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

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

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

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

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

Proof.

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

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

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

We consider the Lagrangian function

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

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

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

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

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

The solutions of this problem are

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

or

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

Therefore,

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

Table 1. Transcription factors linked to ischemia

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

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

Table 2. Information about 3 overlap blocks

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

Exploration with *CODM* allowed us to pick up 3 potentially important *overlap blocks*. This table shows the information for these 3 *overlap blocks*. The "# of UniGenes in cluster of TOL(/SHAM)" is the number of UniGenes which correspond to probes included in a cluster of TOL(/SHAM). The "# of common UniGenes (evaluation value)" is the number of common genes shared between the clusters of TOL and SHAM and its statistical evaluation value. The "similarity f (T, S)" is the similarity of the expression patterns between the clusters of TOL and SHAM. The range of similarity f (T, S) is −1(dissimilar) to 1(similar). The "Binding-sites of transcription factors" shows the name of putative binding-sites of transcription factors, the number of common genes that share the same binding-sites, and the statistical evaluation value of the number of common genes with the same binding-sites, if the evaluation value is 2.0 or higher.

Figures and Figure Legends

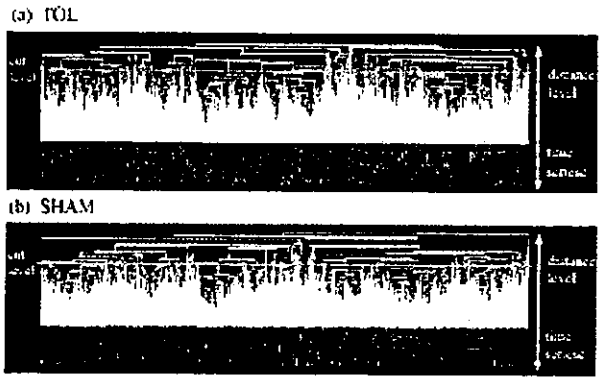

Figure 1. Hierarchical clustering of TOL and SHAM

We obtained time series ({0h, 1h, 3h, 12h, 24h, 48h} x 2) microarray data from rats with induced ischemic tolerance (*tolerant rats*: TOL) and rats with sham operation (sham rats: SHAM). In the analysis, we used these datasets as 12 time-points ({0a, 0b, 1a, 1b, 3a, 3b, ..., 48a, 48b} = $\{T_i\}$ (i = 1,2,...,12)) datasets on TOL and SHAM, respectively. After preprocessing and normalization, hierarchical clustering analysis based on Euclidian distances was then performed for each dataset independently.

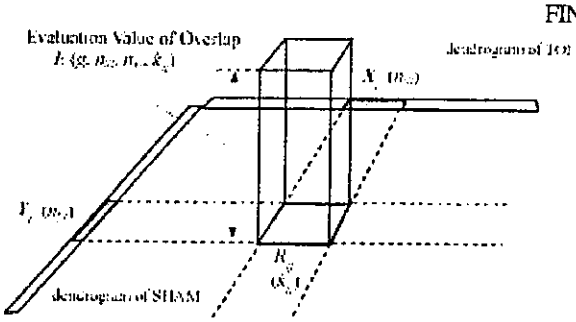

Figure 2. Overlap Block of Two Clusters

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

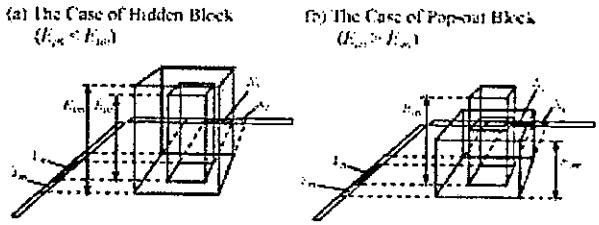

Figure 3. Relationships of Two Blocks

In *CODM*, all of the clusters are dealt with equally, regardless of their difference levels (i.e. their homogeneity). Even if they are included in other clusters, all of the statistical significance of the number of common genes between clusters is simultaneously visualized. Figure 3 shows that there is a risk that a small *overlap blocks* may be hidden in a large block. Assume that the clusters X_j and Y_n are included in X_i and Y_m respectively. Then, if the evaluation value E_{jn} is less than E_{im}, the small block B_{jn} will be hidden within the large block B_{im} (Figure 3a).

