and its advantage is obvious over simple spatial mapping of the expression profiles on chromosomal location. Therefore, the EIM would provide the user with further insight into the genomic structure through mRNA expression.

This work was supported by Grant-in-Aid for Scientific Research on Priority Areas (C) "Genome Information Science" from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

REFERENCES

1. **Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, Loda M, Weber G, Mark EJ, Lander ES, Wong W, Johnson BE, Golub TR, Sugarbaker DJ, and Meyerson M.** Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc Natl Acad Sci USA* 98: 13790–13795, 2001.
2. **Bonner RF, Emmert-Buck M, Cole K, Pohida T, Chuaqui R, Goldstein S, and Liotta LA.** Laser capture microdissection: molecular analysis of tissue. *Science* 278: 1481–1483, 1997.
3. **Fujii T, Dracheva T, Player A, Chacko S, Clifford R, Strausberg LS, Buetow K, Azumi N, Travis WD, and Jen J.** A preliminary transcriptome map of non-small cell lung cancer. *Cancer Res* 62: 3340–3346, 2002.
4. **Hayter AJ.** *Probability and Statistics for Engineers and Scientists* (2nd ed.). Florence, KY: Duxbury Press, 2002.
5. **Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F, and Pinkel D.** Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science* 258: 818–821, 1992.
6. **Lu YJ, Dong XY, Shipley J, Zhang RG, and Cheng SJ.** Chromosome 3 imbalances are the most frequent aberration found in non-small cell lung carcinoma. *Lung Cancer* 23: 61–66, 1999.
7. **Mukasa A, Ueki K, Matsumoto S, Tsutsumi S, Nishikawa R, Fujimaki T, Asai A, Kirino T, and Aburatani H.** Distinction in gene expression profiles of oligodendrogliomas with and without allelic loss of 1p. *Oncogene* 21: 3961–3968, 2002.
8. **Pei J, Balsara BR, Li W, Litwin S, Gabrielson E, Feder M, Jen J, and Testa JR.** Genomic imbalances in human lung adenocarcinomas and squamous cell carcinomas. *Genes Chromosomes Cancer* 31: 282–287, 2001.
9. **Petersen S, Aninat-Meyer M, Schluns K, Gellert K, Dietel M, and Petersen I.** Chromosomal alterations in the clonal evolution to the metastatic stage of squamous cell carcinomas of the lung. *Br J Cancer* 82: 65–73, 2000.
10. **Pollack JR, Perou CM, Alizadeh AA, Eisen MB, Pergamenschikov A, Williams CF, Jeffrey SS, Botstein D, and Brown PO.** Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat Genet* 23: 41–46, 1999.
11. **Reik W and Walter J.** Imprinting mechanisms in mammals. *Curr Opin Genet Dev* 8: 154–164, 1998.
12. **Virtaneva K, Wright FA, Tanner SM, Yuan B, Lemon WJ, Caligiuri MA, Bloomfield CD, de La Chapelle A, and Krahe R.** Expression profiling reveals fundamental biological differences in acute myeloid leukemia with isolated trisomy 8 and normal cytogenetics. *Proc Natl Acad Sci USA* 98: 1124–1129, 2001.

Articles in PresS. Physiol Genomics (February 15, 2005). doi:10.1152/physiolgenomics.00107.2004

A meta-clustering analysis indicates distinct pattern alteration between two series of Gene Expression profiles for induced ischemic tolerance in rats

Makoto Kano[(1)], Shuichi Tsutsumi[(2)], Nobutaka Kawahara [(3)(4)], Yan Wang [(3)], Akitake Mukasa [(2)(3)], Takaaki Kirino [(3)(4)] and Hiroyuki Aburatani[(2)]

[(1)]Intelligent Cooperative System, Department of Information Systems, Research Center for Advanced Science and Technology, University of Tokyo, 153-8904, Japan
[(2)]Genome Science Division, Research Center for Advanced Science and Technology, University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo 153-8904, Japan
[(3)]Department of Neurosurgery, Faculty of Medicine, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan
[(4)]SORST (Solution-Oriented Research for Science and Technology) / JST(Japan Science and Technology), 4-1-8 Honcho, Kawaguchi, Saitama, 332-0012, Japan

Running head: Visualization for Time-series Gene Expression Analysis

To whom requests for proofs are to be addressed:
Makoto Kano,
E-mail: mkano@cyber.rcast.u-tokyo.ac.jp

