296–302) partial sequences of GPC3, respectively, indicating that this band is derived from an NH_2-terminal portion of GPC3. We designated this soluble cleaved fragment of GPC3 as sGPC3.

To further characterize sGPC3, we next tried to precisely identify the undetermined cleavage site by sequencing the residual COOH-terminal portion of GPC3 (designated cGPC3). However, the corresponding band was not visible by SDS-PAGE, presumably due to attachment of heparan sulfate glycosaminoglycan, leading to smearing (Fig. 1*A*). After substituting the two heparan sulfate attachment sites of the expression construct and purifying the resultant GPC3ΔGPIΔHS, we could observe a band of M_r 30,000, as expected (Fig. 1*A*). The NH_2-terminal sequence of this band was identified as SAYYPEDLF, identical to amino acids 359–367 of GPC3. Thus, the cleavage site was identified as being between Arg^{358} and Ser^{359} (Fig. 1*B*). We do not have precise information on the NH_2-terminal sequence of sGPC3 due to modification, but considering that amino acid 1–24 is a putative signal sequence, sGPC3 is likely to consist of amino acids 25–358 with an estimated molecular weight of 38,100, consistent with the M_r 40,000 band observed in SDS-PAGE (Fig. 1*A*).

Soluble GPC3 Is a Major Form of GPC3 Specifically Detected in the Sera of Patients with HCC. We succeeded in generating a number of high-affinity mAbs specific for GPC3 and classified these antibodies into two groups, N-mAbs and C-mAbs, according to their epitopes within amino acids 25–358 or 359–563, respectively (data not shown). These antibodies could also recognize endogenous GPC3 protein in immunoblotting: core protein (M_r 66,000) and glycanated form (smearing) of GPC3 were detected by both N-mAbs and C-mAbs; whereas sGPC3 (M_r 40,000) was detected only by N-mAbs (Fig. 2*A*). An additional M_r 50,000 band was detected strongly in the cell lysate of HepG2 with both N-mAbs and C-mAbs (Fig. 2*A*). This band was only weakly detectable in HuH6 cells and was undetectable in five other hepatoma cell lines (Fig. 2, *A* and *C*; data not shown), suggesting cell-specific variations in the processing of the protein. In the culture supernatant, sGPC3, rather than a core protein or a glycanated form of GPC3, was the major form of GPC3 detected (Fig. 2*A*).

Based on the above *in vitro* finding, we speculated that sGPC3, instead of core protein of GPC3, might be the major form of GPC3 in the sera of HCC patients. To avoid possible interference on immunoblotting by significant migration of albumin or immunoglobulin in the serum, we performed immunoprecipitation before immunoblottng using three N-mAbs (Fig. 2*B*). sGPC3 alone was successfully detected by immunoprecipitation with M18D04 (Fig. 2*C*) or M19B11 (data not shown) followed by immunoblotting with A1836A in the sera of patients with HCC, but not in sera from normal liver (NL). These results clearly demonstrate that sGPC3 is the major diagnostic target specifically detectable in the sera of HCC patients.

Soluble GPC3 Is Useful as a Serological Marker of WD HCC and MD HCC. We next constructed a sandwich ELISA system with these three antibodies to measure the serum level of sGPC3 (Fig. 3*A*). To verify the specificity of the assay, we performed immunoblotting of 10 sera samples from HCC with sGPC3 levels ranging from 4.0 to 55.0 ng/ml and 3 samples from NL with sGPC3 levels of <0.1 mg/ml. We detected only sGPC3 in all 10 HCC samples, whereas no band was detected in 3 samples from NL, indicating high sensitivity and specificity of the assay (Fig. 3*B*). When we examined sera from 69 cases with HCC, 38 cases with LC, and 96 cases with NL, the level of sGPC3 (mean ± SD) was 4.84 ± 8.91 ng/ml for HCC, 1.09 ± 0.74 ng/ml for LC, and 0.65 ± 0.32 ng/ml for NL and was significantly higher in HCC than in NL ($P < 0.001$, Student's t test) or in LC ($P < 0.01$; Fig. 3*C*).

We then evaluated sGPC3 as a general marker for HCC in comparison with AFP. Initial analysis of the receiver-operating characteristic curve using the data from 69 cases with HCC and 38 cases with LC suggested that, used in isolation, sGPC3 is not as good as AFP: the calculated area under the receiver-operating characteristic curve was 0.729 for sGPC3 and 0.799 for AFP (Fig. 3*D*). The sensitivity and specificity of sGPC3 for the diagnosis of HCC (cutoff value, 2.0 ng/ml) were 51% and 90%, respectively, whereas those of AFP measured in parallel (cutoff value, 20 ng/ml) were 55% and 90%, respectively. AFP and sGPC3 were not correlated ($r = 0.13$), and combination measurement of both markers markedly improved sensitivity to 72%.

HCC may be divided into two subgroups correlating to the extent of disease: (*a*) one first treated by surgery, mainly with a solitary tumor or few tumors; and (*b*) the second treated with transcatheter arterial chemoembolization, mostly with multiple and advanced tumors. The serum level of sGPC3 was 2.61 ± 2.69 ng/ml for the former group,

