

Koichi Hirota[2]
hirota@cyber.rcast.u-tokyo.ac.jp

Hiroyuki Aburatani[2]
haburata-tky@umin.ac.jp

Michitaka Hirose[2]
hirose@cyber.rcast.u-tokyo.ac.jp

[1] Graduate School of Engineering, The University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-0033, Japan
[2] Research Center for Advanced Science and Technology, the University of Tokyo, 4-6-1, Komaba, Meguro-ku, Tokyo, 153-8904, Japan

Keywords: visualization, PCA, gene expression analysis, annotation, chromosomal viewer

1 Introduction

Gene expression data has been rapidly accumulated and the methods of these data analysis are required. Statistical methods are used in these data analysis. However, the biological interpretation of the data and the result of the statistical analysis are difficult. Thus we are developing a method of analysis to interpret the data easily. We propose a PCA based analysis method and developed tools based on our proposal.

2 Method and Results

A process of human gene expression data analysis is followings: 1) pre-filtering, 2) statistical analysis, 3) interpretation of the data. First we reduce and eliminate a noise of the data as pre-filtering, that is, we select data, such as genes or samples, according to the liability. Second, we analyze the data statistically. Generally many researchers use hierarchical clustering analysis [1] and principal component analysis (PCA) [3] as statistical method. They check the visualized result and grasp patterns of the gene expression data. A result of clustering analysis is visualized as a dendrogram and it enables us to grasp relations among the elements. However, a result of PCA is visualized as a 2D or 3D scatter plot and it shows us both relations and degree of relations among the elements spatially. Thus we employ PCA from the viewpoint of grasping the total picture. Third, we interpret the data biologically. It is easier if the result of the analysis has its annotations. Thus we develop tools to annotate the data.

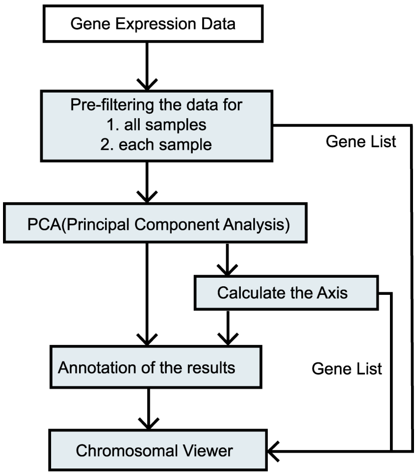

Figure 1: Analysis process we proposed Hatching boxes: developed tools.

On the basis on the gene expression analysis process, we propose the method of analysis using PCA as Fig. 1.

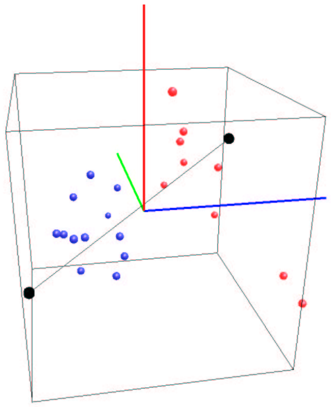

Figure 2: A PCA result of ALL 21 samples using 7,913 genes. Red: good prognosis (upper right), Blue: bad prognosis (lower left).

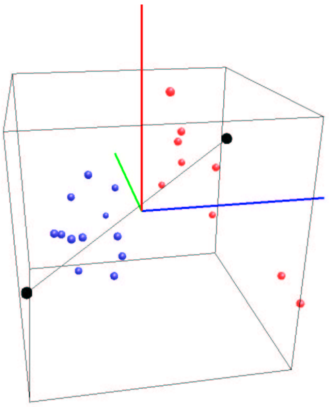

Figure 3: A PCA result of 705 genes using 39 human normal tissue gene expression data.

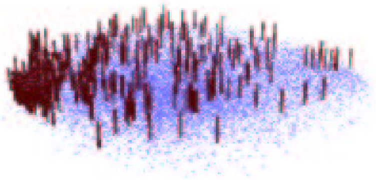

Figure 4: A PCA result of 23,375 genes using 39 human normal tissue gene expression data. 2D visualization of PCA result and GO annotation: RNA binding (GO:0003723).

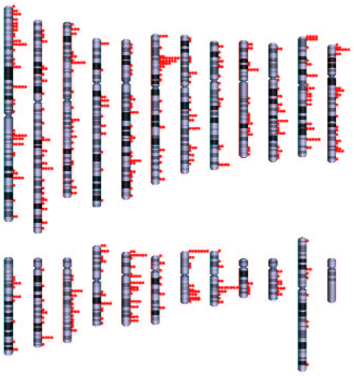

Figure 5: Chromosomal scatter plot of the upper 10% contributed genes (791) relevant to divide ALL prognoses (Fig. 2).

First process is pre-filtering of the gene expression data for all samples and for each sample. Pre-filtering for all samples uses coefficient of variance, standard deviation, maximum and minimum data for each gene. Pre-filtering for each sample uses the filtering thresholds of each sample [2]. After the pre-filtering, PCA is executed for samples (Fig. 2) and for genes (Fig. 3). The result is visualized using 3 principal components, for example, 1^{st}, 2^{nd}, and 3^{rd} principal components. An annotation of the data is added using color such as prognosis (Fig. 2). When the result is visualized 2D, 3^{rd} axis can be used an annotation axis. It is easier to grasp the pattern of annotation using 3^{rd} axis like Fig. 4. In case of Fig. 4, annotation information uses gene ontology (GO). Looking these results, users can set an axis that divides the elements freely (black axis of Fig. 2 and Fig. 3). Contributed components of this axis are calculated using principal components. For example, in case of Fig. 2, genes that contribute to divide groups are calculated and users get a gene list. The last process of our proposal is using chromosomal viewer (Fig. 5). Genes include the gene list are distributed to chromosome and visualized as a histogram for each band.

3 Discussion

We propose a PCA based method of gene expression visual analysis with calculating PCA contribution axis and adding GO annotations. We developed a tool to visualize a chromosomal histogram that can be useful to detect chromosomal imbalances such as chromosomal deletions and amplifications. We will analyze various gene expression data according to our proposal with tools we developed.

References

[1] Eisen M.B., Spellman, P.T., Brown, P.O., and Botstein, D., Cluster analysis and display of genome-wide expression patterns, *Proc. of the National Academy of Science*, 85:14863–14868, 1998.

[2] Nishimura, K., Ishikawa, S., Tsutsumi, S., Aburatani, H., Hirota, K., and Hirose, M., Gene selection using gene expression lata in a virtual environment, *Genome Informatics*, 13:276–277, 2002.

[3] Raychaudhuri, S., Stuart, J.M., and Altman, R.B., Principal components analysis to summarize microarray experiments : application to sporulation time series, *Pacific Symposium on Biocomputing*, 5:452–463, 2000.