Abstract

We have developed a visualization methodology, called a *Cluster Overlap Distribution Map* (*CODM*), for comparing the clustering results of time-series gene expression profiles generated under two different conditions. Although various clustering algorithms for gene expression data have been proposed, there are few effective methods to compare clustering results for different conditions. Using *CODM*, the utilization of three-dimensional space and color allows intuitive visualization of changes in cluster set composition, changes in the expression patterns of genes between the two conditions, and relationship with other known gene information, such as transcription factors. We applied *CODM* to time-series gene expression profiles obtained from Rat 4-vessel occlusion models combined with systemic hypotension and time-matched sham control animals (with sham operation), identifying distinct pattern alteration between the two. Comparison of dynamic changes of time series gene expression levels under different conditions are important in various fields of gene expression profiling analysis, including toxicogenomics and pharmacogenomics. *CODM* will be valuable for various types of analyses within these fields since it integrates and simultaneously visualizes various types of information across clustering results.

Key words: time-series, transcription factor, visualization

1. Introduction

Advances in microarray technologies have made it possible to comprehensively measure gene expression profiles. Observation of dynamic changes of gene expression levels provides important markers to clarify cellular responses, differentiation, and genetic regulatory networks. In particular, a comparison of dynamic changes of time series gene expression levels under various conditions (e.g. administration of different drugs) is expected to make a major contribution to the understanding of complex biological processes. In general, we observe the influence of each condition through the results of clustering analysis, which is the most popular analysis for gene expression profiles. Therefore, a comparison between the results of clustering analyses in different conditions will allow interpretation of different macroscopic phenomenon that occurred under those conditions. However, although many clustering algorithms, including hierarchical clustering (1,2,4,15), k-nearest neighbor (17) and self-organizing maps (10,13,16) have been proposed, there are few effective methods to effectively compare clustering results under different conditions. We have defined four issues to be addressed for a comparison of clustering results, especially for a comparison of time series gene expression data under two different conditions: changes in the composition of the cluster sets, changes in the expression patterns, integration with known other gene information, and threshold problems.

Changes in the composition of the cluster sets

In this report, we focused on hierarchical clustering since it is the most popular method for gene expression analysis. Here we define the composition of a cluster set as the hierarchical structure of clustering results and "cluster set" as the set of all clusters in the structure. A comparison of clusters' compositions shows which clusters are conserved in different conditions and how the genes in a cluster for one condition are distributed into a cluster set under another condition. Genes that cluster under a single condition may possibly be regulated by the same factors for that condition. However, under different conditions, some of those genes would be regulated by other factors and generate different clusters. Thus, changes in the cluster compositions could provide key information for interpreting the effects of the different conditions. To get a full picture of the relationships of two cluster sets, the overlap between each pair of clusters under the two different conditions should be evaluated. However, since clustering analysis, especially hierarchical clustering, almost always generates a great number of clusters, there are a very large number of combinations of clusters. Simple line connections of the genes between the dendrograms of two hierarchical clustering results (14) provides insufficient information about the relationships between the clusters. Therefore, an effective presentation method that provides a full picture of the relationships of the cluster sets would be desirable.

Recently, a statistical model for performing meta-analysis of independent microarray datasets was proposed (12). This model revealed, for example, that four prostate cancer gene expression datasets shared significantly similar results, independent of the method and technology used. However, in a comparison of the cluster sets based on different conditions, the objective is not to confirm that several datasets share significantly similar results, but to detect the differences between them. Several statistical algorithms have been proposed for evaluating how clusters based on expression profiles include genes with well-known functions (3,17). However, the number of clusters that were compared was limited and an effective presentation method was not required in those situations.

Changes in the expression pattern

Where two clusters under different conditions have a statistically meaningful number of genes in common, it is also important to examine the differences in their expression patterns. The differences of macroscopic phenomena that the conditions exhibit result from the differences of expression of multiple, rather than single, genes. Therefore, the genes whose expression patterns changed in a similar fashion between different conditions provide markers for the different phenomena. In other words, if the genes in a certain cluster based on one condition also constitute a cluster for another condition, but the expression patterns are greatly different between the two conditions, these genes are causally implicated in the phenotypic difference.

In general, there will be many false candidate genes whose expression patterns coincidentally match between the two different conditions. Therefore, it is important to simultaneously evaluate the statistical significance of the overlaps between clusters and the differences in their expression patterns.