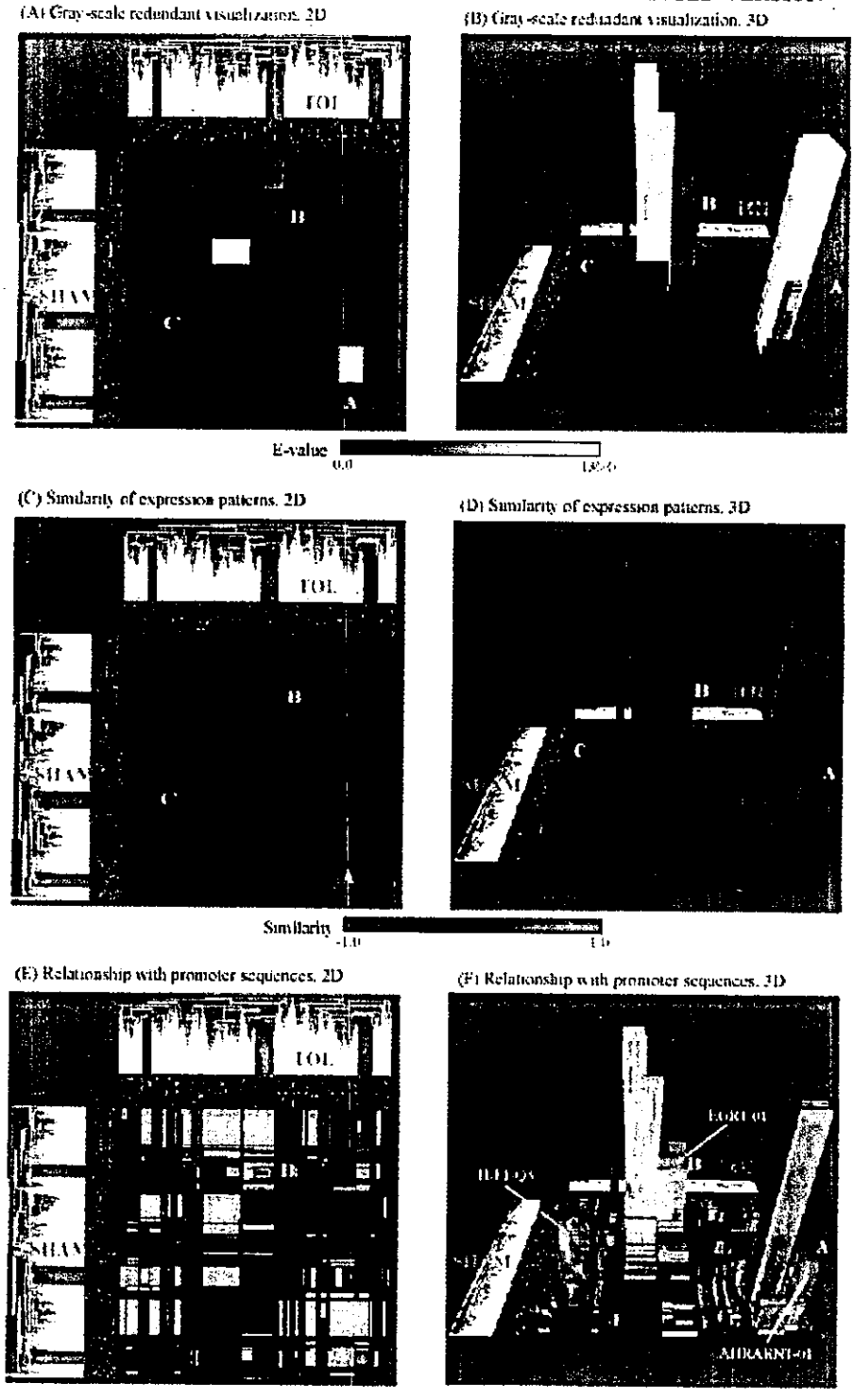

Figure 4. Visualizations for Comparison of Clustering Results of TOL and SHAM

This figure shows visualization results of the comparisons between TOL and SHAM in the mode of redundant visualization (Figures 4a and 4b), similarity of the expression patterns (Figures 4c and 4d), and the relationships with transcription factors (Figures 4e and 4f). In these figures, the *cut level* of the distance for hierarchical clustering was 0.74, and all of the *overlap blocks* with 2.0 or higher evaluation values are displayed as 3D histograms. As the figures show, the *CODM* provides not only a 3D mode (Figures 4b, 4d, and 4f) but also a 2D mode (Figures 4a, 4c, and 4e) where users can see a projected overhead view of the 3D mode.

