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

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

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

We consider the Lagrangian function

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

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

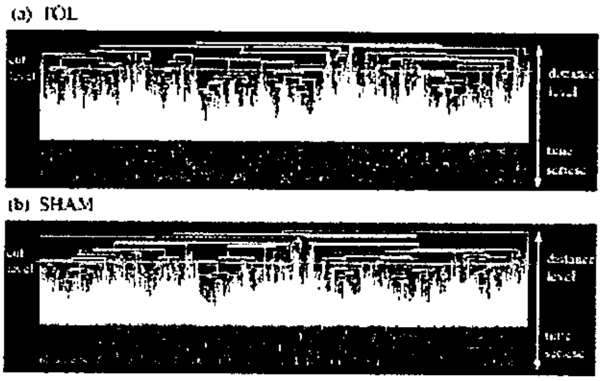


Figure 1. Hierarchical clustering of TOL and SHAM

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.

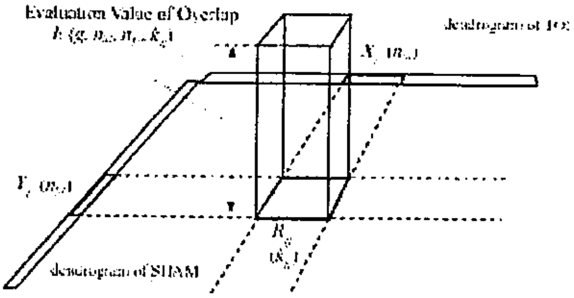
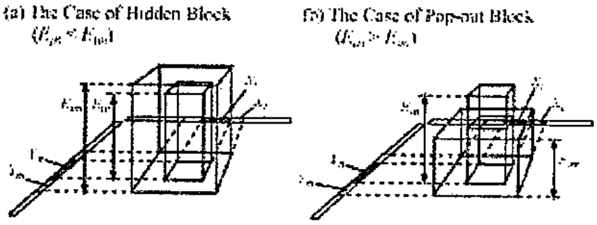


Figure 2. Overlap Block of Two Clusters

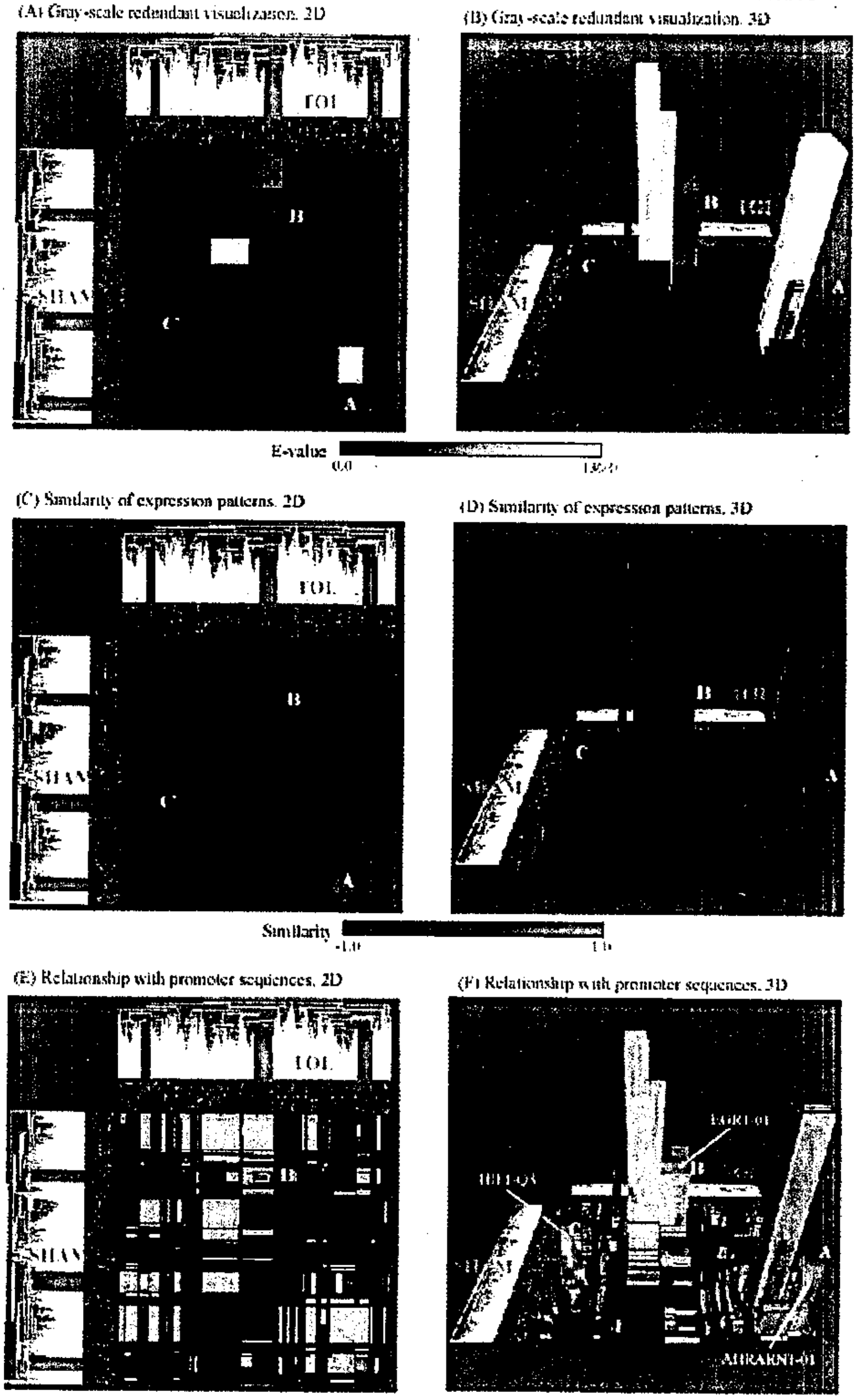
The dendrogram of TOL is mapped to the X-axis and that of SHAM is mapped to the Y-axis. Then, for the area (R_{ij}) determined by a cluster on the X-axis (X_i) and a cluster on the Y-axis (Y_j), a block whose height represents $E(g, n_{xi}, n_{yj,} k_{ij})$ (statistical evaluation values of the overlaps 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).