Integration with other known gene information

In gene expression analysis, it is important to biologically interpret the results after integrating them with other known gene information. Therefore, changes in the composition of the cluster sets and changes in the expression patterns between different conditions should be associated with other known gene information such as transcription factors.

Threshold problems

In a comparison of cluster sets on gene expression profiles, we have to handle four types of thresholds: 1) a threshold for generating clusters for each condition; 2) a threshold for evaluating the number of common genes that two clusters have; 3) a threshold for evaluating the differences in the expression patterns between two clusters; and 4) a threshold for evaluating the relationship with other known gene information. Among these, determining the threshold for generating clusters is most challenging, because the clustering result strongly depends on this threshold, and a change of this threshold greatly affects the number and composition of clusters. It is generally difficult to determine optimal values for these four types of thresholds, and the results of analysis are greatly affected by the threshold values specified. Arbitrary selection of thresholds involves a risk of overlooking important genes, so the number of thresholds should be reduced and, if used, it is necessary to allow users to interactively change the thresholds.

We focused on visualization technology to address these four issues. Interactive visualization is effective for handling ambiguous threshold problems and for providing a wide variety of information at one time. In previous work, we developed a *Cluster Overlap Distribution Map* (*CODM*), which is a visualization method for comparing cluster sets based on different sets of gene expression profiles (7). In this report, we extended it for time-series gene expression analysis. In the *CODM*, the relationships of all possible pairing of clusters can be examined and both the changes in the composition of the cluster sets and the changes in the expression patterns of the clusters can be effectively visualized as 3D histograms, without any arbitrary thresholds. In addition, relationships with other known gene information such as transcription factors can also be elucidated. We applied the *CODM* to a comparison between the gene expression datasets of double ischemia rats and sham control rats (with sham operation), and confirmed that *CODM* identified distinct patterns between the two.

CODM, available on our web site (http://www.genome.rcast.u-tokyo.ac.jp/CODM), runs on a PC with Windows 2000 or Windows XP. Memory requirement is in proportion to the square of the number of genes to be analyzed. The analysis for approximately 4000 genes, represented in this paper, required approximately 250 Mbytes. In addition, since the analysis results of the *CODM* are visualized by use of the OpenGL, a machine with a graphic board with a hardware accelerator for the OpenGL is recommended.

2. Materials and Methods

Experiment Design

In this report, *CODM* is illustrated using time-series gene expression datasets obtained from Rat 4-vessel occlusion models combined with systemic hypotension and time-matched control animals with sham operation. In the experiment, we used 2-minute ischemia rats with induced ischemic tolerance (*tolerant rats*: TOL) and rats with sham operation (*sham rats*: SHAM), after confirming the histological outcomes. Note that the sham rats did not acquire ischemic tolerance. Three days after the operation, we conducted a 6-minute ischemia operation on the two groups. Because of their ischemic tolerance, very little neuronal death of CA1 hippocampal neurons was observed in the tolerant rats (9). Using duplicate assessments of 6 time-points ({0h, 1h, 3h, 12h, 24h, 48h} x 2) after the second ischemia, microdissected CA1 regions from each of the two groups were subjected to oligonucleotide-based microarray analysis.

All animal-related procedures were conducted in accordance with guidelines for the care and use of laboratory animals set out by the National Institutes of Health and approved by the committee for the use of laboratory animals in the University of Tokyo. More detailed experimental design is described in our previous report (8).

Gene Chip experiment

Five μg of total RNA from each sample were used to synthesize biotin-labeled cRNA, which was then hybridized to a high-density oligonucleotide array (GeneChip Rat RG_U34A array, Affymetrix) essentially following a previously published protocol (6). The arrays contain probe sets for 8737 rat genes and ESTs, which were selected from Build 34 of the UniGene Database (derived form GenBank 107, dbEST/11-18-98). Sequences and GenBank accession numbers of all probe sets are available from the Affymetrix home page (http://www.affymetrix.com/index.affx.). Washing and staining was performed in a Fluidics Station 400 (Affymetrix) using the protocol EukGE-WS2. Scanning was performed on an Affymetrix GeneChip scanner to collect primary data. The Affymetrix Microarray Suite v4.0 was used to calculate the average difference for each gene probe on the array, which was shown as an intensity value of gene expression defined by Affymetrix using their algorithm. The average difference has been shown to quantitatively reflect the abundance of a particular mRNA molecule in a population (6). To allow comparison among multiple arrays, the average differences were normalized for each array by assigning the mean of overall average difference values to be 100. This dataset has been submitted as GSE1357 to the National Center for Biotechnology Information's Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/info/linking.html)