In the mode showing the relationships with the transcription factors (Figures 4e and 4f), we considered the relationships with 8 types of transcription factors (HIF, ARNT and EGR families), which are known to mediate response to ischemia. In these figures, only *overlap blocks* with 2.0 or higher evaluation values of the number of genes with putative transcription factor binding sites were color-coded. Where an *overlap block* represents statistical significance for multiple transcription factors' putative binding sites, only the transcription factor with the highest evaluation value was visualized.

Exploration through changing the color-mode and the 2D&3D mode allowed us to pick up 3 potentially important *overlap blocks* which represented high evaluation values of the number of genes with the binding-sites ($E > 2.0$).

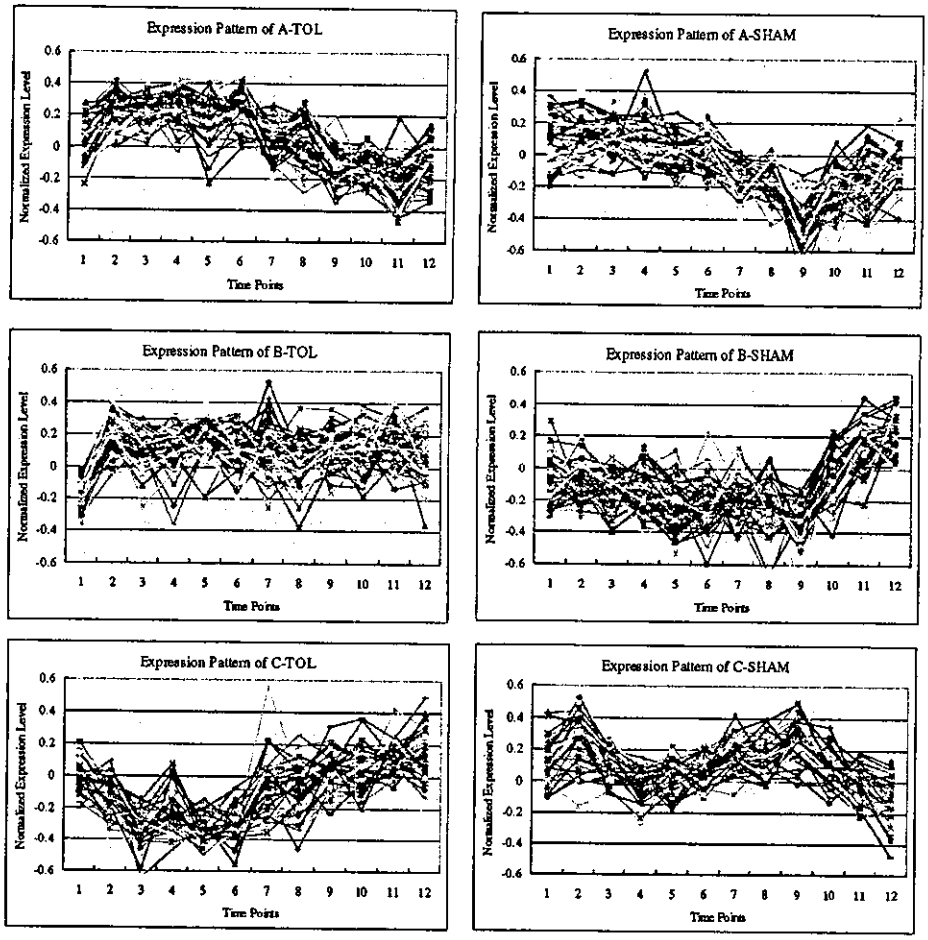

Figure 5. Expression Patterns of genes in the 3 overlap blocks

These figures show the expression patterns of common genes for the 3 *overlap blocks* which were picked up through exploration with *CODM* (Figure 4). The "Expression Patterns of Cluster T_i (/S_i)" (i = a,b,c) are the expression patterns of the common genes of the *overlap block i* in TOL(/SHAM).