(a) The Case of Hidden Block ($E_{jn} < E_{im}$)

(b) The Case of Pop-out Block ($E_{jn} > E_{im}$)

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).



(A) Gray-scale redundant visualization, 2D

(B) Gray-scale redundant visualization, 3D

(C) Similarity of expression patterns, 2D

(D) Similarity of expression patterns, 3D

(E) Relationship with promoter sequences, 2D

(F) Relationship with promoter sequences, 3D

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 ($E > 2.0$).

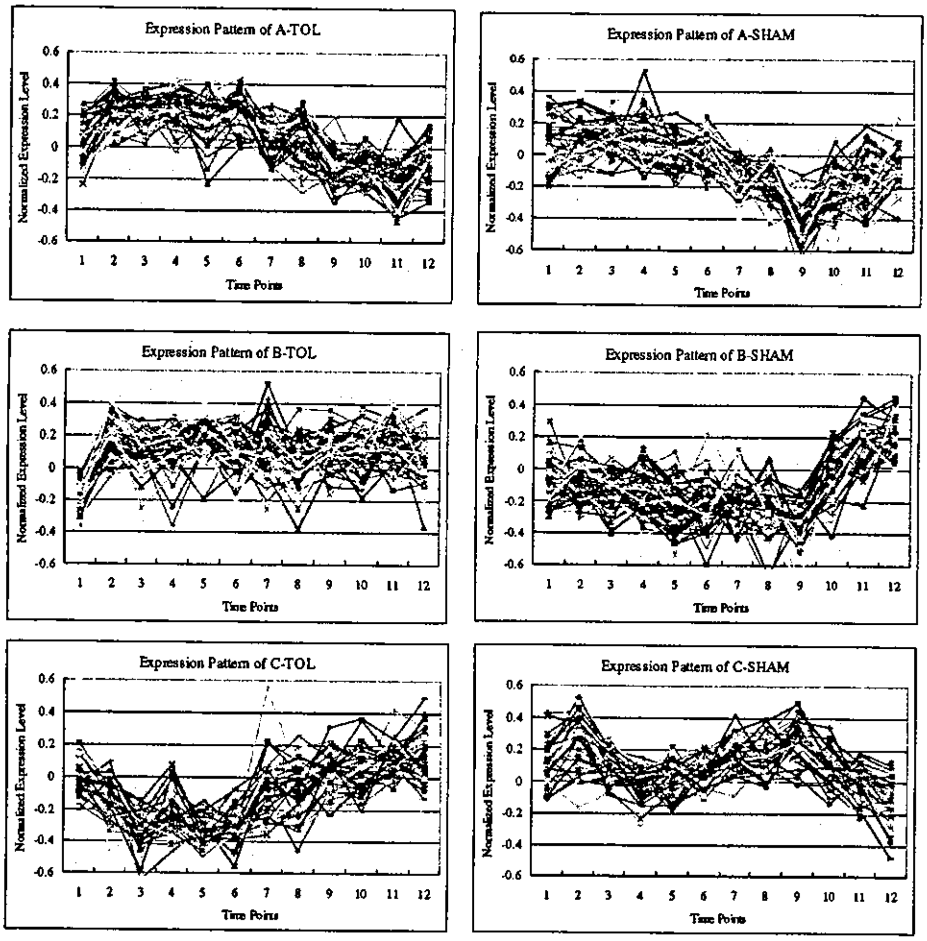


Figure 5. Expression Patterns of genes in the 3 overlap blocks