Preprocessing and clustering

In the following analysis, we used datasets as 12 time-point ($\{0a, 0b, 1a, 1b, 3a, 3b, \ldots.., 48a, 48b\} = \{T_i\}$ ($i = 1,2,\ldots,12$)) datasets on TOL and SHAM, since the *CODM* does not depend on the intervals of the time-points.

Standard clustering analysis for gene expression profiles is based on the correlation coefficients between genes. Therefore, this approach can not handle genes with expression profiles that have almost no changes for a condition. However, if the expression profiles of those genes have meaningful changes in expression levels for other conditions, they provide markers to interpret the influence that the conditions exerted, because they are possibly regulated by different factors. To handle those genes and to align the baselines of the expression patterns between the different datasets, preprocessing (i.e. filtering and normalization) was conducted for all of the datasets where TOL and SHAM were merged. More specifically, 3,363 probes with mean expressions above 50 and coefficient of variance (=standard deviation / mean) above 0.1 were selected. After logarithmic transformation of the gene expression data, the expression levels were normalized to satisfy the following equations:

$$\sum_{i}^{12}(x_i + y_i) = 0 \qquad \textbf{(1)}$$

$$\sum_{i}^{12}(x_i^2 + y_i^2) = 1 \qquad \textbf{(2)}$$

where x_i and y_i are normalized expression levels of a gene at time-point T_i ($i = 1,2,\ldots 12$) on conditions TOL and SHAM, respectively. Using these normalized datasets, hierarchical clustering analysis based on Euclidian distances was then performed for each dataset independently. Clustering analysis using Euclidian distances instead of correlation coefficients allows us to handle genes whose expression levels are down-regulated or up-regulated. In addition, due to the common normalization, gene expression patterns can be compared within a dataset and between datasets.

In general, Euclidian-distance based clustering after normalization, in terms of mean and standard deviation, is equivalent with correlation-coefficient based clustering. That is, a Euclidian-distance based clustering analysis for the merged data of TOL and SHAM with the above preprocessing is equivalent with a correlation-coefficient based clustering analysis for the original merged data. In the analysis of the *CODM*, the preprocessing is conducted for the merged data, and Euclidian-based clustering is individually conducted for each data. Roughly speaking, this analysis provides us with results similar to those of normal correlation-coefficient based clustering, while it allows us to handle genes with expression profiles that have changes for only one condition but not for the other.

As Figures 1a and 1b show, there are a large number of clusters generated at various levels. Although the composition and number of cluster sets depend on the threshold value of the distance, it is generally difficult to identify an optimum value. These aspects make it difficult to compare cluster sets derived from different sources.

The cluster overlap distribution map (*CODM*)

The *CODM* is a visualization methodology for pair-wise comparison between cluster sets generated from different gene expression datasets. In this methodology, two types of cluster sets (i.e. dendrograms of hierarchical clustering results) are mapped respectively to the X-axis and on the Y-axis, and the relationship between them is displayed as a 3D histogram (Figure 2). In this report, the dendrogram of TOL is mapped to the X-axis and that of SHAM is mapped to the Y-axis. The statistical evaluation values of the overlaps between two clusters selected from the respective cluster sets are displayed as the height of the blocks (Figure 2). More specifically, we evaluated the number of common genes between the two different clusters by using hypergeometric probability distributions (17). Assuming that the generation of gene clusters is a random selection from among the total set of genes, the probability of observing at least (k) overlapping genes between randomly selected (n_1) genes and (n_2) genes from among all of the (g) genes is given by:

$$P(g,n_1,n_2,k) = 1 - \sum_{i=k}^{k-1} \frac{{}_{n_2}C_i \cdot {}_{g-n_2}C_{n_1-i}}{{}_{g}C_{n_1}} \quad \left(= P(g,n_2,n_1,k)\right) \qquad (3)$$