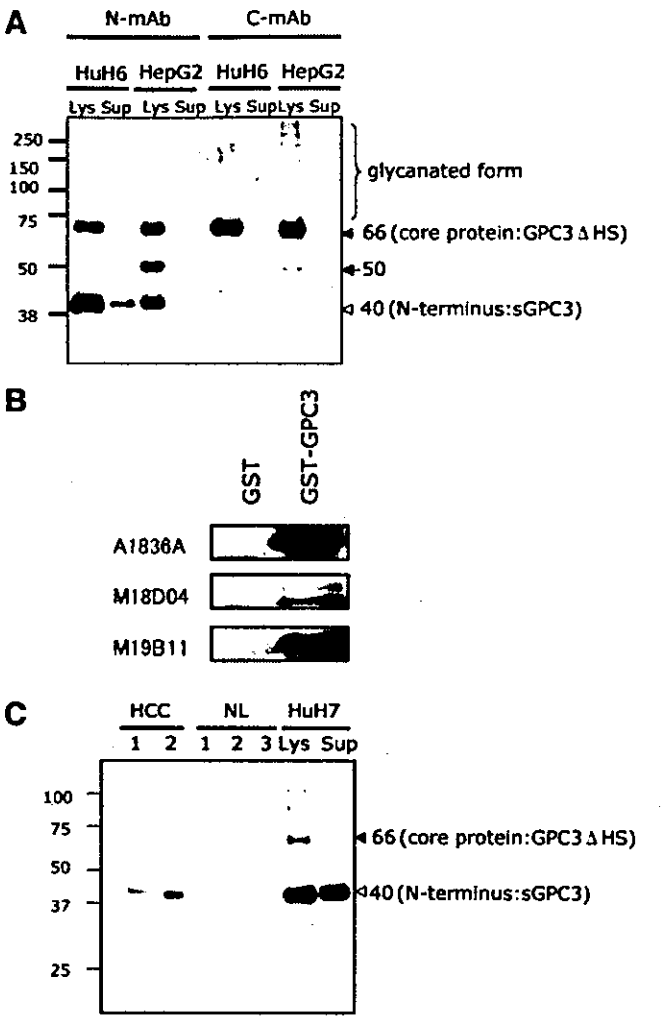

Fig. 2. Characterization of endogenous glypican-3 (GPC3) proteins with monoclonal antibodies (mAbs). *A*, representative immunoblotting of endogenous GPC3 in the cell lysate and culture supernatant with N-mAb and C-mAb. HepG2 and HuH6 were analyzed. Note that soluble GPC3 (sGPC3) alone is detected in the culture supernatant. *Brace*, glycanated GPC3 (smearing); *closed arrowhead*, core protein of GPC3 (M_r 66,000); *open arrowhead*, sGPC3 (M_r 40,000); *arrow*, uncharacterized processed fragment of GPC3 (M_r 50,000). *Lys*, lysate; *Sup*, supernatant of culture media. *B*, immunoblotting analysis with anti-GPC3 antibodies A1836A, M18D04, and M19B11 recognized glutathione *S*-transferase-sGPC3 but did not recognize glutathione *S*-transferase. *C*, detection of sGPC3 alone in the sera of the patients with hepatocellular carcinoma. Sera from two patients with hepatocellular carcinoma and three healthy adults (NL) were analyzed by immunoprecipitation with M18D04 followed by immunoblotting with A1836A. HuH7 cells were analyzed as a reference. *Closed arrowhead*, core protein of GPC3 (M_r 66,000); *open arrowhead*, sGPC3 (M_r 40,000).

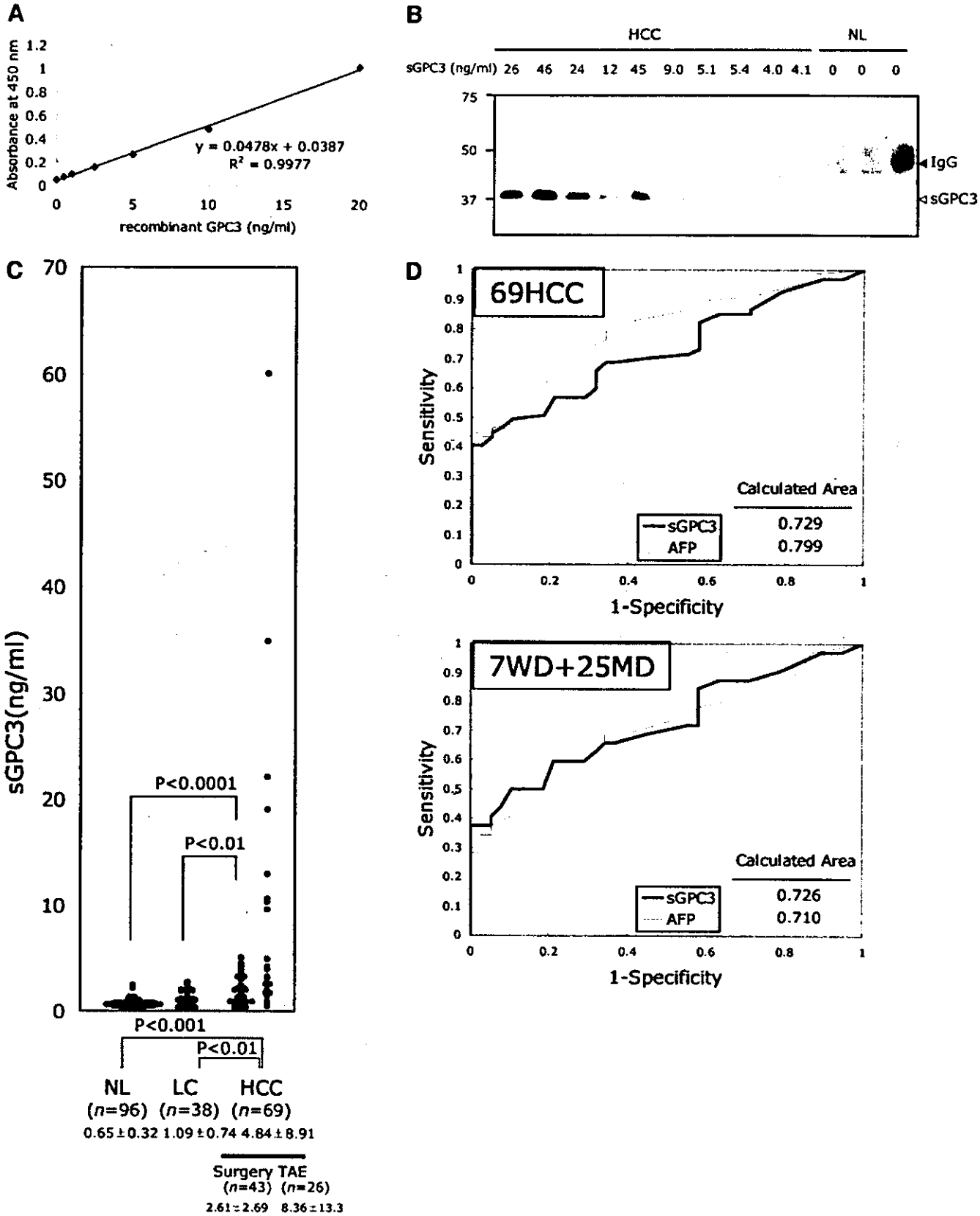

Fig. 3. Evaluation of soluble glypican-3 (sGPC3) as a serological marker of hepatocellular carcinoma (HCC). *A*, standard curve of sandwich ELISA. *B*, high specificity of sandwich ELISA. Specific detection of sGPC3 alone solely in the sera with elevated sGPC3 level measured with sandwich ELISA. Sera from 10 patients with HCC and 3 healthy adults (*NL*) were analyzed by immunoprecipitation with M18D04 followed by immunoblotting with A1836A. Serum sGPC3 level is indicated for each sample. *Open arrowhead*, sGPC3 (M_r 40,000); *closed arrowhead*, IgG. *C*, distribution of sGPC3 in the sera of patients with normal liver, liver cirrhosis (*LC*), and HCC (surgery and transcatheter arterial chemoembolization subgroup). Mean ± SD (ng/ml) of serum sGPC3 is indicated. Number of samples is indicated as *n*. *D*, receiver-operating characteristic curve analysis of sGPC3 (*thick line*) and α-fetoprotein (*thin line*). *Top panel*, all of the 69 HCCs and 38 cases of LC were included in the analysis. *Bottom panel*, 32 HCCs (including 7 well-differentiated and 25 moderately differentiated HCCs) and 38 cases of LC were analyzed. Area under the receiver-operating characteristic curve is indicated.

significantly higher than that for NL ($P < 0.0001$, Student's t test) or LC ($P < 0.01$), and 8.36 ± 13.3 ng/ml for the latter group (Fig. 3*C*), suggesting that the serum level of sGPC3 is elevated in an earlier stage and rises as HCC progresses. We then evaluated sGPC3 as a marker for HCC in relatively early-stage disease. When 43 cases treated by surgery were confined to 32 cases with relatively early-stage HCC (7 cases with WD HCC and 25 cases with MD HCC), calculated areas under the receiver-operating characteristic curve for sGPC3 and AFP were 0.726 and 0.710, respectively, indicating that sGPC3 is superior to AFP (Fig. 3*D*). The sensitivity of sGPC3 and AFP for the diagnosis of WD HCC and MD HCC was 50% and 47%, respectively. Moreover, combination measurement of both markers in WD HCC and MD HCC also markedly improved sensitivity to 72%. These results clearly demonstrate the utility of sGPC3 as a serological marker for HCC, especially for relatively early-stage HCC, and its complementarity to AFP.