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).

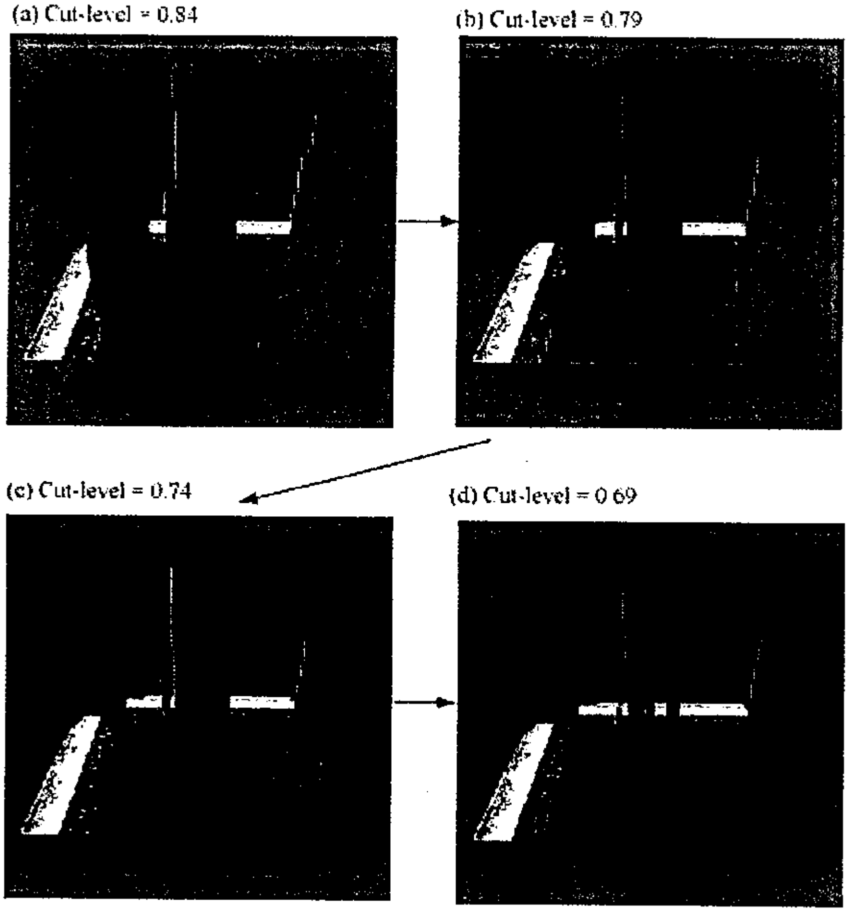


Figure 6. Interactive Changes of Cut-levels

In *CODM*, there is a risk that a small *overlap block* may be hidden in a large block. To avoid this problem, *CODM* allows the user to change the *cut level* interactively. If the user decreases the *cut level*, some small blocks that are hidden in larger blocks will emerge. By considering the homogeneity of clusters and the relationships with other gene information, the user can find important genes displayed as blocks in the *CODM*.

