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.

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

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).

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

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.

19. Yan SF, Lu J, Zou YS, Soh-Won J, Cohen DM, Buttrick PM, Cooper DR, Steinberg SF, Mackman N, Pinsky DJ, and Stern DM. Hypoxia-associated induction of early growth response-1 gene expression. *J Biol Chem* 274: 15030-15040, 1999.

Appendix **Similarity *f(T,S)***

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

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

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

Proof.

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

$$g = \sum_{i=1}^{12} 2x_i y_i \ge -1$$

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

We consider the Lagrangian function

$$L = \sum_{i=1}^{12} 2x_i y_i + \lambda \{ \sum_{i}^{12} (x_i^{\ 2} + y_i^{\ 2}) - 1 \}$$

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

$$\frac{\partial L}{\partial x_i} = 2y_i + 2\lambda \, x_i = 0 \qquad (i = 1....12)$$

$$\frac{\partial L}{\partial y_i} = 2x_i + 2\lambda \, y_i = 0 \qquad (i = 1....12)$$

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

The solutions of this problem are

(i) $x_i = y_i$ $(i = 1,2,\ldots,12)$, $\lambda = -1$ ===> g has the maximum value 1

or

(ii) $x_i = -y_i$ $(i = 1,2,\ldots,12)$, $\lambda = 1$ ===> g has the minimum value -1

Therefore,

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

Table 1. Transcription factors linked to ischemia

transcription factor	# of UniGenes	thresholds
V$AHRARNT_01	540	0.92
V$AHRARNT_02	4	0.91
V$HIF1_Q3	955	0.55
V$HIF1_Q5	507	0.87
V$EGR1_01	143	0.87
V$EGR2_01	92	0.89
V$EGR3_01	26	0.93
V$NGFIC_01	143	0.88

In *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 which contain putative binding sites for the transcription factors, and correspond to probes on RG-U34A (Affymetrix, Santa Clara, CA). This table shows the names of the transcription factors, the number of UniGenes and the thresholds for matching.

Table 2. Information about 3 overlap blocks

Overlap block	# of UniGenes in cluster of TOL	# of UniGenes in cluster of SHAM	# of common UniGenes (evaluation value)	similarity *f*(T,S)	Binding-sites of transcription factors : # of genes (evaluation value)
A	156	147	54 ($E = 46.9$)	0.42	V$AHRARNT_01 : 14 ($E = 2.10$)
B	190	132	60 ($E = 53.3$)	-0.28	V$EGR1_01 : 6 ($E = 2.01$)
C	99	207	43 ($E = 34.8$)	-0.23	V$HIF1_Q3 : 11 ($E = 2.33$)

Exploration with *CODM* allowed us to pick up 3 potentially important *overlap blocks*. This table shows the information for these 3 *overlap blocks*. The "# of UniGenes in cluster of TOL(/SHAM)" is the number of UniGenes which correspond to probes included in a cluster of TOL(/SHAM). The "# of common UniGenes (evaluation value)" is the number of common genes shared between the clusters of TOL and SHAM and its statistical evaluation value. 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 statistical evaluation value of the number of common genes with the same binding-sites, if the evaluation value is 2.0 or higher.

Figures and Figure Legends

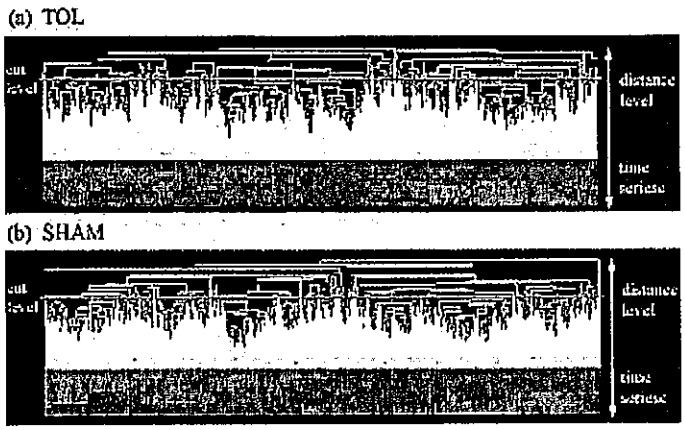

Figure 1. Hierarchical clustering of TOL and SHAM

We obtained time series ({0h, 1h, 3h, 12h, 24h, 48h} x 2) 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 datasets as 12 time-points ({0a, 0b, 1a, 1b, 3a, 3b, ..., 48a, 48b} = $\{T_i\}$ (i = 1,2,...,12)) datasets on TOL and SHAM, respectively. After preprocessing and normalization, hierarchical clustering analysis based on Euclidian distances was then performed for each dataset independently.