DISCUSSION

GPC3 (alternatively called OCI-5 or MXR-7) is a heparan sulfate proteoglycan. The structural characteristics of the glypican family are (*a*) a core protein of approximately M_r 60,000, (*b*) binding to the membrane through GPI anchor, (*c*) heparan sulfate glycosaminoglycan attachment at Ser-Gly sequence within the COOH-terminal portion, and (*d*) a highly conserved pattern of 14 Cys residues (19). *GPC3* was originally isolated as a gene that is developmentally expressed in fetal rat intestine (21, 22). Mutation of *GPC3* is found in Simpson-Golabi-Behmel syndrome characterized by an overgrowth phenotype, hence its putative function was associated with an apoptotic effect (23). Silencing of *GPC3* in some types of cancer (24–26) is in line with this notion.

Overexpression of *GPC3* mRNA in HCC has been reported by ourselves and several other groups (15–18), although the role of GPC3 in carcinogenesis or progression of HCC has yet to be determined. In general, transcription level and protein level do not necessarily correlate. We have succeeded previously in generating an anti-GPC3 mAb against a peptide within the COOH-terminal portion, and we demonstrated using the antibody that the expression level of GPC3 core protein correlated well with its transcription level and that GPC3 was also overexpressed at protein level for the first time (15). Difficulties in making high-affinity antibodies against GPC3 (27), presumably due to its complex structure derived from disulfide bonds between 14 Cys residues, prohibited further analysis. We tried to generate high-affinity mAbs again by using recombinant GPC3 protein expressed in mammalian cells as an immunogen, and we finally succeeded in generating numerous high-affinity mAbs; to our knowledge, this is the first establishment of mAbs that can react with sGPC3. We did not recognize sGPC3 in a previous study (15) because we used a mAb against a relatively COOH-terminal portion (amino acids 355–371).

In the present work, we have precisely characterized GPC3 and demonstrated that the M_r 40,000 protein, sGPC3, derives from the NH_2-terminal portion of GPC3 and is cleaved between Arg^{358} and Ser^{359}. The M_r 40,000 protein was previously described by Mast *et al.* (19), who were searching for the binding protein on the plasma membrane of HepG2 cells for tissue factor pathway inhibitor. They purified a M_r 40,000 protein from culture supernatant of HepG2 cells and showed that it was derived from the NH_2-terminal portion of GPC3. They did not identify a cleavage site for the protein, unlike our study, but it is highly likely that the soluble protein they observed is sGPC3. They described purification of a M_r 40,000 protein only when protease inhibitors were used throughout the procedure, strongly suggesting that GPC3 cleavage is mediated by a protease (19). In addition, they found that washing the cells with dextran sulfate or heparin released significantly higher amounts of GPC3 than seen before treatment, strongly suggesting that most GPC3 is noncovalently attached to the cell surface after cleavage of the GPI anchor, but not in the culture supernatant (19). Our finding that sGPC3 alone is the major form of GPC3 in the culture supernatant of hepatoma cells and the serum of patients with HCC is consistent with these findings.

Very recently, two other groups reported elevated levels of GPC3 in the serum of HCC patients. The results still seem preliminary, although they are quite similar to ours. Here, we have made significant improvements in the reliability of the assay. Nakatsura *et al.* (28) used a polyclonal antibody raised against 303–464 amino acids of GPC3 in their analysis. The specificity of their ELISA is to be confirmed because it is not sandwich ELISA, despite the many nonspecific bands the antibody detected in their immunoblotting. Moreover, the standard used in the assay was not recombinant GPC3 but a supernatant of HepG2 cells that is a mixture of many heterogeneous proteins. It is possible that they are measuring a mixture of nonspecific but HCC-related proteins. Capurro *et al.* (29) used a polyclonal antibody and a mAb, both raised against the last 70 amino acids of the COOH-terminal portion of GPC3, to detect glycanated GPC3 in serum with their sandwich ELISA. However, the major detectable form of GPC3 in serum is sGPC3, which cannot be detected with these antibodies against the COOH-terminal portion, as shown clearly in the present study. In fact, we examined many combinations of mAbs in our sandwich ELISA, but we could detect signal only when we used a combination of two N-mAbs (data not shown). Furthermore, the only evidence reported previously for the extracellular localization of glycanated GPC3 is immunoblotting of HepG2 cell culture supernatant, rather than serum from HCC patients. Here, we demonstrated that sGPC3 is in the culture supernatant and serum of the HCC patients using both immunoblotting and sandwich ELISA with the same combination of mAbs. One possible interpretation of the result, obtained by Capurro *et al.*, is that they are detecting some short fragments derived from a COOH-terminal portion but not the glycanated form of GPC3, and this issue should be further investigated.

We have delineated the usefulness of sGPC3 as highly sensitive to early-stage HCC. In addition, there were several cases with elevated serum sGPC3 among LC patients, although not included in this study, where HCC developed within 6 months after serum examination or some tumor was already detected by ultrasound without final diagnosis of HCC by computed tomography or angiography. We have also demonstrated the complementarity of sGPC3 to another HCC marker, AFP. These findings promise future bedside use of sGPC3 as a serological marker of HCC. Another attractive aspect of GPC3 is that the membrane-anchored portion is a potential target for antibody therapy. In this context, diagnosis with serum sGPC3 is useful not only in early detection of HCC but also for future identification of patients with high sGPC3 levels for tailor-made HCC therapy. Thus, further investigation into the clinical aspects of GPC3 in HCC is warranted.

ACKNOWLEDGMENTS

We thank H. Meguro and S. Fukui for excellent technical assistance and H. Satoh for providing pGEX-5X-sGPC3 construct.