ORIGINAL PAPER

Vol. 21 no. 4 2005, pages 439–444
doi:10.1093/bioinformatics/bti188

Multidimensional support vector machines for visualization of gene expression data

D. Komura [1,*], H. Nakamura[1], S. Tsutsumi[1], H. Aburatani[2] and S. Ihara[1]

[1]Research Center for Advanced Science and Technology and [2]Genome Science Division, Center for Collaborative Research, University of Tokyo, Tokyo 153-8904, Japan

Received on May 21, 2004; accepted on November 11, 2004
Advance Access publication December 17, 2004

ABSTRACT

Motivation: Since DNA microarray experiments provide us with huge amount of gene expression data, they should be analyzed with statistical methods to extract the meanings of experimental results. Some dimensionality reduction methods such as Principal Component Analysis (PCA) are used to roughly visualize the distribution of high dimensional gene expression data. However, in the case of binary classification of gene expression data, PCA does not utilize class information when choosing axes. Thus clearly separable data in the original space may not be so in the reduced space used in PCA.

Results: For visualization and class prediction of gene expression data, we have developed a new SVM-based method called multidimensional SVMs, that generate multiple orthogonal axes. This method projects high dimensional data into lower dimensional space to exhibit properties of the data clearly and to visualize a distribution of the data roughly. Furthermore, the multiple axes can be used for class prediction. The basic properties of conventional SVMs are retained in our method: solutions of mathematical programming are sparse, and nonlinear classification is implemented implicitly through the use of kernel functions. The application of our method to the experimentally obtained gene expression datasets for patients' samples indicates that our algorithm is efficient and useful for visualization and class prediction.

Contact: komura@hal.rcast.u-tokyo.ac.jp

1 INTRODUCTION

DNA microarray has been the key technology in modern biology and helped us to decipher the biological system because of its ability to monitor the expression levels of thousands of genes simultaneously. Since DNA microarray experiments provide us with huge amount of gene expression data, they should be analyzed with statistical methods to extract the meanings of experimental results.

A great number of supervised learning algorithms have been proposed and applied to classification of gene expression data (Golub *et al.*, 1999; Tibshirani *et al.*, 2002; Khan *et al.*, 2001). Support Vector Machines (SVMs) have been paid attention in recent years because of their good performance in various fields, especially in the area of bioinformatics including classification of gene expression data (Furey *et al.*, 2000). However, SVMs predict a class of test samples by projecting the data into one-dimensional space based on a decision function. As a result, information loss of the original data is enormous.

Some methods are used for projecting high dimensional data into lower dimensional space to clearly exhibit the properties of the data and to roughly visualize the distribution of the data. Principal Component Analysis (PCA) (Fukunaga, 1990) and its derivatives, e.g. Nonlinear PCA (Diamantaras and Kung, 1996) and Kernel PCA (Schölkopf *et al.*, 1998), are most widely used for this purpose (Huang *et al.*, 2003). One drawback of PCA analysis is, however, that class information is not utilized for class prediction because PCA chooses axes based on the variance of overall data. Thus clearly separable data in the original space may not be so in the reduced space used in PCA. Another method for visualization and reducing dimension of data is discriminant analysis. It chooses axes based on class information in terms of within- and between-class variance. However, it is reported that SVMs often outperform discriminant analysis (Brown *et al.*, 2000).

The main purpose of this paper is to cover the shortcoming of SVMs by introducing multiple orthogonal axes for reducing dimensions and visualization of gene expression data. To this end, we have developed multidimensional SVMs (MD-SVMs), a new SVM-based method that generates multiple orthogonal axes based on margin between two

[*]To whom correspondence should be addressed.

Komura *et al.* (2004) Multidimensional Support Vector Machines for Visualization of Gene Expression Data. Symposium on Applied Computing, Proceedings of the 2004 ACM symposium on Applied computing, 175-179; http://doi.acm.org/10.1145/967900.967936 classes to minimize generalization errors. The axes generated by this method reduce dimensions of original data to extract information useful in estimating the discriminability of two classes. This method fulfills the requirement of both visualization and class prediction. The basic properties of SVMs are retained in our method: solutions of mathematical programming are sparse, and nonlinear classification of data is implemented implicitly through the use of kernel functions.