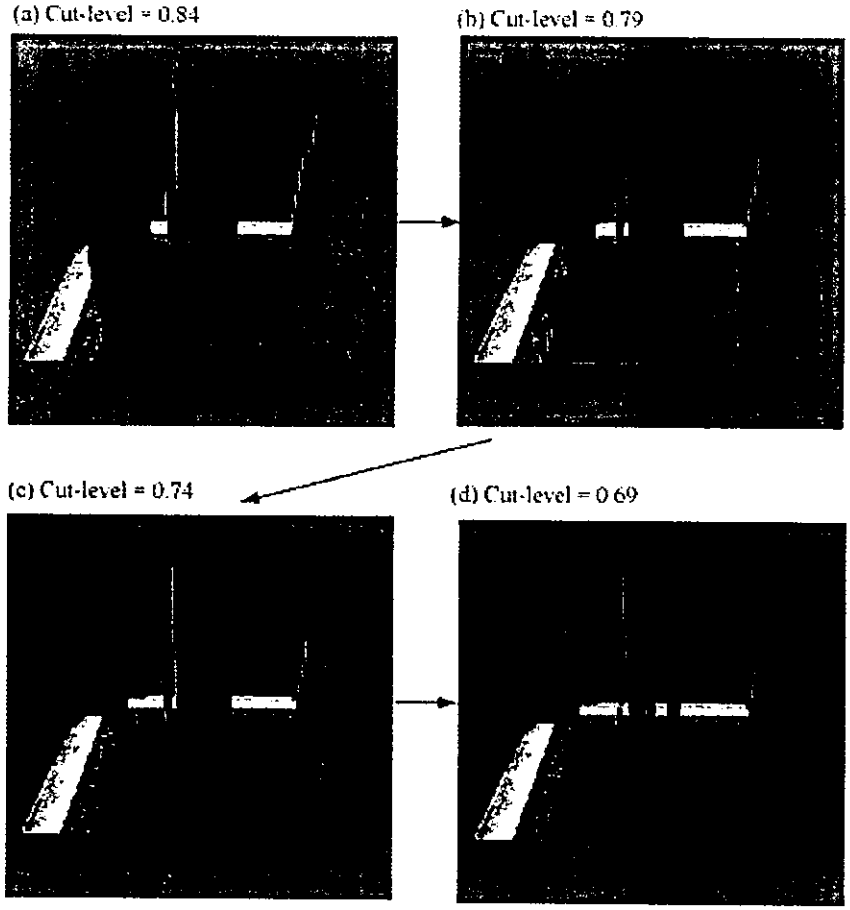

Figure 6. Interactive Changes of Cut-levels

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

ORIGINAL PAPER

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

Multidimensional support vector machines for visualization of gene expression data

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

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

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

ABSTRACT

Motivation: Since DNA microarray experiments provide us with huge amount of gene expression data, they should be analyzed with statistical methods to extract the meanings of experimental results. Some dimensionality reduction methods such as Principal Component Analysis (PCA) are used to roughly visualize the distribution of high dimensional gene expression data. However, in the case of binary classification of gene expression data, PCA does not utilize class information when choosing axes. Thus clearly separable data in the original space may not be so in the reduced space used in PCA.
Results: For visualization and class prediction of gene expression data, we have developed a new SVM-based method called multidimensional SVMs, that generate multiple orthogonal axes. This method projects high dimensional data into lower dimensional space to exhibit properties of the data clearly and to visualize a distribution of the data roughly. Furthermore, the multiple axes can be used for class prediction. The basic properties of conventional SVMs are retained in our method: solutions of mathematical programming are sparse, and nonlinear classification is implemented implicitly through the use of kernel functions. The application of our method to the experimentally obtained gene expression datasets for patients' samples indicates that our algorithm is efficient and useful for visualization and class prediction.
Contact: komura@hal.rcast.u-tokyo.ac.jp

1 INTRODUCTION

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

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

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

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

*To whom correspondence should be addressed.

Komura *et al.* (2004) Multidimensional Support Vector Machines for Visualization of Gene Expression Data. Symposium on Applied Computing, Proceedings of the 2004 ACM symposium on Applied computing, 175-179; http://doi.acm.org/10.1145/967900.967936

classes to minimize generalization errors. The axes generated by this method reduce dimensions of original data to extract information useful in estimating the discriminability of two classes. This method fulfills the requirement of both visualization and class prediction. The basic properties of SVMs are retained in our method: solutions of mathematical programming are sparse, and nonlinear classification of data is implemented implicitly through the use of kernel functions.