REFERENCES

1. Befeler AS, Di Bisceglie AM. Hepatocellular carcinoma: diagnosis and treatment. Gastroenterology 2002;122:1609–19.

2. Gebo KA, Chander G, Jenckes MW, et al. Screening tests for hepatocellular carcinoma in patients with chronic hepatitis C: a systematic review. Hepatology 2002;36: S84–92.
3. Johnson PJ. The role of serum α-fetoprotein estimation in the diagnosis and management of hepatocellular carcinoma. Clin Liver Dis 2001;5:145–59.
4. Taketa K. α-Fetoprotein: reevaluation in hepatology. Hepatology 1990;12:1420–32.
5. Kanetaka K, Sakamoto M, Yamamoto Y, et al. Overexpression of tetraspanin CO-029 in hepatocellular carcinoma. J Hepatol 2001;35:637–42.
6. Kondoh N, Shuda M, Tanaka K, et al. Enhanced expression of S8, L12, L23a, L27 and L30 ribosomal protein mRNAs in human hepatocellular carcinoma. Anticancer Res 2001;21:2429–33.
7. Scuric Z, Stain SC, Anderson WF, Hwang JJ. New member of aldose reductase family proteins overexpressed in human hepatocellular carcinoma. Hepatology 1998; 27:943–50.
8. Tanaka K, Kondoh N, Shuda M, et al. Enhanced expression of mRNAs of antisecretory factor-1, gp96, DAD1 and CDC34 in human hepatocellular carcinomas. Biochim Biophys Acta 2001;1536:1–12.
9. Shirota Y, Kaneko S, Honda M, Kawai HF, Kobayashi K. Identification of differentially expressed genes in hepatocellular carcinoma with cDNA microarrays. Hepatology 2001;33:832–40.
10. Okabe H, Satoh S, Kato T, et al. Genome-wide analysis of gene expression in human hepatocellular carcinomas using cDNA microarray: identification of genes involved in viral carcinogenesis and tumor progression. Cancer Res 2001;61:2129–37.
11. Smith MW, Yue ZN, Geiss GK, et al. Identification of novel tumor markers in hepatitis C virus-associated hepatocellular carcinoma. Cancer Res 2003;63:859–64.
12. Xu XR, Huang J, Xu ZG, et al. Insight into hepatocellular carcinogenesis at transcriptome level by comparing gene expression profiles of hepatocellular carcinoma with those of corresponding noncancerous liver. Proc Natl Acad Sci USA 2001;98: 15089–94.
13. Chen X, Cheung ST, So S, et al. Gene expression patterns in human liver cancers. Mol Biol Cell 2002;13:1929–39.
14. Chuma M, Sakamoto M, Yamazaki K, et al. Expression profiling in multistage hepatocarcinogenesis: identification of HSP70 as a molecular marker of early hepatocellular carcinoma. Hepatology 2003;37:198–207.
15. Midorikawa Y, Ishikawa S, Iwanari H, et al. Glypican-3, overexpressed in hepatocellular carcinoma, modulates FGF2 and BMP-7 signaling. Int J Cancer 2003;103: 455–65.
16. Hsu HC, Cheng W, Lai PL. Cloning and expression of a developmentally regulated transcript MXR7 in hepatocellular carcinoma: biological significance and temporospatial distribution. Cancer Res 1997;57:5179–84.
17. Zhu ZW, Friess H, Wang L, et al. Enhanced glypican-3 expression differentiates the majority of hepatocellular carcinomas from benign hepatic disorders. Gut 2001;48: 558–64.
18. Zhou XP, Wang HY, Yang GS, et al. Cloning and expression of MXR7 gene in human HCC tissue. World J Gastroenterol 2000;6:57–60.
19. Mast AE, Higuchi DA, Huang ZF, et al. Glypican-3 is a binding protein on the HepG2 cell surface for tissue factor pathway inhibitor. Biochem J 1997;327:577–83.
20. Niwa H, Yamamura K, Miyazaki J. Efficient selection for high-expression transfectants with a novel eukaryotic vector. Gene (Amst.) 1991;108:193–9.
21. Filmus J, Church JG, Buick RN. Isolation of a cDNA corresponding to a developmentally regulated transcript in rat intestine. Mol Cell Biol 1988;8:4243–9.
22. Filmus J. Glypicans in growth control and cancer. Glycobiology 2001;11:19R–23R.
23. Pilia G, Hughes-Benzie RM, MacKenzie A, et al. Mutations in GPC3, a glypican gene, cause the Simpson-Golabi-Behmel overgrowth syndrome. Nat Genet 1996;12: 241–7.
24. Lin H, Huber R, Schlessinger D, Morin PJ. Frequent silencing of the GPC3 gene in ovarian cancer cell lines. Cancer Res 1999;59:807–10.
25. Xiang YY, Ladeda V, Filmus J. Glypican-3 expression is silenced in human breast cancer. Oncogene 2001;20:7408–12.
26. Kim H, Xu GL, Borczuk AC, et al. The heparan sulfate proteoglycan GPC3 is a potential lung tumor suppressor. Am J Respir Cell Mol Biol 2003;6:694–701.
27. Filmus J, Shi W, Wong ZM, Wong MJ. Identification of a new membrane-bound heparan sulphate proteoglycan. Biochem J 1995;311:561–5.
28. Nakatsura T, Yoshitake Y, Senju S, et al. Glypican-3, overexpressed specifically in human hepatocellular carcinoma, is a novel tumor marker. Biochem Biophys Res Commun 2003;306:16–25.
29. Capurro M, Wanless IR, Sherman M, et al. Glypican-3: a novel serum and histochemical marker for hepatocellular carcinoma. Gastroenterology 2003;125:89–97.

Toolbox

Physiol Genomics 21: 274–283, 2005.
First published February 15, 2005; doi:10.1152/physiolgenomics.00107.2004.

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

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

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

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

Kano, Makoto, Shuichi Tsutsumi, Nobutaka Kawahara, Yan Wang, Akitake Mukasa, Takaaki Kirino, and Hiroyuki Aburatani. A meta-clustering analysis indicates distinct pattern alteration between two series of gene expression profiles for induced ischemic tolerance in rats. *Physiol Genomics* 21: 274–283, 2005. First published February 15, 2005; doi:10.1152/physiolgenomics.00107.2004.—We have developed a visualization methodology, called a "cluster overlap distribution map" (CODM), for comparing the clustering results of time series gene expression profiles generated under two different conditions. Although various clustering algorithms for gene expression data have been proposed, there are few effective methods to compare clustering results for different conditions. With CODM, the utilization of three-dimensional space and color allows intuitive visualization of changes in cluster set composition, changes in the expression patterns of genes between the two conditions, and relationship with other known gene information, such as transcription factors. We applied CODM to time series gene expression profiles obtained from rat four-vessel occlusion models combined with systemic hypotension and time-matched sham control animals (with sham operation), identifying distinct pattern alteration between the two. Comparisons of dynamic changes of time series gene expression levels under different conditions are important in various fields of gene expression profiling analysis, including toxicogenomics and pharmacogenomics. CODM will be valuable for various types of analyses within these fields, because it integrates and simultaneously visualizes various types of information across clustering results.

time series; transcription factor; visualization

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

Changes in the Composition of the Cluster Sets

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

Article published online before print. See web site for date of publication (http://physiolgenomics.physiology.org).

Address for reprint requests and other correspondence: M. Kano, Intelligent Cooperative System, Dept. of Information Systems, Research Center for Advanced Science and Technology, Univ. of Tokyo, Tokyo 153-8904, Japan (E-mail: mkano@cyber.rcast.u-tokyo.ac.jp).

tween 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

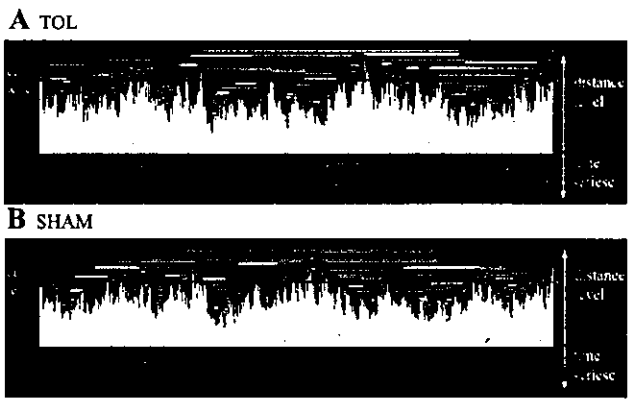

Fig. 1. Hierarchical clustering of TOL (*A*) and SHAM (*B*). We obtained time series ({0 h, 1 h, 3 h, 12 h, 24 h, 48 h} × 2) microarray data from rats with induced ischemic tolerance (tolerant rats, TOL) and rats with sham operation (sham rats, SHAM). In the analysis, we used these data sets as 12 time point ({0a, 0b, 1a, 1b, 3a, 3b, . . ., 48a, 48b} = {T_i} (i = 1,2,. . .,12)) 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.

population (6). To allow comparison among multiple arrays, the average differences were normalized for each array by assigning the mean of overall average difference values to be 100. This 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, ..., 48a, 48b} = {T_i} (i = 1,2,. . .,12)) 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:

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

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

where x_i and y_i are normalized expression levels of a gene at time point T_i (i = 1,2,. . .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:

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

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

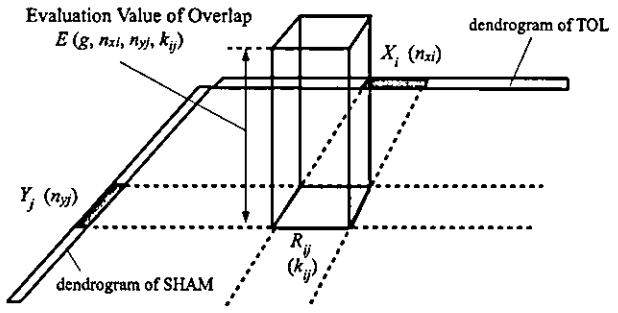

Fig. 2. Overlap block of two clusters. The dendrogram of TOL is mapped to the x-axis, and that of SHAM is mapped to the y-axis. Then, for the area (R_{ij}) determined by a cluster on the x-axis (X_i) and a cluster on the y-axis (Y_j), a block whose height represents $E(g,n_{xi},n_{yj},k_{ij})$ (statistical evaluation values of the overlaps between X_i and 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.