This paper is organized as follows. In Section 2, we introduce the fundamental of SVMs. In Section 3, we describe the algorithm of MD-SVMs. In Section 4 and 5, we show numerical experiments on real gene expression datasets and reveal that our algorithm is effective for data visualization and class prediction.

1.1 Notation

$\mathbb{R}$ is defined as the set of real numbers. Each component of a vector $x \in \mathbb{R}^n, i = 1,\ldots,m$ will be denoted by $x_j, j = 1,\ldots,n$. The inner product of two vectors $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^n$ will be denoted by $x \cdot y$. For a vector $x \in \mathbb{R}^n$ and a scalar $a \in \mathbb{R}, a \leq x$ is defined as $a \leq x_i$ for all $i = 1,\ldots,n$. For an arbitrary variable x, x^k is just a name of the variable with upper suffix, not defined as k-th power of x.

2 SUPPORT VECTOR MACHINES

Since details of SVMs are fully described in the articles (Vapnik, 1998; Cristianini and Shawe-Taylor, 2000), we briefly introduce the fundamental principle of SVMs in this section. We consider a binary classification problem, where a linear decision function is employed to separate two classes of data based on m training samples $x_i \in \mathbb{R}^n, i = 1,\ldots,m$ with corresponding class values $y_i \in \{\pm 1\}, i = 1,\ldots,m$. SVMs map a data $x \in \mathbb{R}^n$ into a higher, probably infinite, dimensional space $\mathbb{R}^N$ than the original space with an appropriate nonlinear mapping $\phi : \mathbb{R}^n \rightarrow \mathbb{R}^N, n < N$. They generate the linear decision function of the form $f(x) = \text{sign}(w \cdot \phi(x) + b)$ in the high dimensional space, where $w \in \mathbb{R}^N$ is a weight vector which defines a direction perpendicular to the hyperplane of the decision function, while $b \in \mathbb{R}$ is a bias which moves the hyperplane parallel to itself. The optimal decision function given by SVMs is a solution of an optimization problem

$$\min_{w,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i,$$
$$\text{s.t. } y_i(w \cdot \phi(x_i) + b) \geq 1 - \xi_i, \quad i = 1,\ldots,m, \ \xi \geq 0, \quad (1)$$

with $C > 0$. Here, $\xi \in \mathbb{R}^m$ is a vector whose elements are slack variables and $C \in \mathbb{R}$ is a regularization parameter for penalizing training errors. When $C \rightarrow \infty$, no training errors are allowed, and thus this is called hard margin classification. When $0 < C < \infty$, this is called soft margin classification because it allows some training errors. Note that a geometric margin γ between two classes is defined as $\frac{1}{\|w\|^2}$. The optimization problem formalizes the tradeoff between maximizing margin and minimizing training errors. The problem is transformed into its corresponding dual problem by introducing lagrange multiplier $\alpha \in \mathbb{R}^m$ and replacing $\phi(x_i) \cdot \phi(x_j)$ by kernel function $K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$ to be solved in an elegant way of dealing with a high dimensional vector space. The dual problem is

$$\max_{\alpha} -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j K(x_i, x_j) + \sum_{i=1}^{m}\alpha_i,$$
$$\text{s.t. } 0 \leq \alpha \leq C, \sum_{i=1}^{m}\alpha_i y_i = 0. \quad (2)$$

By virtue of the kernel function, the value of the inner product $\phi(x_i) \cdot \phi(x_j)$ can be obtained without explicit calculation of $\phi(x_i)$ and $\phi(x_j)$. Finally, the decision function becomes $f(x) = \text{sign}(\sum_{i=1}^{m}\alpha_i y_i K(x_i, x) + b)$. by using kernel functions between training samples $x_i, i = 1,\ldots,m$ and a test sample x.