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

1.1 Notation

$\mathbb{R}$ is defined as the set of real numbers. Each component of a vector $\boldsymbol{x} \in \mathbb{R}^n, i = 1,\ldots,m$ 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

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

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

with $C > 0$. Here, $\boldsymbol{\xi} \in \mathbb{R}^m$ 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

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

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

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

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

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

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

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 $\alpha^k, \gamma^k \in \mathbf{R}^m$, $\beta^k \in \mathbf{R}^{k-1}$ 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 $x \in \mathbf{R}^m$ 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 $K(x_i, x_j) = x_i \cdot x_j$ and RBF kernel $K(x_i, x_j) = \exp -\gamma \| x_i - x_j \|^2$ with $\gamma = 0.001$ in the experiments of MD-SVMs.

4.2 Materials

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

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

5 RESULTS AND DISCUSSION

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

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

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

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

6 CONCLUSION

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

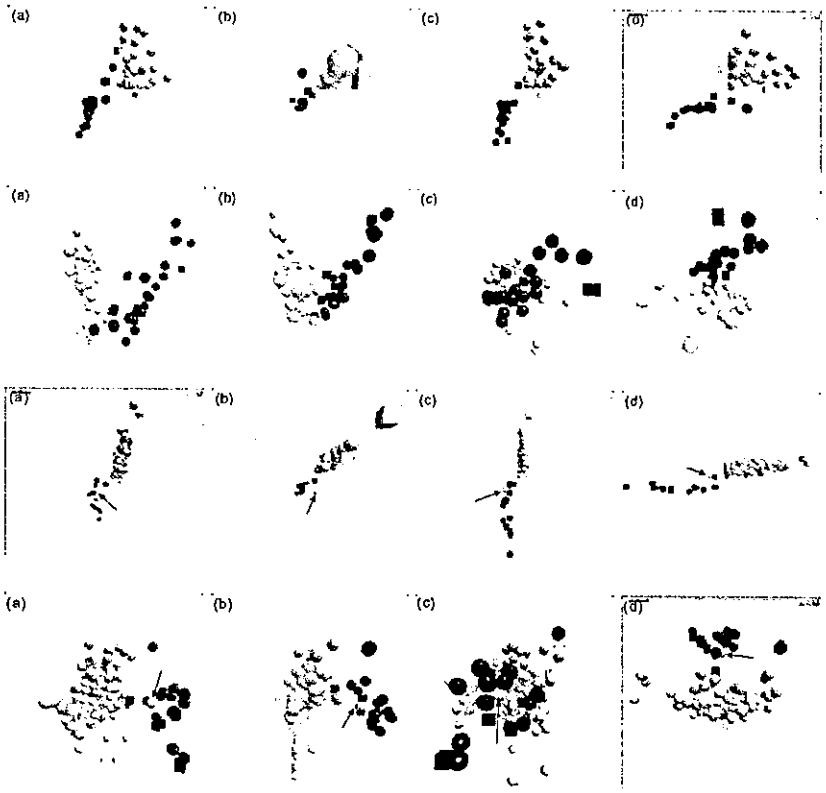

Fig. 1. (Top row) Distribution obtained by MD-SVMs for the leukemia dataset with linear kernel. (Second row) Distribution obtained by PCA on the leukemia dataset. (Third row) Distribution obtained by MD-SVMs for the lung tissues dataset with linear kernel. The sample indicated by arrows appears to be 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.

Carcinogenesis vol.25 no.8 pp.1495–1505, 2004
doi:10.1093/carcin/bgh155

Global gene expression analysis of rat colon cancers induced by a food-borne carcinogen, 2-amino-1-methyl-6-phenylimidazo[4,5-*b*]pyridine

Kyoko Fujiwara, Masako Ochiai, Tsutomu Ohta[1], Misao Ohki[1], Hiroyuki Aburatani[2], Minako Nagao, Takashi Sugimura and Hitoshi Nakagama[3]

Biochemistry Division and [1]Medical Genetics Division, National Cancer Center Research Institute, 5-1-1 Tsukiji, Chuo-ku, Tokyo 104-0045, Japan and [2]Research Center for Advanced Science and Technology, The University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo 153-8904, Japan

[3]To whom correspondence should be addressed. Tel: +81 3 3547 5239; Fax: +81 3 3542-2530;
Email: hnakagam@gan2.res.ncc.go.jp

