# Visualization for Genome Function Analysis

Makoto Kanou, Kunihiro Nishimura, Koichi Hirota, Michitaka Hirose
Hiroyuki Aburatani, Takao Hamakubo and Tatsuhiko Kodama
*Research Center for Advanced Science and Technology, the University of Tokyo*
*E-Mail: {mkano, kuni, hirota, hirose}@cyber.rcast.u-tokyo.ac.jp*
*haburata-tky@umin.ac.jp*
*{hamakubo, kodama}@med.rcast.u-tokyo.ac.jp*

## Abstract

*In this paper, we discuss application possibilities of virtual reality technology such as immersive projection technology to the field of genome science. The prototype of the visualization environment is implemented and used in analysis to elucidate the important genes in categorizing liver cancer.*

## 1. Introduction

Recently, the analysis of genome function as a postgenome problem has come to attract researchers' interest. For the analysis, what is needed is a tool that will enable to display of the entire gene correlation intuitively. In this context, we believe that virtual reality (VR) technology can significantly contribute to genome science.

Our first target of research is to develop a visualization environment, by using immersive projection technology (IPT) such as CABIN (Figure1)[1]. This was implemented in our campus as a powerful VR environment that can be used for the purpose of enabling researchers to understand the gene expression level intuitively and interactively.

**Fig.1 CABIN**

As an example, we used 20 sample data of liver cancer, which can be categorized into three classes based on its progression: "cirrhosis of the liver" (L) (n=8), "well differentiated" (W) (n=6) and "moderately differentiated" (M) (n=6). Each sample data consists of 6817 types of gene expression level, which has been obtained from a DNA chip (Affymetrix Co.)

## 2. Methodology

The purposes of our visualization environment are to elucidate the important genes called "predictors" in categorizing liver cancer and to display differences between categories by using the predictors. Although it is difficult to identify predictors directly for the three categories (L, W, M), it is relatively easy to identify those for two categories. In combination with this pair-comparison method, we should be able to determine differences in gene expression patterns among three categories. Thus, we looked at three cases, (L-W), (L-M), and (W-M).

Our analysis process consists of three steps; "filtering," "selection" and "PCA" (principal component analysis).

**Filtering:** Measurement noise may be large at low gene expression values. Thus, we filtered out unreliable genes using the two thresholds below (eq. (1)).

$$T_{diff} = Max - min$$
$$T_{ratio} = Max / min \qquad (1)$$

where

*Max* : maximum expression of all genes of two classes
*min* : minimum expression of all genes of two classes

However, it is difficult to precisely to determine the threshold level. In order to solve this problem, we visualized the gene expression data as a set of vector $n(r,\phi,\theta)$. The average gene expression level of each sample can be mapped to r, and gene id g is encoded as $\phi$ and $\theta$. In addition, the function of the gene is mapped to the color of vector n. By implementing this visualization in the virtual environment

displayed in CABIN, the user can analyze the 3D gene expression data represented by the burr-like image. Researchers can compare the shapes of the two "burr" like objects. By changing Tdiff and Tratio, several "delicate" genes can be removed from the object, and the user can observe the change in the shape of the burr. In particular, normalizing the r of one burr to the r of the other burr helps researchers understand significant genes.

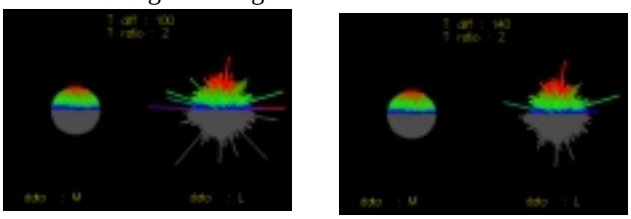

**Fig.2 Burr Model**
**(Left: Tdiff=100, Right:Tdiff=140)**

**Selection:** After filtering, we used the correlation function *P(g)* shown below (eq. (2)) to select predictors[2]. Then, the genes whose |*P(g)*| are sufficiently high can be considered as predictors.

$$P(g) = \frac{\mu_A - \mu_B}{\sigma_A + \sigma_B} \qquad (2)$$

where

g : gene ID

$\mu_i$ : average of "g's" expression data in category i

$\sigma_i$ : standard deviation of "g's" expression data in category i

However, since gene expression data are biological, they always have errors and extraordinary values. Therefore, if selection is carried out automatically, it is inevitable that unsuitable genes will be mixed with the predictors. Thus our visualization system displays the matrix shown in Figure 3. The candidates of predictors whose |*P(g)*| is high are arranged along the vertical axis of the matrix, and the samples for each candidate are arrayed along the horizontal axis. The system asks users to choose "adopt," "discard" or "pending."

The expression level of each gene is shown in color from blue to red. Blue means that the expression is weak, and red means that the expression is strong. In addition, a cross section of the row indicated by a small arrow is displayed on the matrix.

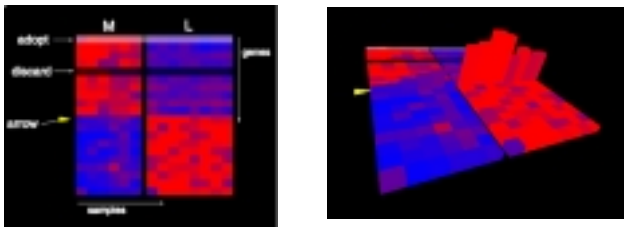


**Fig.3 Selection Matrix**

**PCA:** After selection, PCA is performed using the selected predictors. We adopt the first principal component vector only.

This process is repeated for (M-L), (W-M) and (L-W). The result of the analysis is displayed in CABIN. The x axis is set by the first principal component vector of (M-L), the y axis by that of (W-M) and the z axis by that of (L-W). Using the 3D virtual environment, researchers can check data separation from various points of view. Figure 4 shows that the three categories of the samples can be clearly divided in a 3D space.

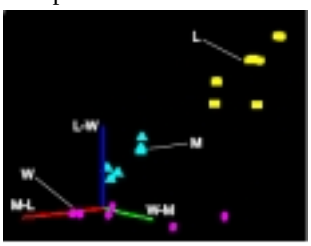


**Fig.4 The Result of PCA**

**Fig.5 The Implementation in CABIN**

## 3. Conclusion

This study is our first attempt to apply VR technology to the field of genome science. Unlike the conventional bioinformatics approach, we place importance on human capability rather than on automated computing. Using CABIN, it should be possible to have a visualization environment with a wide field of view that enables simultaneous viewing of all data.

In that context, we developed a prototype VR environment for identifying predictors that are characteristic of each category of cancer. When this environment was demonstrated to DNA researchers in our research center (RCAST), their response was very positive, and valuable suggestions helped define our future research direction. Although this environment is still in its initial stage, we believe that this kind of environment will be very useful in DNA research.

## References

[1] Hirose M., Ogi T., Ishiwata S., Yamada T., Development and Evaluation of the Imersive Multiscreen Display CABIN, Systems and Computers in Japan, Vol.30, No.1, pp.13-22, 1999.

[2] T.R. Golub, et al: "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring". Science, Oct.15, 1999. pp.531-537