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 AkitaK Mukasa,2,3 Takaaki Kirino,3,4 and Hiroyuki Aburatani2

1Intelligent Cooperative System, Department of Information Systems, Research Center for Advanced Science and Technology, University of Tokyo, Tokyo; 2Genome Science Division, Research Center for Advanced Science and Technology and 3Department of Neurosurgery, Faculty of Medicine, University of Tokyo, Tokyo; and
4Solution-Oriented Research for Science and Technology/Japan Science and Technology, Kawaguchi, Saitama, Japan

Submitted 5 May 2004; accepted in final form 11 February 2005

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 different conditions is 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, because it integrates and simultaneously visualizes various types of information across clustering results.

time series; transcription factor; visualization

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) provide 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 data sets was proposed (12). This model revealed, for example, that four prostate cancer gene expression data sets 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 data sets share significantly similar results, but to detect the differences be-
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, then 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 three-dimensional (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 data sets 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 ~4,000 genes, represented in this paper, required ~250 megabytes. In addition, since the analysis results of the CODM are visualized by use of the OpenGL, a machine with a graphics board with a hardware accelerator for the OpenGL is recommended.

MATERIALS AND METHODS

Experiment Design

In this report, CODM is illustrated using time series gene expression data sets obtained from rat four-vessel occlusion models combined with systemic hypotension and time-matched control animals with sham operation. In the experiment, we used 2-min 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-min 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). With duplicate assessments of 6 time points (0 h, 1 h, 3 h, 12 h, 24 h, 48 h) × 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 were 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).

GeneChip Experiment

Five micrograms of total RNA from each sample was 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 8,737 rat genes and expressed sequence tags (ESTs), which were selected from Build 34 of the UniGene Database (derived from 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
In the following analysis, we used data sets as 12 time point \(\{0a, 0b, 1a, 1b, 3a, 3b, \ldots, 48a, 48b\}\) microarray data from rats with induced ischemic tolerance (tolerant rats, TOL) and rats with sham operation (sham rats, SHAM). In the analysis, we used these data sets as 12 time point \(\{0a, 0b, 1a, 1b, 3a, 3b, \ldots, 48a, 48b\}\) data sets on TOL and SHAM, respectively. After preprocessing and normalization, hierarchical clustering analysis based on Euclidian distances was then performed for each data set independently.

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 data set has been submitted as GSE1357 to the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/info/linking.html)

**Preprocessing and Clustering**

In the following analysis, we used data sets as 12 time point \(\{0a, 0b, 1a, 1b, 3a, 3b, \ldots, 48a, 48b\}\) data sets 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 cannot 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, then these provide markers to interpret the influence that the conditions exerted, because these are possibly regulated by different factors. To handle those genes and to align the baselines of the expression patterns between the different data sets, preprocessing (i.e., filtering and normalization) was conducted for all of the data sets where TOL and SHAM were merged. More specifically, 3,363 probes with mean expressions above 50 and coefficient of variance (CV = 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:

1. \[\sum_{i=1}^{12} (x_i + y_i) = 0\]  
2. \[\sum_{i=1}^{12} (x_i^2 + y_i^2) = 1\]

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 data sets, we performed hierarchical clustering analysis based on Euclidian distances, for each data set independently. Clustering analysis using Euclidian distances instead of correlation coefficients allows us to handle genes whose expression levels are downregulated or upregulated. In addition, due to the common normalization, gene expression patterns can be compared within a data set and between data sets.

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 Fig. 1, A and B, shows, 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**

The CODM is a visualization methodology for pair-wise comparison between cluster sets generated from different gene expression data sets. In this methodology, two types of cluster sets (i.e., dendrograms of hierarchical clustering results) are mapped, respectively, to the x-axis and to the y-axis, and the relationship between them is displayed as a 3D histogram (Fig. 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 (Fig. 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:

\[P(g, n_1, n_2; k) = 1 - \sum_{i=0}^{k-1} \frac{n_1^C_i \cdot (g-n_1)^C_{n_2-i}}{g^C_{n_2}} \]  

(3)

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

\[E(g, n_1, n_2, k) = \frac{X_i (n_{2i})}{X_i (n_{2i}) + Y_i (n_{2j})} \]

where \(X_i\) and \(Y_i\) are the number of overlap genes between cluster sets on TOL and SHAM, respectively.