Colon cancers develop after accumulation of multiple genetic and epigenetic alterations in colon epithelial cells. To shed light on global changes in gene expression of colon cancers and to gain further insight into the molecular mechanisms underlying colon carcinogenesis, we have conducted a comprehensive microarray analysis of mRNA using a rat colon cancer model with the food-borne carcinogen, 2-amino-1-methyl-6-phenylimidazo[4,5-*b*]pyridine (PhIP). Of 8749 genes or ESTs on a high density oligonucleotide microarray, 27 and 46 were over- and underexpressed, respectively, by ≥3-fold in colon cancers in common in two rat strains with distinct susceptibility to PhIP carcinogenesis. For example, genes involved in inflammation and matrix proteases and a cell cycle regulator gene, *cyclin D2*, were highly expressed in colon cancers. In contrast, genes encoding structural proteins, muscle-related proteins, matrix-composing and mucin-like proteins were underexpressed. Interestingly, a subset of genes whose expression is characteristic of Paneth cells, i.e. the *defensins* and *matrilysin*, were highly overexpressed in colon cancers. The presence of *defensin 3* and *defensin 5* transcripts in cancer cells could also be confirmed by *in situ* mRNA hybridization. Furthermore, Alcian blue/periodic acid Schiff base (AB-PAS) staining and immunohistochemical analysis with an anti-lysozyme antibody demonstrated Paneth cells in the cancer tissues. AB-PAS-positive cells were also observed in high grade dysplastic aberrant crypt foci, which are considered to be preneoplastic lesions of the colon. Our results suggest that Paneth cell differentiation in colon epithelial cells could be an early morphological change in cryptic cells during colon carcinogenesis.

Introduction

The development of colon cancers comprises multiple steps requiring the accumulation of genetic and epigenetic alterations in colon epithelial cells, and these changes further affect expression of a variety of downstream genes and may cause considerable changes in gene expression profiles in cancer cells as a consequence. Inactivation of the *adenomatous polyposis coli* (*APC*) gene, β-*catenin*, K-*RAS*, *SMAD2*, SMAD4, *p53* and mismatch repair genes by genetic alterations, for example, play key roles (1,2). Furthermore, alterations of gene expression profiles by perturbation of CpG island methylation in promoter regions and/or the histone acetylation/deacetylation status of chromatin also have a substantial impact on colon carcinogenesis (2).

Oral administration of 2-amino-1-methyl-6-phenylimidazo[4,5-*b*]pyridine (PhIP), one of the most abundant heterocyclic amines produced while cooking meat and fish (3,4), induces aberrant crypt foci (ACF) (5,6), putative preneoplastic lesions of the colon (7,8), in experimental animals within a short period and colon adenomas and adenocarcinomas after 1 or 2 years in rats, preferentially in males (9). A number of studies have revealed that PhIP-induced rat colon cancers resemble human neoplasms with regard to observed histological features and genetic alterations (10–14). There are several advantages with the use of animal cancer models to dissect the molecular basis of colon carcinogenesis. For example, inbred experimental animals share a common genetic background within the strain and, furthermore, carcinogenesis experiments using these animals can be carried out under well-controlled conditions. Genetic and/or epigenetic alterations in colon cancers induced in experimental animals are therefore expected to be more uniform compared with those in humans with diverse genetic backgrounds. Colon cancers induced by PhIP indeed demonstrate β-catenin accumulation in both cytoplasm and nucleus (13) and β-*catenin* mutations are observed at codons 32, 34, 36 or 38 in exon 2, the majority being G → T transversions (12,13). In the *Apc* gene, 5′-GGGA-3′ sites in exons 14 and 15 and a 5′-agGGGGG-3′ site at the junction of intron 10 and exon 11 are mutation hot-spots (10,13). Using a model system, we have recently revealed sequential progression from dysplastic ACF to colon cancer (14,15). Although the PhIP-induced rat colon cancer model has provided cancer researchers with a powerful tool for dissecting molecular events involved in the formation of colon cancers with relevance to human colon carcinogenesis, extensive studies aimed at the elucidation of early genetic events in colon cancer development have hitherto not been conducted.

In the present study we therefore performed a global gene expression analysis of rat colon cancers induced by PhIP using high density oligonucleotide microarrays (GeneChip; Affymetrix, Santa Clara, CA). To eliminate detection of strain-specific changes, but rather to detect specific gene expression profiles essential for colon cancer development, two rat strains, F344 and ACI, were subjected to analysis, the former being the more susceptible to PhIP-induced colon carcinogenesis. A considerable number of genes were found to be differentially expressed in colon cancers compared with normal counterpart epithelium, including examples characteristic of Paneth cells. Global

Abbreviations: AB-PAS, Alcian blue/periodic acid Schiff base; ACF, aberrant crypt foci; DIG, digoxigenin; APC, adenomatous polyposis coli; EST, expressed sequence tagged; H&E, hematoxylin and eosin; PhIP, 2-amino-1-methyl-6-phenylimidazo[4,5-*b*]pyridine.

changes in gene expression profiles are also discussed in comparison with those reported in human colon cancers. Another focus is on the appearance of Paneth cells in ACF, especially in dysplastic ones, and its biological significance.