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

Then in the area (R_{ij}) determined by a cluster on the x-axis (X_i) and a cluster on the y-axis (Y_j), a block whose height represents $E(g,n_{xi},n_{yj},k_{ij})$ is displayed, where n_{xi} is the number of genes in X_i, n_{yj} is the number of genes in Y_j, and k_{ij} is the number of overlapping genes between X_i and Y_j (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_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}, then the small block B_{jn} will be hidden in the large block B_{im} (Fig. 3*A*). 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 overlaps 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_{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:

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

Since $\{x_{ti}\}$ and $\{y_{si}\}$ (i = 1,2,. . .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):

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

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.,

$$\forall x_{ki} \approx 0 \text{ or } \forall y_{ki} \approx 0$$

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, 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_1 is the number of genes that 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 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 the RG-U34A microarray. The 1.1-kb sequences of the 2,816 UniGenes were searched to determine whether they correspond to the TRANSFAC 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

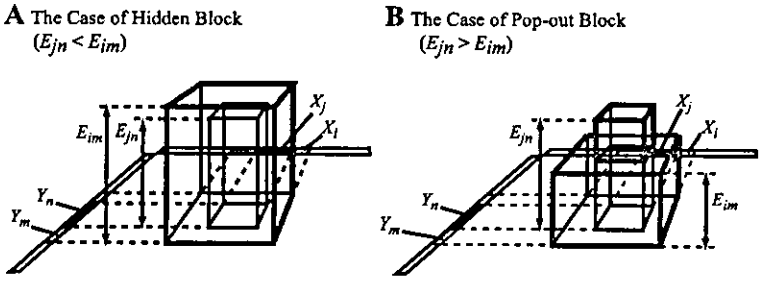

Fig. 3. Relationships of two blocks. In CODM, all of the clusters are dealt with equally, regardless of their difference levels (i.e., their homogeneity). Even if they are included in other clusters, all of the statistical significance of the number of common genes between clusters is simultaneously visualized. There is a risk that a small overlap block 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} (*A*).

Table 1. *Transcription factors linked to ischemia*

Transcription Factor	No. 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$EGR3_01	26	0.93
V$ENGFIC_01	143	0.88

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.

coded overlap blocks that contain statistically meaningful numbers of genes with putative transcription factor binding sites. If an overlap block represents statistical significance for multiple transcription factors' putative binding sites, then 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.

RESULTS AND DISCUSSION

Figure 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 Fig. 4, 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 Fig. 4 shows, the CODM provides not only a 3D mode but also a two-dimensional (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 somewhat 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 three potentially important overlap blocks (Fig. 4). The information for these three overlap blocks is shown in Table 2, their gene lists are shown in the Supplemental Material, and their expression patterns are shown in Fig. 5. (The Supplemental Material is available at the *Physiological Genomics* web site.)[1]

[1]The Supplemental Material (Supplemental Tables S1–S3) for this article is available online at **http://physiolgenomics.physiology.org/cgi/content/full/00107.2004/DC1**.

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 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. 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

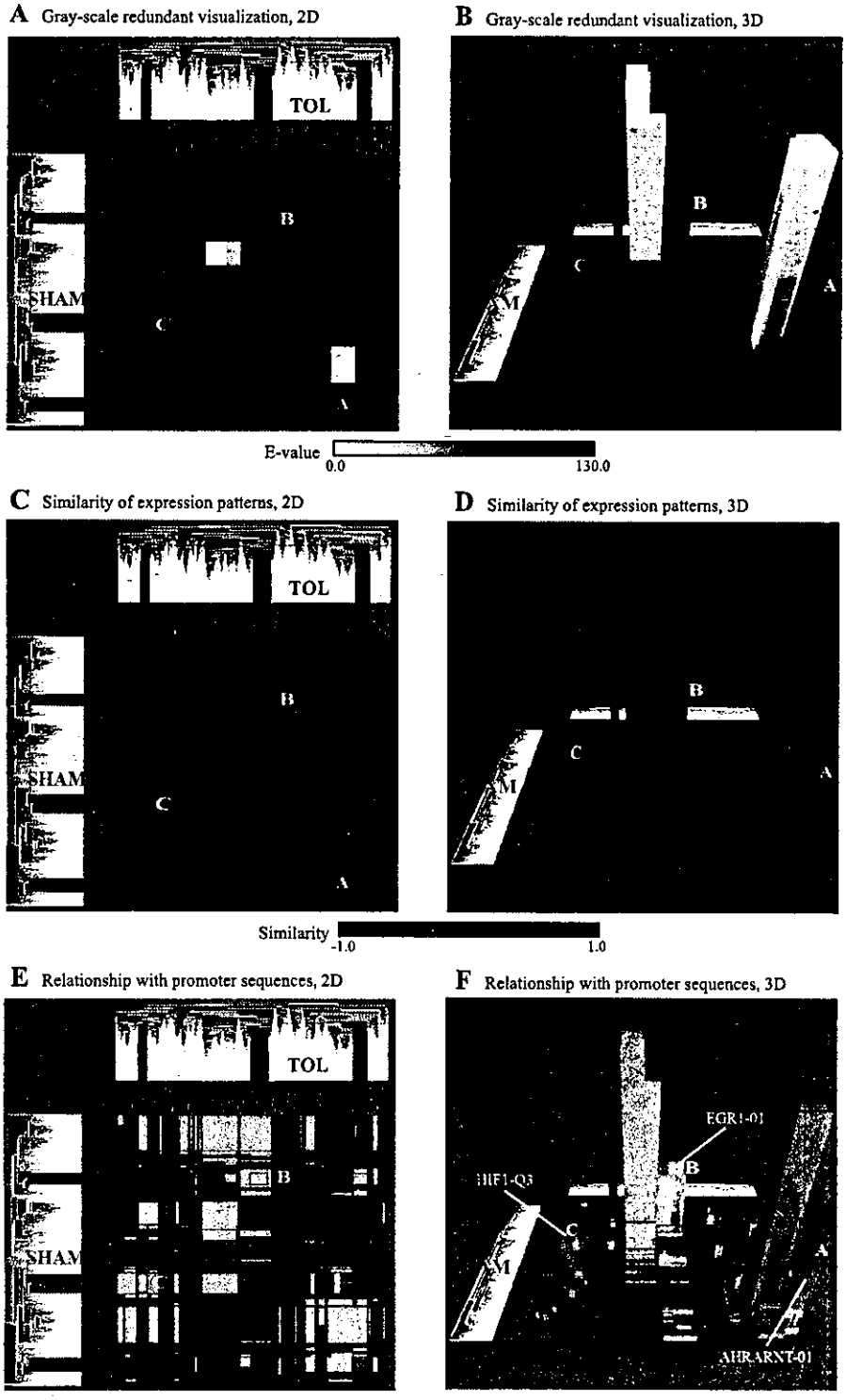

Fig. 4. Visualizations for comparison of clustering results of TOL and SHAM. These are visualization results of the comparisons between TOL and SHAM in the mode of redundant visualization (*A* and *B*), similarity of the expression patterns (*C* and *D*), and the relationships with transcription factors (*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 ($E > 2.0$).