**Fig. 2.** Overlap block of two clusters. The dendrogram of TOL is mapped to the x-axis, and that of SHAM is mapped to the y-axis. Then, for the area \((R_g)\) determined by a cluster on the x-axis \((X)\) and a cluster on the y-axis \((Y)\), a block whose height represents \(E(g, n_1, n_2, k)\) (statistical evaluation values of the overlaps between \(X_i\) and \(Y_i\)) is displayed, where \(g\) is the total number of genes, \(n_{2i}\) is the number of genes in \(X_i\), \(n_{2j}\) is the number of genes in \(Y_i\), and \(k\) is the number of overlap genes between \(X_i\) and \(Y_i\).
Then in the area \( R_{xy} \) determined by a cluster on the \( x \)-axis \( (X_i) \) and a cluster on the \( y \)-axis \( (Y_j) \), a block whose height \( h_{xy} \) is displayed, where \( n_{XY} \) 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 \( Y_j \) (Fig. 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 8,737 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 (Fig. 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 subclusters 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 \( X_1 \) and \( Y_m \) are included in \( X_i \) and \( Y_n \), respectively. Then, if the evaluation value \( E_{im} \) is less than \( E_{in} \), then the small block \( B_{im} \) will be hidden in the large block \( B_{in} \) (Fig. 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.

1) **Redundant visualization.** The first mode is a representation of the evaluation values of overlap blocks using a gray scale. This redundant representation helps users comprehend the distribution of the relative evaluation values of overlaps.

2) **Similarity of expression patterns.** The second mode 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 \( N_T \) 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:

\[
f(T,S) = 1 - \frac{1}{N_T} \sum_{k=1}^{N_T} \sum_{i=1}^{12} (x_{ki} - y_{ki})^2
\]

Since \( x_{ki} \) and \( y_{ki} \) \( i = 1, 2, \ldots, 12 \) satisfy Eqs. 1 and 2, the range of \( f(T,S) \) is \(-1 \) to \( 1 \), and \( f(T,S) \) can be rewritten as follows (See APPENDIX):

\[
f(T,S) = \frac{1}{N_T} \sum_{k=1}^{N_T} \sum_{i=1}^{12} 2x_{ki}y_{ki}
\]

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.,

\[
\forall x_{ki} = 0 \text{ or } \forall y_{ki} = 0
\]

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, then 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.

3) **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, then 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 Eq. 3, where \( g \) is the total number of genes that are classified by the known classification, \( n_{1j} \) is the number of genes that are classified by the known classification among overlapping genes, \( n_{2j} \) 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 eight types of transcription factors (HIF, ARNT, and EGR families) that 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 microarray. The 1.1-kb sequences of 1,084 transcription factors (HIF, ARNT, and EGR families) that 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 microarray. The 1.1-kb sequences of 1,084 transcription factors (HIF, ARNT, and EGR families) that 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 microarray. The 1.1-kb sequences of 1,084 transcription factors (HIF, ARNT, and EGR families) that 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 microarray.
Visualization for Time Series Gene Expression Analysis

Table 1. Transcription factors linked to ischemia

<table>
<thead>
<tr>
<th>Transcription Factor</th>
<th>No. of UniGenes</th>
<th>Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>VSAHRRARNT_01</td>
<td>540</td>
<td>0.92</td>
</tr>
<tr>
<td>VSAHRRARNT_02</td>
<td>4</td>
<td>0.91</td>
</tr>
<tr>
<td>VSHIF1_Q3</td>
<td>955</td>
<td>0.55</td>
</tr>
<tr>
<td>VSHIF1_Q5</td>
<td>507</td>
<td>0.87</td>
</tr>
<tr>
<td>VSEGRI_01</td>
<td>143</td>
<td>0.87</td>
</tr>
<tr>
<td>VSEGRI_02</td>
<td>92</td>
<td>0.89</td>
</tr>
<tr>
<td>VSEGRI_03</td>
<td>26</td>
<td>0.93</td>
</tr>
<tr>
<td>VSEGRI_04</td>
<td>143</td>
<td>0.88</td>
</tr>
</tbody>
</table>

In the cluster overlap distribution map (CODM), changes in the composition of the cluster sets and changes in the expression patterns between different conditions were associated with 8 types of transcription factors (HIF, ARNT, and EGR families), which are all known to mediate response to ischemia. We extracted UniGenes that contain putative binding sites for the transcription factors and correspond to probes on RG-U34A GeneChips (Affymetrix, Santa Clara, CA). Shown are the names of the transcription factors, the number of UniGenes, and the thresholds for matching.