Materials and methods

Animals and diets

PhIP was purchased from the Nard Institute (Osaka, Japan) in the form of PhIP-HCl and added to AIN-93G basal diet (7% w/w soybean oil; Dyets, Bethlehem, PA) at a concentration of 400 p.p.m. A high fat diet (AIN-93G basal diet supplemented with 23% w/w hydrogenated vegetable oil) was also purchased from Dyets. Five-week-old male F344 and ACI strain rats were purchased from CLEA Japan (Tokyo, Japan) and housed 3 per cage in an air-conditioned animal room with a 12 h light/dark cycle. Prior to the experiment, all the animals were acclimatized to the housing environment and the AIN-93G basal diet for 1 week.

Experimental protocol and tissue samples

Starting at the age of 6 weeks, rats were fed a diet containing PhIP following an intermittent PhIP feeding protocol (13). At experimental week 60 all animals were killed and colons were removed. When colon cancers with polypoid growth were detected by the naked eye, cancerous parts were resected with a razor blade, bisected and one half was embedded in O.C.T. compound (Tissue-Tek; Sakura Finetechnical Co., Tokyo, Japan), frozen and stored at −80°C until use for frozen section preparation and RNA extraction. The remaining halves were fixed in neutral 10% formalin overnight at 4°C and embedded in paraffin blocks according to standard procedures. Normal counterparts were collected from the surrounding normal parts of the colon and separately embedded in O.C.T. compound and samples were snap-frozen in liquid nitrogen and stored at −80°C until use for RNA extraction. In separate experiments using the intermittent PhIP feeding protocol, ACF were assayed at experimental weeks 18 and 25, after fixation of tissue in formalin and embedding in paraffin blocks as described above.

High density oligonucleotide microarray analysis

Twelve colon cancer tissues, six each from F344 and ACI rats, and 12 normal counterparts were collected by digging them out of frozen O.C.T. blocks using 18 gauge needles. Total RNA was extracted from ~1 mg of tissue with TRIZOL reagent (Invitrogen, Carlsbad, CA). Two of six colon cancer tissues from F344 rats, however, did not provide sufficient amounts of good quality RNA. The remaining four samples from F344 and six from ACI rats were subjected to the following experiments. cRNA was synthesized, labeled with biotin and hybridized to high density oligonucleotide microarrays, Rat Genome U34A (RG U34A; Affymetrix), as described previously. The average hybridization intensity for each array was scaled to 1000 to reliably compare multiple arrays. Prior to statistical analysis, genes were filtered according to the following criteria. For genes overexpressed in cancers, for example, they should have 'present (P)' or 'marginal (M)' calls in at least half of the colon cancer samples of the respective rat strains. For genes underexpressed in cancers, in contrast, they should have P or M calls at least in half of the normal counterpart samples. To assess statistical differences in gene expression between colon cancers and normal tissues, average signal intensity and standard variation were calculated for each group and GeneSpring 4.3 (Silicon Genetics, Redwood City, CA) was employed for the Mann–Whitney *U*-test. The significant *P* value was set at 0.05. Then, genes which were differentially expressed between cancer and normal tissue at ≥3-fold were selected and subjected to further analysis, including Venn diagrams, hierarchical clustering analysis, functional classification and comparison with expression profiles of human colon cancers. Permutation analysis was also carried out to assess the statistical significance of genes differentially expressed between the two rat strains.

Histological analysis

For hematoxylin and eosin (H&E) staining, paraffin sections were prepared at 3.5 μm thickness following standard procedures. Histological evaluation of colonic lesions was performed as described previously (13). For Alcian blue (pH 2.5)/periodic acid Schiff base (AB-PAS) staining to evaluate the presence of Paneth cells, both frozen (10 μm thickness) and paraffin (3.5 μm thickness) sections were used. The staining was carried out according to conventional methods.