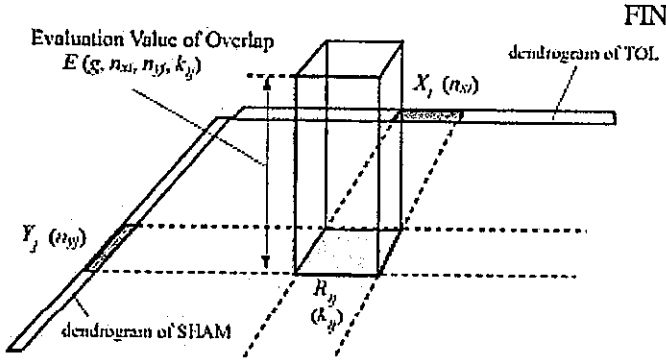

Figure 2. Overlap Block of Two Clusters

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_{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 $E(g, n_{xi}, n_{yj}, k_{ij})$ (statistical evaluation values of the overlaps between X_i and a Y_j,) is displayed, where (g) is the total number of genes, (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 overlap genes between (X_i) and (Y_j).

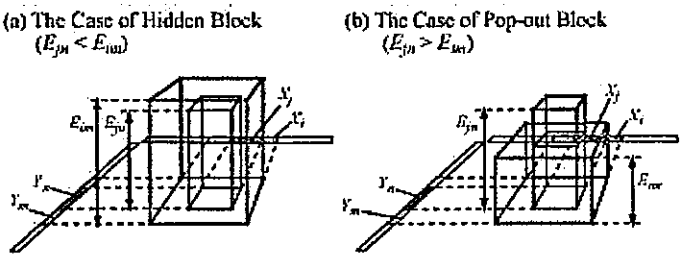

Figure 3. Relationships of Two Blocks

In *CODM*, all of the clusters are dealt with equally, regardless of their difference levels (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. Figure 3 shows that there is a risk that a small *overlap blocks* may be hidden in a large block. 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 within the large block B_{im} (Figure 3a).