3 MULTIDIMENSIONAL SUPPORT VECTOR MACHINES

In order to overcome the drawback that SVMs cannot generate more than one decision function, we propose a SVM-based method that can be used for both data visualization and class prediction in this section. We call this method multidimensional SVMs (MD-SVMs). We deal with the same problem as mentioned in Section 2. Conventional SVMs give an optimal solution set (w, b, ξ) which corresponds to a decision function, while our MD-SVMs give the multiple sets $(w^k, b^k, \xi^k), k = 1, 2, \ldots, l$ with $l \leq n$, so that all the directions w_k are orthogonal to one another. The orthogonal axes can be used for reducing the dimension of original data and data visualization in three dimensional space by means of projection. Here the first set (w^1, b^1, ξ^1) is equivalent to that obtained by conventional SVMs. Now we only refer to the steps of obtaining $(w^k, b^k, \xi^k), k = 2, 3, \ldots, l$. In practice, the k-th set $(w^k, b^k, \xi^k)k = 2, 3, \ldots, l$ are found with iterative computations of the optimization problem

$$\min_{w^k,\xi^k} \frac{1}{2}\|w^k\|^2 + C\sum_{i=1}^{m}\xi_i^k,$$
$$\text{s.t. } y_i(w^k \cdot \phi(x_i) + b^k) \geq 1 - \xi_i^k, i = 1,\ldots,m,$$
$$\xi^k \geq 0, w^k \cdot w^j = 0, j = 1,\ldots,k-1. \quad (3)$$

This problem differs from that of conventional SVMs in the last constraint $w^k \cdot w^j = 0$. The weight vector $w^j, j = 1,\ldots,k-1$ should be computed in advance by solving other optimization problems (3). The optimization problem is modified by introducing lagrange multipliers $\boldsymbol{\alpha}^k, \boldsymbol{\gamma}^k \in \mathbb{R}^m$, $\boldsymbol{\beta}^k \in \mathbb{R}^{k-1}$ and kernel functions. The primal Lagrangian is

$$\begin{aligned} L(w^k, b^k, \xi^k) = \frac{1}{2} \| w^k \|^2 + C \sum_{i=1}^{m} \xi_i^k \\ + \sum_{i=1}^{m} \alpha_i^k (1 - \xi_i^k - y_i (w^k \cdot \phi(x_i) + b^k)) \\ + \sum_{j=1}^{k-1} \beta_j^k (w^k \cdot w^j) - \sum_{i=1}^{m} \gamma_i^k \xi_i. \end{aligned} \tag{4}$$

Consequently, the optimization problem is

$$\begin{aligned} \max_{\boldsymbol{\alpha}^k, \boldsymbol{\beta}^k} -\frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i^k \alpha_j^k y_i y_j K(x_i, x_j) \\ + \frac{1}{2} \sum_{i=1}^{k-1} \beta_i^k \beta_i^k (w^i \cdot w^i) + \sum_{i=1}^{m} \alpha_i^k, \\ \text{s.t. } 0 \le \boldsymbol{\alpha}^k \le C, \sum_{i=1}^{m} \alpha_i^k y_i = 0, \\ \sum_{i=1}^{m} \alpha_i^k y_i (\phi(x_i) \cdot w^j) = 0, j = 1, \ldots, k-1 \end{aligned} \tag{5}$$

Here $\phi(x_p) \cdot w^q$ and $w^p \cdot w^p$ are calculated recursively as follows:

$$\phi(x_p) \cdot w^q = \sum_{i=1}^{m} \alpha_i^q y_i K(x_p, x_i) - \sum_{i=1}^{q-1} \beta_i^q (\phi(x_p) \cdot w^i), \tag{6}$$

$$\begin{aligned} w^p \cdot w^p = \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i^p \alpha_j^p y_i y_j K(x_i, x_j) \\ - \sum_{i=1}^{m} \sum_{j=1}^{p-1} \alpha_i^p y_i \beta_j^p (\phi(x_i) \cdot w^j) + \sum_{i=1}^{p-1} \beta_i^p \beta_i^p (w^i \cdot w^i) \\ - \sum_{i=1}^{m} \sum_{j=1}^{p-1} \alpha_i^p y_i \beta_j^p (\phi(x_i) \cdot w^j), \end{aligned} \tag{7}$$

where $\phi(x_p) \cdot w^1 = \sum_{i=1}^{m} \alpha_i^1 y_i K(x_p, x_i)$ and $w^1 . w^1 = \sum_{i=1}^{m} \alpha_i^1 y_i (\phi(x_i), w^1)$. As can be seen, there is no need to calculate nonlinear map of data $\phi(x)$ in problem (5) because all nonlinear mappings can be replaced with kernel functions.