In situ mRNA hybridiztion for defensin genes

In situ mRNA hybridization was carried out as described previously (16,17) under contract by Genostaff (Tokyo, Japan) using frozen sections prepared at 10 μm thickness. A 293 bp cDNA fragment of the rat neutrophil *defensin 3* gene was amplified by PCR with primers 5′-CTCCCTGCATACGCCAAAG-3′ (forward) and 5′-AACAGAGTCGGTAGATGCG-3′ (reverse) and a 335 bp cDNA fragment of the *defensin 5* gene with primers 5′-AACTTGTCCTCCTT TCTGCC-3′ (forward) and 5′-AACATCAGCATCGGTGGCC-3′ (reverse). Amplified fragments were cloned into pCRII (Invitrogen) and digoxigenin (DIG)-labeled RNA probes were generated by an *in vitro* transcription method using DIG-labeling mix (Roche Molecular Biochemicals, Tokyo, Japan). Hybridized probes were detected by an IgG antibody against the DIG label and visualized with NBT/BCIP solution (Roche Molecular Biochemicals). Nuclear counterstaining was performed with Kernechtrot Stain Sol (Muto Chemical, Tokyo, Japan).

Semi-quantitative RT–PCR

Extracted RNA was transcribed to cDNA using an oligo$(dT)_{12-18}$ primer and SuperScript™ II reverse transcriptase (Invitrogen) and the cDNAs produced were divided into aliquots in tubes and stored at −20°C until analyzed. Each aliquot of cDNA was subjected to semi-quantitative reverse transcription (RT)–PCR with the primer sequences listed in Table I. A set of semi-quantitative RT–PCR reactions for representative genes was carried out within 1 day to avoid the effects of degradation of cDNA templates. For reference, expression of the β-*actin* and *glyceraldehyde 3-phosphate dehydrogenase* (*G3PDH*) genes was also quantified for each sample. PCR amplification was carried out at 94°C for 30 s, 60°C for 30 s and 72°C for 1 min using Advantage Taq (Clontech, Palo Alto, CA) under the conditions recommended by the manufacturer. PCR cycles were set at 25 for β-*actin* and *G3PDH*, 35 for α-*defensin NP4* and β-*defensin* 2 and 30 cycles for the other genes. PCR products were also analyzed by gel electrophoresis on a 2% agarose gel in 0.5× TBE (89 mM Tris, 89 mM boric acid, 1.9 mM EDTA). The amounts of PCR products were quantified by analysis performed on a Macintosh iBook G3 computer using the public domain NIH Image program (developed at the US National Institutes of Health and available on the Internet by anonymous ftp from zippy.nimh.nih.gov. or on floppy disk from the National Technical Information Service, Springfield, VA, part no. PB95-500195GEI). PCR reactions for individual genes were

Table I. List of primers used for RT–PCR

Gene name	Forward primer	Reverse primer
Matrilysin	5′-TTCGCAAGGGGAGATCACG-3′	5′-AACAGAAGAGTGACCCAGAC-3′
Mash2	5′-TTACCCATGCTGTCTAGTGC-3′	5′-AGTCCTCCAGCAGTTCAAGT-3′
Oct1A	5′-CCTTCATCATCCTGGTCAC-3′	5′-ATGAAGGGGGTGAAGATCC-3′
Carbonic anhydrase IV	5′-GGTAAACGAGGGCTTCCAG-3′	5′-TGAGACCTGAACACCTGGC-3′
AA799832	5′-GCGATCATGCCTTGCTTAAC-3′	5′-TTCCAGCGGCAGATGAAGG-3′
Defensin NP1 like	5′-TGCTGTTCAAGATTTACGCG-3′	5′-ACCTTGATAGCCGAATGCAGC-3′
Defensin NP3	5′-CTCCCTGCATACGCCAAAG-3′	5′-AACAGAGTCGGTAGATGCG-3′
Defensin α5	5′-AACTTGTCCTCCTTTCTGCC-3′	5′-AACATCAGCATCGGTGGCC-3′
Defensin NP4	5′-GACACTCACTCTGCTCATCA-3′	5′-ATGACAAATGGCTTCTTCTC-3′
Defensin β1	5′-CTTGGACGCAGAACAGATCA-3′	5′-AAACCACTGTCAACTCCTGC-3′
β-Actin	5′-GACTTCGAGCAAGAGATGGC-3′	5′-AGGAAGGAAGGCTGGAAGAG-3′
G3PDH	5′-TCATGACCACAGTCCATGCC-3′	5′-CTCAGTGTAGCCCAGGATGC-3′