When the P-value is small, the overlap is regarded as statistically meaningful. Thus, we defined the evaluation value of the overlap as:

$$E(g,n_1,n_2,k) = -\log_{10} P(g,n_1,n_2,k) \qquad (4)$$

Then in the area (R_{ij}) determined by a cluster on the X-axis (X_i) and a cluster on the Y-axis (Y_j), a block whose height represents $E(g, n_{xi}, n_{yj,} k_{ij})$ is displayed, where (n_{xi}) is the number of genes in (X_i), (n_{yj}) is the number of genes in (Y_j), and (k_{ij}) is the number of overlapping genes between (X_i) and (Y_j) (Figure 2). We term this block an *overlap block.* Note that the number of UniGenes, to which probes in a cluster correspond through their original GenBank accession number, was used as the number of genes. In this report, all 8737 probes on RG-U34A were corresponding to 5,249 UniGenes (g = 5,249).

For hierarchical clustering, there are a large number of clusters generated at various distance levels. Our algorithm examines the overlaps of the genes between all combinations of two clusters with smaller *distance level* values than the *cut level,* which is a threshold value specified by users (Figure 1). In other words, we evaluated and visualized any clusters with a smaller distance level than the *cut level,* even if they were included in other clusters. Note that conventional hierarchical clustering does not focus on sub-clusters that are included in other clusters. Since all of the statistically significant combinations between cluster sets can be visualized simultaneously, users can grasp the overall picture of the relationships between the two different cluster sets.

In the *CODM*, all of the clusters are dealt with equally without regard to their difference level (i.e. their homogeneity). Even if they are included in other clusters, all of the statistical significance of the number of common genes between clusters is simultaneously visualized. Therefore, there is a risk that a small *overlap block* may be hidden by a large block. For example, assume that the clusters X_j and Y_n are included in X_i and Y_m respectively. Then, if the evaluation value E_{jn} is less than E_{im}, the small block B_{jn} will be hidden in the large block B_{im} (Figure 3a). To avoid this problem, the *CODM* allows the user to change the *cut level* interactively. That is, if the user decreases the *cut level,* some small blocks that are hidden in larger blocks will emerge. Therefore, in consideration of the homogeneity of clusters and the relationships with other gene information, the user can find important genes displayed as blocks in the *CODM.*

Color of Each Overlap Block

Since the statistical significance of the number of common genes between two different clusters is represented as the height of a block, the color of a block can be used to represent other information. In the current prototype, the *CODM* provides three color modes.

(a) *Redundant Visualization*

The first is a representation of the evaluation values of overlaps using a gray scale. This redundant representation helps users comprehend the distribution of the relative evaluation values of overlaps.

(b) *Similarity of Expression Patterns*

The second is a representation of the similarity of expression patterns between two clusters, from red to blue. The similarity $f(T, S)$ of expression patterns between cluster T on TOL and cluster S on SHAM was defined using the average of the square of the Euclidean distance between them. Assuming that N_{TS} is the number of common genes in T and S, x_{ki} and y_{ki} are normalized expression levels of a common gene k at time T_i on TOL and SHAM, respectively. The similarity $f(T, S)$ was defined as follows.

$$f(T,S)=1-\frac{1}{N_{TS}}\sum_{k=1}^{N_{TS}}\sum_{i=1}^{12}(x_{ki}-y_{ki})^2 \tag{5}$$

Since $\{x_{ti}\}$ and $\{y_{si}\}$ ($i = 1,2,\ldots 12$) satisfy Equations 1 and 2, the range of $f(T, S)$ is –1 to 1 and $f(T, S)$ can be rewritten as follows (See Appendix).

$$f(T,S)=\frac{1}{N_{TS}}\sum_{k=1}^{N_{TS}}\sum_{i=1}^{12}2x_{ki}y_{ki} \tag{6}$$

In the *CODM*, the similarity $f(T, S)$ was represented as the color of the block from red ($f(T, S) = 1$) to blue ($f(T, S) = -1$). Roughly speaking, red indicates that expression patterns between the two clusters are similar and blue indicates they have a negative correlation. In addition, purple ($f(T, S) = 0$) indicates they have no correlation or genes of one cluster have no changes in expression levels (i.e. $\forall x_{ki} \approx 0$ or $\forall y_{ki} \approx 0$)

As mentioned above, if genes in a certain cluster based on SHAM also constitute a cluster in TOL, but the expression level in SHAM is significantly different from that in TOL, these genes provide potential markers for the cause of ischemic tolerance. Strong candidates will appear as tall blue or purple blocks. *CODM* allows users to easily look for such blocks, with interactively controlling the thresholds.