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 Fig. 4, A and B, 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 subclusters, and these subclusters 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, then these subclusters 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 differences. 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. Figure 4, C and D, shows 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 blocks B and C, would be good candidates, since their similarities 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 patterns, as shown in Fig. 5. The expression of genes in TOL in block B was upregulated, compared with SHAM, at early stage, i.e., 1 h, 3 h, and 12 h. On the other hand, the expression of genes in TOL in block C was downregulated, compared with SHAM, at early stage, i.e., 1 h, and 3 h. 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. Figure 4, E and F, shows the relationships between eight types of transcription factors (HIF, ARNT, and EGR families; see Table 1) that were reported to have a relationship with ischemia (5, 8, 18, 19). In Fig. 4, 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 a relationship with the transcription factors $E > 2.0$. This example illustrates the utility of representing relationships with other known gene-associated information by

Fig. 4. Visualizations for comparison of clustering results of TOL and SHAM. These are visualizations results of the comparisons between TOL and SHAM in the mode of redundant visualization (A and B), similarity of the expression patterns (C and D), and the relationships with transcription factors (E and F). Here, the cut level of the distance for hierarchical clustering was 0.74, and all of the overlap blocks with 2.0 or higher evaluation values are displayed as three-dimensional (3D) histograms. As shown, the CODM provides not only a 3D mode (B, D, and F) but also a two-dimensional (2D) mode (A, C, and E) where users can see a projected overhead view of the 3D mode. In the mode showing the relationships with the transcription factors (E and F), we considered the relationships with 8 types of transcription factors (HIF, ARNT, and EGR families) that are known to mediate response to ischemia. Here, only overlap blocks with 2.0 or higher evaluation values of the number of genes with putative transcription factor binding sites were color coded. Where an overlap block represents statistical significance for multiple transcription factors’ putative binding sites, only the transcription factor with the highest evaluation value was visualized. Exploration through changing the color mode and the 2D and 3D mode allowed us to pick up three potentially important overlap blocks that represented high evaluation values of the number of genes with the binding sites ($E > 2.0$).
Table 2. Information about 3 overlap blocks

<table>
<thead>
<tr>
<th>Overlap Block</th>
<th>No. of UniGenes in Cluster of TOL</th>
<th>No. of UniGenes in Cluster of SHAM</th>
<th>No. of Common UniGenes (Evaluation Value)</th>
<th>Similarity $f(T,S)$</th>
<th>Binding Sites of Transcription Factors: No. of Genes (Evaluation Value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>156</td>
<td>147</td>
<td>54 ($E = 46.9$)</td>
<td>0.42</td>
<td>VSAHRARNT_01:14 ($E = 2.10$)</td>
</tr>
<tr>
<td>B</td>
<td>190</td>
<td>132</td>
<td>60 ($E = 53.3$)</td>
<td>−0.28</td>
<td>VSEG1_01:6 ($E = 2.01$)</td>
</tr>
<tr>
<td>C</td>
<td>99</td>
<td>207</td>
<td>43 ($E = 34.8$)</td>
<td>−0.23</td>
<td>VSHIF1_Q3:11 ($E = 2.33$)</td>
</tr>
</tbody>
</table>

Exploration with CODM allowed us to pick up 3 potentially important “overlap blocks.” The “No. of UniGenes in Cluster of TOL(SHAM)” is the number of UniGenes which correspond to probes included in a cluster of TOL(SHAM). The “No. of Common UniGenes” is the number of common genes shared between the clusters of TOL and SHAM, and its statistical evaluation value, ($E$), is shown in parentheses. The “Similarity $f(T,S)$” is the similarity of the expression patterns between the clusters of TOL and SHAM. The range of similarity $f(T,S)$ is $-1$ (dissimilar) to $1$ (similar). The “Binding Sites of Transcription Factors” shows the name of putative binding sites of transcription factors, the number of common genes that share the same binding sites, and the $E$ value of the number of common genes with the same binding sites, if the evaluation value is 2.0 or higher. TOL, induced ischemic tolerance; SHAM, sham operation.