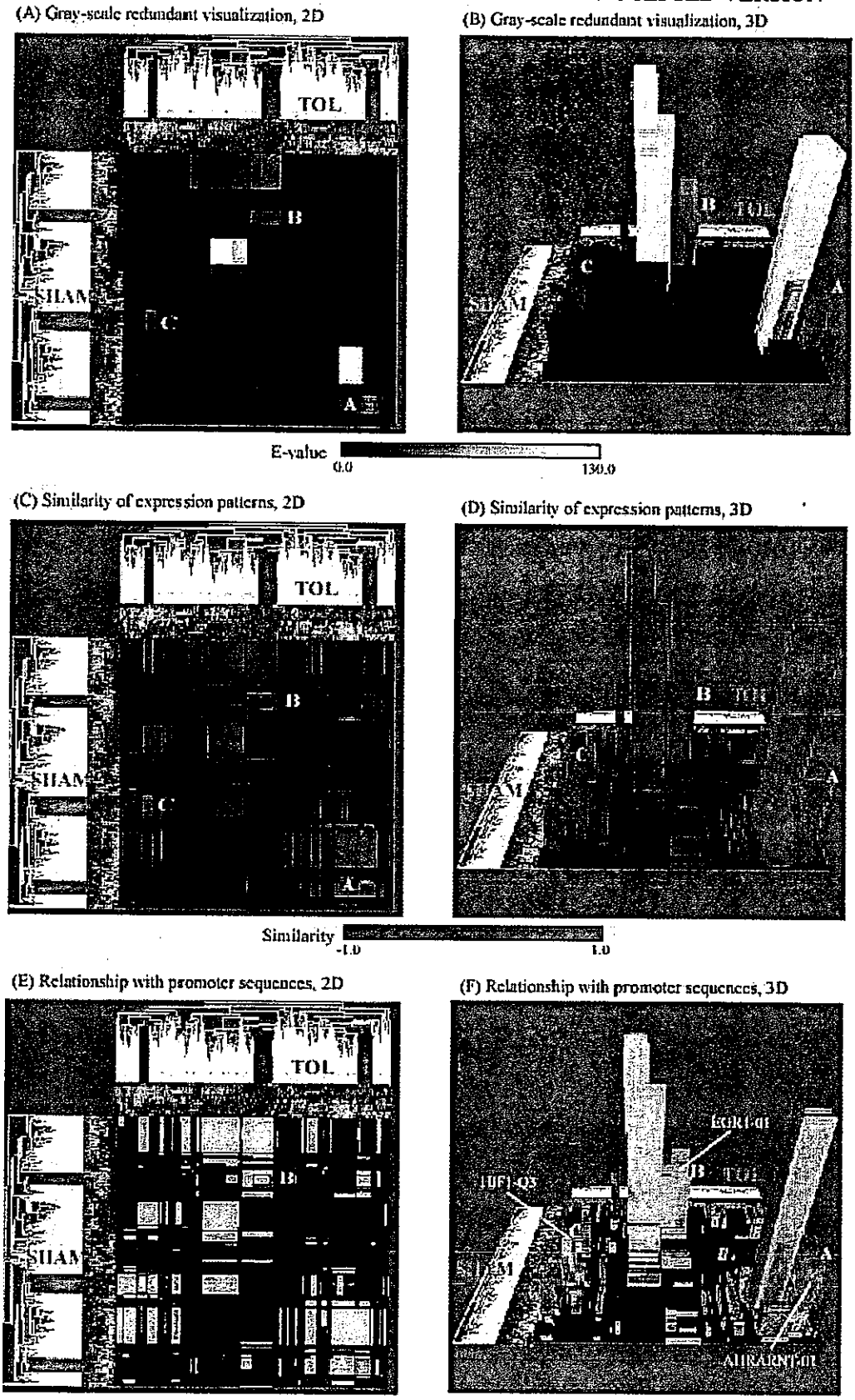

Figure 4. Visualizations for Comparison of Clustering Results of TOL and SHAM

This figure shows visualization results of the comparisons between TOL and SHAM in the mode of redundant visualization (Figures 4a and 4b), similarity of the expression patterns (Figures 4c and 4d), and the relationships with transcription factors (Figures 4e and 4f). In these figures, 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 3D histograms. As the figures show, the *CODM* provides not only a 3D mode (Figures 4b, 4d, and 4f) but also a 2D mode (Figures 4a, 4c, and 4e) where users can see a projected overhead view of the 3D mode.

In the mode showing the relationships with the transcription factors (Figures 4e and 4f), we considered the relationships with 8 types of transcription factors (HIF, ARNT and EGR families), which are known to mediate response to ischemia. In these figures, 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&3D mode allowed us to pick up 3 potentially important *overlap blocks* which represented high evaluation values of the number of genes with the binding-sites ($E > 2.0$).

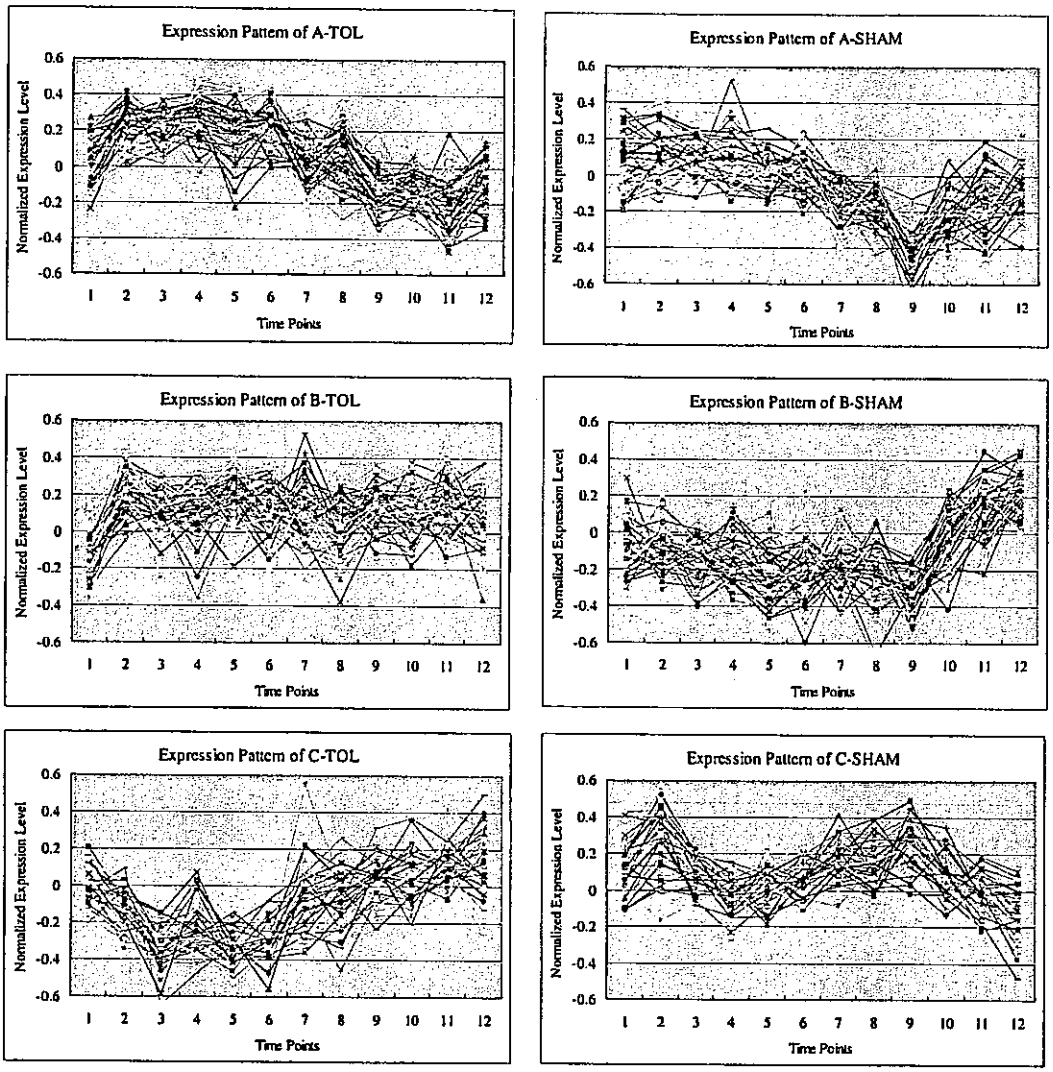

Figure 5. Expression Patterns of genes in the 3 overlap blocks

These figures show the expression patterns of common genes for the 3 *overlap blocks* which were picked up through exploration with *CODM* (Figure 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).