(c) *Relationship with a Known Gene Classification*

The third type of information is a representation of the relationship between overlapping genes and a known gene classification. If statistically significant representation of genes within a particular class is observed among the overlapping genes, the block is color-coded according to the class. The level of statistical significance of the representation of genes within a particular class is evaluated using Equation 3, where (g) is the total number of genes that are classified by the known classification, (n_1) is the number of genes which are classified by the known classification among overlapping genes, (n_2) is the total number of genes within a class based on the known gene classification, and (k) is the observed number of genes found in both the given overlapping genes and the given class according to the known gene classification.

In this report, we associated overlapping genes with 8 types of transcription factors (HIF, ARNT and EGR families), which were reported to have a relationship with ischemia (5,8,18,19). We extracted complete sequences of 1.0 kb upstream and 0.1 kb downstream for 2,816 UniGenes among the 5,249 UniGenes corresponding to 8,737 probes on RG-U34A. The 1.1 kb sequences of the 2,816 UniGenes were searched to determine if they correspond to the TRASFAC matrices v7.2 (11) with the threshold set to "Minimum False Negative". Table 1 shows the names of the transcription factors, the number of UniGenes that correspond to each transcription factor, and the thresholds for matching. In *CODM*, we color-coded *overlap blocks* which contain statistically meaningful number of genes with putative transcription factor binding sites. If an *overlap block* represents statistical significance for multiple transcription factors' putative binding sites, only a single transcription factor with the highest evaluation value was visualized. However, the *CODM* allows users to click *overlap blocks* and browse description messages (in a console window) for the relationships with all of the transcription factors.

3. Results and Discussion

Figures 4 shows the visualization results of the comparison between TOL and SHAM in the mode of redundant visualization, the similarity of the expression patterns, and the relationships with known gene classifications (transcription factors). In the figure, the *cut level* for the distance for hierarchical clustering was 0.74, and all *overlap blocks* with 2.0 or higher evaluation values are displayed as a 3D histogram. As the figure shows, the *CODM* provides not only a 3D mode but also a 2D mode where users can see a projected overhead view of the 3D mode. In the 3D mode, the statistical significance of the overlaps between clusters and the differences in expression levels between the clusters can be simultaneously represented, since we can use the height and color of blocks. However, it is a little difficult to recognize the expression patterns of clusters that generate an overlapping block. For this purpose, the 2D mode is better, although the 2D mode of *CODM* can visualize only a single species of information at a time, i.e. the statistical significance of the overlaps or the differences in expression levels between clusters, or relationships with known gene classification. Therefore, it is useful to interactively change the mode as required. Exploration by changing the color-mode and the 2D and 3D modes allowed us to pick up 3 potentially important *overlap blocks* (Figure 4). The information for these 3 *overlap blocks* is shown in Table 2, their gene lists are shown in the Supplement Tables, and their expression patterns are shown in Figure 5.

As stated above, we assumed that there are four issues for a comparison of clustering results: changes in the composition of the cluster sets, changes in the expression patterns, relationships with other known gene information, and threshold problems. The *CODM* enables us to address these issues as follows.

Changes in the composition of the cluster sets

As shown in Figures 4a and 4b, the *CODM* can intuitively visualize changes in the composition of the cluster sets as 3D histograms. That is, the dissimilarity of the expression level under SHAM divides each cluster on TOL into specific sub-clusters and these sub-clusters are displayed along the Y-axis. In the same manner, the relationships between each cluster of SHAM and all of the clusters of TOL are displayed on the X-axis. If a clustering analysis is conducted for the merged data of TOL and SHAM, these sub-clusters would be scattered and it would be difficult to intuitively observe the relationships of the compositions of the cluster sets.