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

Table 2. *Information about 3 overlap blocks*

Overlap Block	No. of UniGenes in Cluster of TOL	No. of UniGenes in Cluster of SHAM	No. of Common UniGenes (Evaluation Value)	Similarity $f(T,S)$	Binding Sites of Transcription Factors: No. 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.' 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, shamoperation.

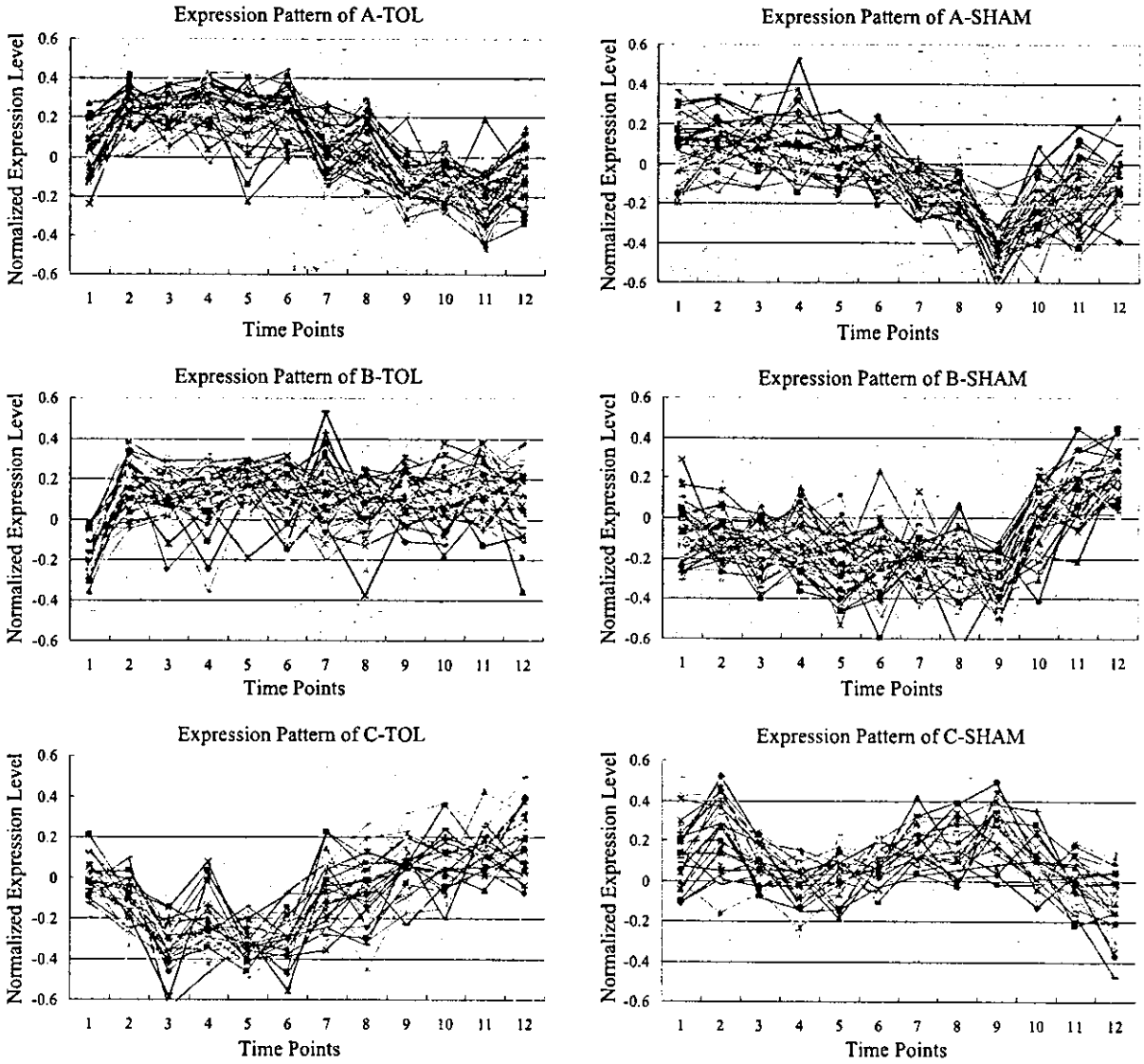

Fig. 5. Expression patterns of genes in the three overlap blocks. These are the expression patterns of common genes for the three overlap blocks that were picked up through exploration with CODM (Fig. 4). The "Expression Patterns of Cluster $T_i(/S_i)$" (i = a,b,c) are the expression patterns of the common genes of the *overlap block i* in TOL(/SHAM).

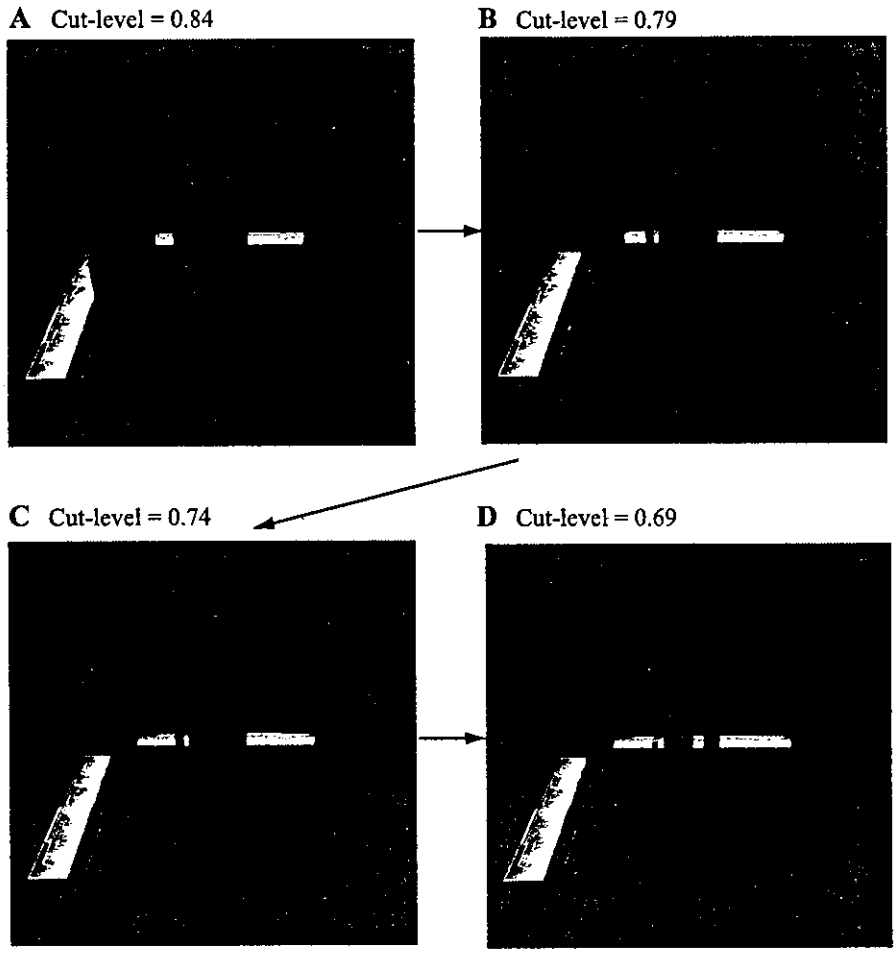

Fig. 6. Interactive changes of cut levels. In CODM, there is a risk that a small overlap block may be hidden in a large block. To avoid this problem, CODM allows the user to change the cut level interactively. If the user decreases the cut level, then some small blocks that are hidden in larger blocks 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.