Note that this optimization problem is a nonconvex quadratic problem when k is more than 1. As a consequence, the optimal solutions are not easy to be obtained. In Section 4, we use local optimum for numerical experiments when k is 2 or 3. We note the experimental results are still encouraging.

The corresponding Karush–Kuhn–Tucker conditions are

$$\alpha_i^k \{1 - \xi_i^k - y_i (w^k \cdot \phi(x_i) + b^k)\} = 0, \tag{8}$$

$$\xi_i^k (\alpha_i^k - C) = 0, i = 1, \ldots, m. \tag{9}$$

These are exactly the same as conventional SVMs. We highlight the other properties conserved from conventional SVMs:

- Projecting data into high dimensional space is implicit, using kernel functions to replace inner products.
- The solutions $\boldsymbol{\alpha}^k$ of the optimization problem is sparse. Then the corresponding decision function depends only on few 'Support Vectors'.

Since each decision function is normalized independently to hold $w^k \cdot \phi(x_i) + b^k = y_i$ for $i = 1, \ldots, m$, data scales of the axes should be aligned with first axis ($k = 1$) for visualization. The margin γ^k, the L2-distance between support vectors of each class of k-th axis, is

$$\left(\sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i^k \alpha_j^k y_i y_j K(x_i, x_j) - \sum_{i=1}^{k-1} \beta_i^k \beta_i^k (w^i \cdot w^i) \right)^{-\frac{1}{2}}. \tag{10}$$

So a scaling factor $s^k = \gamma^1 / \gamma^k$ is

$$\sqrt{\frac{\sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i^1 \alpha_j^1 y_i y_j K(x_i, x_j)}{\sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i^k \alpha_j^k y_i y_j K(x_i, x_j) - \sum_{i=1}^{k-1} \beta_i^k \beta_i^k (w^i \cdot w^i)}}. \tag{11}$$

The decision function of k-th step has the form $f^k(x) = \text{sign}\left(\sum_{i=1}^{m} \alpha_i^k y_i K(x_i, x) + b^k\right)$. Since the right hand side of the equation has the function of projecting original data into one dimensional space, the data can be plot in up to three dimensional space for visualization. The coordinate of data $x \in \mathbb{R}^m$ in three dimensional space is

$$(s^{k_1} g^{k_1}(x), s^{k_2} g^{k_2}(x), s^{k_3} g^{k_3}(x)), \tag{12}$$

where $g^k(x) = \sum_{i=1}^{m} \alpha_i^k y_i K(x_i, x) + b^k$. The space represents a distribution of data clearly based on the margin between two classes.

4 NUMERICAL EXPERIMENTS

4.1 Method

In order to confirm the effectiveness of our algorithm, we have performed numerical experiments. MD-SVMs can generate multiple axes, up to the number of features. Here we choose three axes, $k = 1, 2, 3$, to simplify the experiments. When k is

2 or 3, we use local optimum in problem (5) since it is difficult to obtain the global solutions. In our experiments, we carry out hold-out validation because cross-validation changes decision functions every time the dataset is split. Then we compare the results obtained by MD-SVMs with those obtained by PCA.

In the experiments, the expression values for each of the genes are normalized such that the distribution over the samples has a zero mean and unit variance. Before normalization, we discard genes in the dataset with the overall average value less than 0.35. Then we calculate a score $F(x(j)) = |(\mu^+(j)-\mu^-(j))/(\sigma^+(j)+\sigma^-(j))|$, for the remaining genes. Here $\mu^+(j)(\mu^-(j))$ and $\sigma^+(j)(\sigma^-(j))$ denote the mean and standard deviation of the j-th gene of the samples labeled $+1(-1)$, respectively. This score becomes the highest when the corresponding expression levels of the gene differ most in the two classes and have small deviations in each class. We select 100 genes with the highest scores and use them for hold-out validation. These procedures for gene selection are done only for training data for fair experiments.