Changes in the expression pattern

A comparison of the dynamic changes of gene expression level across time under various conditions provides a useful tool for interpreting complex biological processes. However, there are generally many false candidate genes whose expression patterns between two different conditions are different purely by chance. For the comparison between TOL and SHAM, only 357 probes (of the 3,363 selected probes) had 0.8 or higher correlation coefficient values of expression pattern between the two conditions. On the other hand, 756 probes had negative correlation coefficient values. As stated above, the difference of macroscopic phenomena that the conditions exhibit results from the difference of expression of not a single gene, but of multiple genes. Therefore, it is quite important to search for genes whose expression patterns changed in a similar fashion between different conditions. Figures 4c and 4d show that the *CODM* can simultaneously depict the statistical significance of the overlaps between clusters and the differences in their expression patterns. In this mode, tall blocks colored blue or purple, such as block B and C, would be good candidates, since their similarity of expression patterns were negative (-0.28 and -0.23), while the two clusters under different conditions share a statistically meaningful number of common genes (E = 53.3 and E = 34.8). Note that, the objective of the *CODM* is to identify such potentially important pairs of clusters from massive combinations. To further understand the significance of the expression patterns., it would be a desirable approach to combine *CODM* with other visualization tools for line graphical view of expression patters, as shown in Figure 5. The expression of genes in TOL in block B was up-regulated, compared to SHAM, at early-stage, i.e. 1h, 3h, and 12h. On the other hand, the expression of genes in TOL in block C was down-regulated, compared to SHAM, at early-stage, i.e. 1h, and 3h. Once again, *CODM* enabled us to easily detect candidate genes of this type.

Integration with other known gene information

In gene expression analysis, interpretation and validation of the results should be performed in the context of what is already known about the genes being analyzed. *CODM* allows us to associate the results with other such gene information and narrow down candidates. Figures 4e and 4f show the relationships between 8 types of transcription factors (HIF, ARNT and EGR families —see Table 1), which were reported to have a relationship with ischemia (5,8,18,19). In the figures, *overlap blocks* with 2.0 or higher evaluation values for the representation of genes with putative transcription factor binding sites were color-coded. Table 2 shows that *overlap blocks* A, B, and C implied relationship with the transcription factors (E>2.0). This example illustrates the utility of representing relationships with other known gene associated information by use of the color of *overlap blocks*, although it may be difficult to extract biological conclusions due to the limited number of genes with the putative binding sites in the *overlap blocks*. If binding-site information from more genes becomes available, more detailed analysis of results will be possible. Furthermore, representation of relationships with other known gene classifications should provide us with deeper insights.

Threshold problems

Arbitrary selection of thresholds involves a risk of overlooking important genes. In a comparison of cluster sets on gene expression profiles, there are four types of thresholds: 1) a threshold for generating clusters for each condition; 2) a threshold for evaluating the number of common genes that two clusters share; 3) a threshold for evaluating the differences in the expression patterns between two clusters; and

4) a threshold for evaluating the relationship with other known gene information. The *CODM* reduces the number of thresholds and allows users to interactively change the thresholds as follows.

1) Threshold for generating clusters for each condition

Since conventional hierarchical clustering does not focus on sub-clusters that are included in other clusters, there is a risk that the important sub-clusters could be overlooked. In the *CODM*, overlaps of genes between any two clusters of TOL and SHAM are statistically evaluated, even if they are included in other clusters. In addition, the *CODM* allows users to interactively change the *cut level*, in order to reduce the risk that a small *overlap block* may be hidden in a large block (Figure 6). Therefore, by considering the homogeneity of clusters and the relationships with other known gene information, the user should be able to find the important genes displayed as blocks.

2) Threshold for evaluating the number of common genes shared by two clusters.

In *CODM*, the statistical significance of the number of common genes between two different clusters is represented as the height of a block, and statistical significance of the overlap of all combinations of clusters are displayed as a 3D histogram at the same time. Therefore, without the selection of an arbitrary threshold, the distribution of the statistical significance of the overlap is effectively displayed. Although (to reduce the rendering load) Figure 4 shows only *overlap blocks* with 2.0 or higher evaluation values of the overlap, users can interactively change this value.

3) Threshold for evaluating the differences in the expression patterns between two clusters

CODM represents the differences in the expression patterns between two clusters by the color of the blocks ranging from red to blue. Therefore, the distribution of differences in the expression patterns of all combinations of clusters is displayed at the same time, without any selection of an arbitrary threshold.

4) Threshold for evaluating the relationships with other known gene information.

Although only *overlap blocks* with 2.0 or higher evaluation values for the representation of genes with putative transcription factor binding sites were color-coded in Figures 4e and 4f, users can interactively change this value.