![Fig. 5. Expression patterns of genes in the three overlap blocks. These are the expression patterns of common genes for the three overlap blocks that were picked up through exploration with CODM (Fig. 4). The “Expression Patterns of Cluster $T_i(S_i)$” ($i = a,b,c$) are the expression patterns of the common genes of the overlap block $i$ in TOL(SHAM).](http://physiolgenomics.physiology.org/)

Physiol Genomics • VOL 21 • www.physiolgenomics.org
use of the color of overlap blocks, although it may be difficult to extract biological conclusions because of the limited number of genes with the putative binding sites in the overlap blocks. If binding site information from more genes becomes available, then 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 subclusters that are included in other clusters, there is a risk that the important subclusters could be overlooked. In the CODM, overlaps of genes between any two clusters of TOL and SHAM are statistically evaluated, even if these are included in other clusters. In addition, the CODM allows users to interactively change the cut level, to reduce the risk that a small overlap block may be hidden in a large block (Fig. 6). Therefore, by considering the homogeneity of clusters and the relationships with other known gene information, the user can find important genes displayed as blocks in the CODM.

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 significances 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) Fig. 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 Fig. 4E and Fig. 4F, users can interactively change this value.

Conclusion

In this report we described the characteristics of the CODM method, a visualization tool for comparing clustering results of gene expression profiles under two different conditions. In CODM, the utilization of 3D 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.

APPENDIX

Similarity $f(T,S)$

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

$$= 1 - \frac{1}{N_{TS}} \sum_{k=1}^{N_{TS}} \left( \sum_{i=1}^{12} (x_{ki}^2 + y_{ki}^2) - \sum_{i=1}^{12} 2x_{ki}y_{ki} \right)$$

$$= 1 - \frac{1}{N_{TS}} \sum_{k=1}^{N_{TS}} \left( 1 - \sum_{i=1}^{12} 2x_{ki}y_{ki} \right)$$

$$= \frac{1}{N_{TS}} \sum_{k=1}^{N_{TS}} \sum_{i=1}^{12} 2x_{ki}y_{ki}$$

The similarity $f(T,S)$ satisfies the following inequality:

$$-1 \leq f(T,S) \leq 1$$

Proof. Since $f(T,S) \leq 1$ is obvious, we only need to prove $-1 \leq f(T,S)$. We begin by showing that

$$g = \sum_{i=1}^{12} 2x_{i}y_{i} \geq -1$$

where

$$\sum_{i=1}^{12} (x_{i}^2 + y_{i}^2) = 1$$

We consider the Lagrangian function

$$L = \sum_{i=1}^{12} 2x_{i}y_{i} + \gamma \left( \sum_{i=1}^{12} (x_{i}^2 + y_{i}^2) - 1 \right)$$

where $\gamma$ is a Lagrange undetermined multiplier. By taking the derivative, we convert the constrained optimization problem into an unconstrained problem as follows:

$$\frac{\partial L}{\partial x_i} = 2y_i + 2\gamma x_i = 0 \quad (i = 1 \ldots 12)$$

$$\frac{\partial L}{\partial y_i} = 2x_i + 2\gamma y_i = 0 \quad (i = 1 \ldots 12)$$

$$\frac{\partial L}{\partial \gamma} = \sum_{i=1}^{12} (x_i^2 + y_i^2) - 1 = 0$$

The solutions of this problem are

$$x_i = y_i \quad (i = 1, 2, \ldots, 12), \quad \gamma = -1 \rightarrow g$$

or

$$x_i = -y_i \quad (i = 1, 2, \ldots, 12), \quad \gamma = 1 \rightarrow g$$

Therefore,

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

REFERENCES