The regularization parameter C in problem (5) is set to 1000. This value is rather large but finite because we would like to avoid ill-posed problems in a hard margin classification. We choose linear kernel $K(x_i,x_j) = x_i \cdot x_j$ and RBF kernel $K(x_i,x_j) = \exp -\gamma\|x_i - x_j\|^2$ with $\gamma = 0.001$ in the experiments of MD-SVMs.

4.2 Materials

Leukemia dataset (Golub et al., 1999) This gene expression dataset consists of 72 leukemia samples, including 25 acute myeloid leukemia (AML) samples and 47 acute lymphoblastic leukemia (ALL) samples. They are obtained by hybridization on the Affymetrix GeneChip containing probe sets for 7070 genes. Training set contains 20 AML samples and 42 ALL samples. Test set contains 5 AML samples and 5 ALL samples. AML samples are labeled +1 and ALL samples are labeled −1.

Lung tissue dataset (Bhattacharjee et al., 2001) This dataset consists of 203 samples from lung tissue, including 16 samples from normal tissue and 187 samples from cancerous tissue, and is obtained by hybridization on the Affymetrix U95A Genechip containing probe sets for 12558 genes. Training set includes 13 samples from normal tissue and 157 samples from cancerous tissue. Test set includes 3 samples from normal tissue and 30 samples from cancerous tissue. Samples from normal tissue are labeled +1 and samples from cancerous tissue are labeled −1.

5 RESULTS AND DISCUSSION

The results of numerical experiments are shown in Figure 1, and Tables 1 and 2. The distributions obtained by MD-SVMs on the leukemia dataset and the lung tissues dataset are given in Figure 1-(1) and 1-(3), respectively. Those obtained by PCA are given in Figure 1-(2) and 1-(4), respectively. The number of misclassified samples by MD-SVMs are summarized in Table 1 and 2. In these tables, the class of the samples is predicted based on decision functions $f^k(x), k = 1,2,3,$ corresponding to each of the three axes.

Figure 1-(1) and 1-(3) illustrate that MD-SVMs are likely to separate the samples of each class in all the three directions. However, as shown in Figure 1-(2) and 1-(4), PCA does not separate the samples in the directions of the 2nd or the 3rd axis. These axes by PCA are dispensable with the objective of visualization for class prediction. In other words, MD-SVMs gather the plots of the samples into the appropriate clusters of each class, while PCA rather scatters them. Furthermore, in the distribution by MD-SVMs for the lung tissues dataset, one sample outlies from correct clusters (indicated by arrows in Figure 1-(3)). Though this sample also seems to be an outlier in the distribution by PCA (also indicated in Figure 1-(4)), the outlier significantly deviates in MD-SVMs. This may arise from the fact that MD-SVMs can separate the samples in all the directions. These observations indicate that MD-SVMs are well suited for visualizing in binary classification problems.

The significant advantage of MD-SVMs over PCA is the ability to predict the classes. MD-SVMs can predict the classes of samples based on the decision functions $f^k(x)$ without extra computation, while PCA cannot. The predicted class of a sample should be matched by the all the decision functions in an ideal case. However that does not always occur as seen in Tables 1 and 2. In such cases, the simplest method for prediction is to use only the 1st axis, which corresponds to the decision function generated by conventional SVMs. The idea is supported by the fact that the 1st decision function classifies the samples most correctly in almost all cases in Tables 1 and 2. The more advanced method is weighted voting. Scaling factor or normalized objective values in problem (5) are the candidate of the weight.

Multiple decision functions generated by MD-SVMs are useful for outlier detection. Samples misclassified by multiple decision functions may be mis-labeled or categorized into unknown classes. For example, see the column '3 axes' of test sample of the lung tissues dataset with RBF kernel in Table 2. This sample is misclassified by all decision functions, so we can say that this data contains some experimental error. The hierarchical clustering method also supports our result. These results indicate that MD-SVMs can be used for finding candidates of outliers.

6 CONCLUSION

For both visualization and class prediction of gene expression data, we propose a new method called Multidimensional Support Vector Machines. We formulate the method as a quadratic program and implement the algorithm. This is motivated by the following facts: (1) SVMs perform better than the other classification algorithms, but they generate only one axis for class prediction. (2) PCA chooses multiple