4. Conclusion

In this report we described the characteristics of the *Cluster Overlap Distribution Map* (*CODM*) method, a visualization tool for comparing clustering results of gene expression profiles under two different conditions. In *CODM*, the utilization of three-dimensional space and color allows us to intuitively visualize changes in the composition of cluster sets, changes in the expression patterns of genes between the two conditions, and the relationships with a known gene classification such as transcription factors. Comparison of dynamic changes of gene expression levels across time under different conditions is required in a wide variety of fields of gene expression analysis, including toxicogenomics and pharmacogenomics. Since *CODM* integrates and simultaneously visualizes various types of information across clustering results, it can be applied to various analyses in these fields.

References

1. Alizadeh AA, and Staudt LM. Genomic-scale gene expression profiling of normal and malignant immune cells. *Curr Opin Immunol* 12: 219-225, 2000.

2. Chiang LW, Grenier JM, Ettwiller L, Jenkins LP, Ficenec D, Martin J, Jin F, DiStefano PS, and Wood A. An orchestrated gene expression component of neuronal programmed cell death revealed by cDNA array analysis. *Proc Natl Acad Sci USA* 98: 2814-2819, 2001.

3. Cho RJ, Huang M, Campbell MJ, Dong H, Steinmetz L, Sapinoso L, Hampton G, Elledge SJ, Davis RW, and Lockhart DJ. Transcriptional regulation and function during the human cell cycle. *Nat Genet* 27: 48-54, 2001.

4. Eisen MB, Spellman PT, Brown PO, and Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* 95: 14863-14868, 1998.

5. Huang LE, Arany Z, Livingston DM, and Bunn HF. Activation of hypoxia-inducible transcription factor depends primarily upon redox-sensitive stabilization of its alpha subunit. *J Biol Chem* 271: 32253-32259, 1996.

6. Ishii M, Hashimoto S, Tsutsumi S, Wada Y, Matsushima K, Kodama T, and Aburatani H. Direct comparison of GeneChip and SAGE on the quantitative accuracy in transcript profiling analysis. *Genomics* 68: 136-143, 2000.

7. Kano M, Nishimura K, Tsutsumi S, Aburatani H, Hirota K, and Hirose M. Cluster overlap distribution map: visualization for gene expression analysis using immersive projection yechnology. *Presence: Teleoperators and Virtual Environments* 12: 96-109, 2003.

8. Kawahara N, Wang Y, Mukasa A, Furuya K, Shimizu T, Hamakubo T, Aburatani H, Kodama T, and Kirino T. Genome-wide gene expression analysis for induced ischemic tolerance and delayed neuronal death following transient global ischemia in rats. *J Cereb Blood Flow Metab* 24: 212-223, 2004.

9. Kirino T. Ischemic tolerance. *J Cereb Blood Flow Metab* 22: 1283-96, 2002.

10. Manger ID, and Relman DA. How the host 'sees' pathogens: global gene expression responses to infection. *Curr Opin Immunol* 12: 215-218, 2000.

11. Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos DU, Land S, Lewicki-Potapov B, Michael H, Munch R, Reuter I, Rotert S, Saxel H, Scheer M, Thiele S, and Wingender E. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res* 31: 374-378, 2003.

12. Rhodes DR, Barrette TR, Rubin MA, Ghosh D, and Chinnaiyan AM. Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Res* 62: 4427-4433, 2002.

13. Saban MR, Hellmich H, Nguyen NB, Winston J, Hammond TG, and Saban R. Time course of LPS-induced gene expression in a mouse procmodel of genitourinary inflammation. *Physiol Genom* 5: 147-160, 2001.

14. Seo J, and Shneiderman B. Interactively Exploring Hierarchical Clustering Results. *IEEE Computer* 35: 80-86, 2002.

15. Shiffman D, Mikita T, Tai JT, Wade DP, Porter JG, Seilhamer JJ, Somogyi R, Liang S, and Lawn RM. Large scale gene expression analysis of cholesterol-loaded macrophages. *J Biol Chem* 275: 37324-37332, 2000.

16. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, and Golub TR. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci USA* 96: 2907-2912, 1999.

17. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, and Church GM. Systematic determination of genetic network architecture. *Nat Genet* 22: 281-285, 1999.

18. Wang GL, Jiang BH, Rue EA, and Semenza GL. Hypoxia-inducible factor 1 is a basic-helix-loop-helix-PAS heterodimer regulated by cellular O2 tension. *Proc Natl Acad Sci USA* 92: 5510-5514, 1995.