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 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 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. 4*E* and Fig. 4*F*, 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)

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

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

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

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

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

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

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

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

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

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

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

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

The solutions of this problem are

$x_i = y_i \quad (i = 1,2,...,12), \quad \gamma = -1 \Rightarrow g$ has maximum value 1

or

$x_i = -y_i \quad (i = 1,2,...,12), \quad \gamma = 1 \Rightarrow g$ has the minimum value -1

Therefore,

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

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 technology. *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: 1283–1296, 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 model of genitourinary inflammation. *Physiol Genomics* 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 O_2 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.

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 $\boldsymbol{x} \in \mathbb{R}^n, i = 1,\ldots,m$ will be denoted by $x_j, j = 1,\ldots,n$. The inner product of two vectors $\boldsymbol{x} \in \mathbb{R}^n$ and $\boldsymbol{y} \in \mathbb{R}^n$ will be denoted by $\boldsymbol{x} \cdot \boldsymbol{y}$. For a vector $\boldsymbol{x} \in \mathbb{R}^n$ and a scalar $a \in \mathbb{R}, a \leq \boldsymbol{x}$ 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 $\boldsymbol{x}_i \in \mathbb{R}^n, i = 1,\ldots,m$ with corresponding class values $y_i \in \{\pm 1\}, i = 1,\ldots,m$. SVMs map a data $\boldsymbol{x} \in \mathbb{R}^n$ into a higher, probably infinite, dimensional space $\mathbb{R}^N$ than the original space with an appropriate nonlinear mapping $\phi : \mathbb{R}^n \to \mathbb{R}^N, n < N$. They generate the linear decision function of the form $f(\boldsymbol{x}) = \mathrm{sign}(\boldsymbol{w} \cdot \phi(\boldsymbol{x}) + b)$ in the high dimensional space, where $\boldsymbol{w} \in \mathbb{R}^N$ 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

$$\min_{w,\xi} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{m}\xi_i,$$

$$\text{s.t. } y_i(\boldsymbol{w} \cdot \phi(\boldsymbol{x}_i) + b) \geq 1 - \xi_i, \quad i = 1,\ldots,m, \xi \geq 0, \quad (1)$$

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 $C \to \infty$, 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 $\frac{1}{\|\boldsymbol{w}\|^2}$. 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 $\boldsymbol{\alpha} \in \mathbb{R}^m$ and replacing $\phi(\boldsymbol{x}_i) \cdot \phi(\boldsymbol{x}_j)$ by kernel function $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \phi(\boldsymbol{x}_i) \cdot \phi(\boldsymbol{x}_j)$ to be solved in an elegant way of dealing with a high dimensional vector space. The dual problem is

$$\max_{\alpha} -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) + \sum_{i=1}^{m}\alpha_i,$$

$$\text{s.t. } 0 \leq \boldsymbol{\alpha} \leq C, \sum_{i=1}^{m}\alpha_i y_i = 0. \quad (2)$$

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

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 $(\boldsymbol{w}, b, \xi)$ which corresponds to a decision function, while our MD-SVMs give the multiple sets $(\boldsymbol{w}^k, b^k, \xi^k), k = 1,2,\ldots,l$ with $l \leq n$, so that all the directions $\boldsymbol{w}_k$ 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 $(\boldsymbol{w}^1, b^1, \xi^1)$ is equivalent to that obtained by conventional SVMs. Now we only refer to the steps of obtaining $(\boldsymbol{w}^k, b^k, \xi^k), k = 2,3,\ldots,l$. In practice, the k-th set $(\boldsymbol{w}^k, b^k, \xi^k) k = 2,3,\ldots,l$ are found with iterative computations of the optimization problem

$$\min_{w^k,\xi^k} \frac{1}{2}\|\boldsymbol{w}^k\|^2 + C\sum_{i=1}^{m}\xi_i^k,$$

$$\text{s.t. } y_i(\boldsymbol{w}^k \cdot \phi(\boldsymbol{x}_i) + b^k) \geq 1 - \xi_i^k, i = 1,\ldots,m,$$

$$\xi^k \geq 0, \boldsymbol{w}^k \cdot \boldsymbol{w}^j = 0, j = 1,\ldots,k-1. \quad (3)$$

This problem differs from that of conventional SVMs in the last constraint $\boldsymbol{w}^k \cdot \boldsymbol{w}^j = 0$. The weight vector $\boldsymbol{w}^j, j = 1,\ldots,k-1$ should be computed in advance by solving

other optimization problems (3). The optimization problem is modified by introducing lagrange multipliers $\alpha^k, \gamma^k \in \mathbb{R}^m$, $\beta^k \in \mathbb{R}^{k-1}$ and kernel functions. The primal Lagrangian is

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

Consequently, the optimization problem is

$$\max_{\alpha^k,\beta^k} -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i^k\alpha_j^k y_i y_j K(x_i, x_j) + \frac{1}{2}\sum_{i=1}^{k-1}\beta_i^k\beta_i^k(w^i \cdot w^i) + \sum_{i=1}^{m}\alpha_i^k,$$

$$\text{s.t. } 0 \le \alpha^k \le C, \sum_{i=1}^{m}\alpha_i^k y_i = 0,$$

$$\sum_{i=1}^{m}\alpha_i^k y_i(\phi(x_i) \cdot w^j) = 0, j = 1, \ldots, k-1 \quad (5)$$

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

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

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

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

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

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

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 α^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

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

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

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

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

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

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 $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i \cdot \boldsymbol{x}_j$ and RBF kernel $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp -\gamma \| \boldsymbol{x}_i - \boldsymbol{x}_j \|^2$ 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 $f^k(\boldsymbol{x}), k = 1,2,3$, 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 $f^k(\boldsymbol{x})$ 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

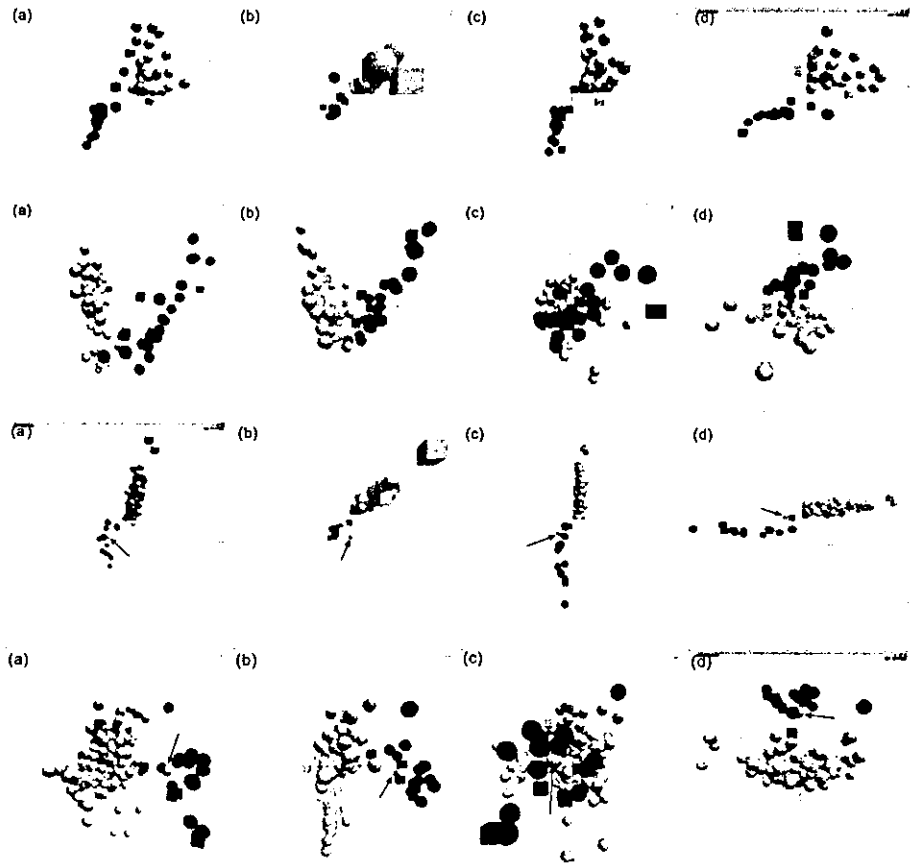

Fig. 1. (Top row) Distribution obtained by MD-SVMs for the leukemia dataset with linear kernel. (Second row) Distribution obtained by PCA on the leukemia dataset. (Third row) Distribution obtained by MD-SVMs for the lung tissues dataset with linear kernel. The sample indicated by arrows appears to be an outlier. (Fourth row) Distribution obtained by PCA for the lung tissues dataset. The sample indicated by arrows is the same as in the third row but with less deviates. (a) Cross shot, (b) 1st axis (x axis) and 2nd axis (y axis), (c) 2nd axis (x axis) and 3rd axis (y axis), (d) 3rd axis (x axis) and 1st axis (y axis). Black objects and white objects indicate AML samples (or normal tissues) ALL samples (or cancreous tissues), respectively. Training data and test data are expressed as a sphere and a cube, respectively.

Table 1. Number of classification errors in the MD-SVMs for the leukemia dataset. The columns 'n-th axis', $n = 1, 2, 3$, indicates the number of samples misclassified by n-th decision function. The columns 'n axes', $n = 1, 2, 3$, indicates the number of samples misclassified by n decision functions

Kernel	Sample	# of samples	1st axis	2nd axis	3rd axis	1 axis	2 axes	3 axes
Linear	Training	62	0	1	2	1	1	0
RBF	Training	62	0	2	7	5	2	0
Linear	Test	10	1	1	2	2	1	0
RBF	Test	10	0	2	0	2	0	0

Table 2. Number of classification errors in the MD-SVMs on the lung dataset. See the caption of Table 1 for other explanation

Kernel	Sample	# of samples	1st axis	2nd axis	3rd axis	1 axis	2 axes	3 axes
Linear	Training	170	0	1	1	0	1	0
RBF	Training	170	0	3	5	2	3	0
Linear	Test	33	1	0	0	1	0	0
RBF	Test	33	1	1	1	0	0	1

orthogonal axes, but it cannot predict classes of samples without other classification algorithms. We have tried to cover the shortcomings of both methods. MD-SVMs choose multiple orthogonal axes, which correspond to decision functions, from high dimensional space based on a margin between two classes. These multiple axes can be used for both visualization and class prediction.

Numerical experiments on real gene expression data indicate the effectiveness of MD-SVMs. All axes generated by MD-SVMs are taken into account for separating class of samples, while the 2nd and the 3rd axes by PCA are not. The samples in the distributions by MD-SVMs gather into appropriate clusters more vividly than those by PCA. MD-SVMs can predict the classes of the samples with multiple decision functions. We also indicate that MD-SVMs are useful for outlier detection with multiple decision functions.

There are several future works to be done on MD-SVMs: (1) application of our method to wider variety of gene expression datasets, (2) investigation of gene selection for preprocess of analysis and (3) investigation on class prediction method with multiple decision functions. Firstly, the use of more suitable samples may show that the axes chosen by MD-SVMs separate samples more clearly than those by PCA. Secondly, since the conventional SVMs show good generalization performance especially with large number of features, it is expected that MD-SVMs show much better performance than PCA with increasing the number of genes used in the numerical experiments. Since the element of weight vector generated by SVMs is one of the measures of discrimination power of the corresponding genes (Guyon *et al.*, 2002), that generated by MD-SVMs can be used for gene selection. Thirdly, the classification with probability as well as the weighted voting mentioned in Section 4 may be achieved in our scheme since the conventional SVMs have been already expanded for the purpose with sigmoid functions (Platt, 1999). We hope that our method sheds some lights on the future study of gene expression experiments.

REFERENCES

Bhattacharjee,A., Richards,W., Staunton,J., Li,C., Monti,S., Vasa,P., Ladd,C., Beheshti,J., Bueno,R., Gillette,M. *et al.* (2001) Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl Acad. Sci. USA*, **98**, 13790–13795.

Brown,M., Grundy,W., Lin,D., Cristianini,N., Sugnet,C., Furey,T., Ares,M. and Haussler,D. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl Acad. Sci. USA*, **97**, 262–267.

Cristianini,N. and Shawe-Taylor,J. (2000) *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods.* Cambridge University Press, NY.

Diamantaras,K. and Kung,S. (1996) *Principal Component Neural Networks Theory and Applications.* John Wiley & Sons, NY.

Fukunaga,K. (1990) *Introduction to Statistical Pattern Recognition.* Academic Press, NY.

Furey,T., Cristianini,N., Duffy,N., Bednarski,D., Schummer,M. and Haussler,D. (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, **16**, 906–914.

Golub,T., Slonim,D., Tamayo,P., Huard,C., Gaasenbeek,M., Mesirov,J., Coller,H., Loh,M., Downing,J., Caligiuri,M., Bloomfield,C. and Lander,E. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Guyon,I., Weston,J., Barnhill,S. and Vapnik,V. (2002) Gene selection for cancer classification using support vector machines. *J. Machine Learn.*, **46**, 389–422.

Huang,E., Ishida,S., Pittman,J., Dressman,H., Bild,A., Kloos,M., D'Amico,M., Pestell,R., West,M. and Nevins,J. (2003) Gene expression phenotypic models that predict the activity of oncogenic pathways. *Nat. Genet.*, **34**, 226–230.

Khan,J., Wei,J., Ringnér,M., Saal,L., Ladanyi,M., Westermann,F., Berthold,F., Schwab,M., Antonescu,C., Peterson,C. and Meltzer,P. (2001) Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat. Med.*, **7**, 673–679.

Platt,J. (1999) *Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods.* MIT Press, Cambridge, MA.

Schölkopf,B., Smola,A. and Müller,K. (1998) Non-linear component analysis as a kernel eigenvalue problem. *Neural Comput.*, **10**, 1299–1319.

Tibshirani,R., Hastie,T., Narasimhan,B. and Chu,G. (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl Acad. Sci. USA*, **99**, 6567–6572.

Vapnik,V. (1998) *Statistical Learning Theory.* John Wiley & Sons, NY.