BMC Bioinformatics

Research article

Classification of heterogeneous microarray data by maximum entropy kernel

Wataru Fujibuchi*†1 and Tsuyoshi Kato†2,1

Address: [1]National Institute of Advanced Industrial Science and Technology (AIST), Computational Biology Research Center, 2-42 Aomi, Koto-ku, Tokyo 135-0064, Japan and [2]Graduate School of Frontier Sciences, University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8562, Japan

Email: Wataru Fujibuchi* - w.fujibuchi@aist.go.jp; Tsuyoshi Kato - kato-tsuyoshi@cb.k.u-tokyo.ac.jp

* Corresponding author †Equal contributors

Published: 26 July 2007

BMC Bioinformatics 2007, **8**:267 doi:10.1186/1471-2105-8-267

Received: 16 January 2007
Accepted: 26 July 2007

This article is available from: http://www.biomedcentral.com/1471-2105/8/267

Abstract

Background: There is a large amount of microarray data accumulating in public databases, providing various data waiting to be analyzed jointly. Powerful kernel-based methods are commonly used in microarray analyses with support vector machines (SVMs) to approach a wide range of classification problems. However, the standard vectorial data kernel family (linear, RBF, etc.) that takes vectorial data as input, often fails in prediction if the data come from different platforms or laboratories, due to the low gene overlaps or consistencies between the different datasets.

Results: We introduce a new type of kernel called maximum entropy (ME) kernel, which has no pre-defined function but is generated by kernel entropy maximization with sample distance matrices as constraints, into the field of SVM classification of microarray data. We assessed the performance of the ME kernel with three different data: heterogeneous kidney carcinoma, noise-introduced leukemia, and heterogeneous oral cavity carcinoma metastasis data. The results clearly show that the ME kernel is very robust for heterogeneous data containing missing values and high-noise, and gives higher prediction accuracies than the standard kernels, namely, linear, polynomial and RBF.

Conclusion: The results demonstrate its utility in effectively analyzing promiscuous microarray data of rare specimens, e.g., minor diseases or species, that present difficulty in compiling homogeneous data in a single laboratory.

Background

Microarray has become a standard tool in many biological studies. Typically, classification analyses, where gene expressions of distinct biological groups are compared and classified according to their gene expression characteristics, are frequently performed in various clinical situations such as tumor diagnosis [1,2], anti-cancer drug response analysis [3,4], and prognosis analysis [5,6]. Kernel methods [7] play important roles in such disease analyses, especially when classifying data with support vector machines (SVMs) [8] based on the feature or marker genes that are correlated with the characteristics of the groups. In most of those studies, only standard kernels such as linear, polynomial, and RBF (radial basis function), which take vectorial data as input, have been popularly used and generally successful.

Other than the above *vectorial data kernel* family, there is another family called *structured data kernel* family that has been studied in many other fields including bioinformat-

ics and machine learning [9-12]. The structured data kernel family conveys structural or topological information with or without numerical data as input to describe data. For example, the string kernel for text classification [9], the marginalized count kernel [10] for biological sequences, the diffusion kernel [11] and the maximum entropy (ME) kernel [12] for graph structures are well known in the biological field.

In microarray analysis, one of the main issues that hamper accurate and realistic predictions is the lack of repeat experiments, often due to financial problems or rarity of specimens such as minor diseases. Utilization of public or old data together with one's current data could solve this problem; many studies combining several microarray datasets have been performed [13-15]. However, due to the low gene overlaps and consistencies between different datasets, the vectorial data kernels are often unsuccessful in classifying data from various datasets if naïvely integrated [14].

Among the structured data kernels, the ME kernel can take any distance data as input, and is thus applicable to vectorial data as well when converted into the Euclidean or other types of distance relationships among vectors. Since the ME kernel increases the distances among different sample vectors (or samples hereafter), while keeping similar samples in close distance, discriminative boundaries may be found more explicitly than in the case of the vectorial kernels due to the sparse distribution of heterogeneous samples (Figure 1). Furthermore, the ME kernel has, unlike the RBF kernel, a special property of excluding arbitrary gene values composing vectorial data in calculating the distances among samples. Hence, by ignoring only spurious gene values in each sample without deleting those genes entirely from a dataset, the ME kernel can effectively utilize gene expression information in heterogeneous data containing mosaic-like missing or noisy values.

This paper is constructed as follows. We first show how the ME kernel can effectively work in heterogeneous microarray data using the Euclidean distance among sample vectors. Then, we show the unique and powerful noise reduction ability of the ME kernel in microarray data. Finally, we demonstrate that the ME kernel performs better than the standard kernels in classifying practical microarray data, namely, squamous cell carcinoma metastasis in the oral cavity.

Results

We describe herein the classification performance of the ME kernel, compared to that of the three standard kernels, linear, polynomial and RBF. We also test two types of distance-based kernels, EKM and Saigo [16,17], for comparison. The schematic view of the entire analysis process is shown in Figure 2. Note that the RBF kernel also uses Euclidean distance as the metric of sample (dis-) similarities but cannot use the *k*-nearest neighbor gene distance (*k*NND) since it violates the positive semidefiniteness of kernels.

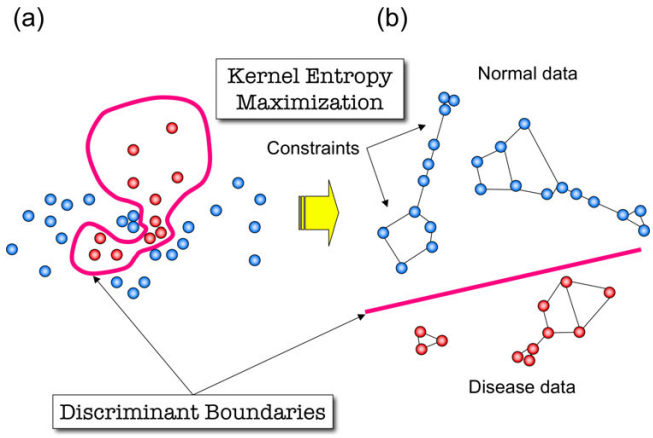

Figure 1
Maximum entropy kernel for heterogeneous data. Samples and their distance constraints in the feature space are drawn schematically as graph nodes and edges, respectively. (a) The heterogeneous data are entangled in the feature space, making it difficult to find the discriminant boundary. (b) After kernel entropy maximization, the distances among samples are expanded in the feature space under constraints that hold only similar samples closely, making it easier to find the discriminant boundary.

We first use *heterogeneous* kidney carcinoma data to confirm the ME kernel's superior discrimination ability against a highly mixed heterogeneous dataset. Then, we demonstrate the ME kernel's interesting denoising ability based on *k*NND using *homogeneous* leukemia microarray data with artificial noise. Finally, we further apply the ME kernel with *k*NND denoising to a more practical problem, i.e., *heterogeneous* data of squamous cell carcinoma metastasis in the oral cavity, to assess its total performance.

Data normalization and classification analysis

Before testing the performance, all the data are properly normalized by being first log-transformed, and then scaled to mean 0 and standard deviation 1 (i.e., Z-normalization) in each sample and then each gene. All the normalized datasets are available for free at our Internet server [18]. Also the ME program that runs on Linux OS is available upon request. Many genes have a large number of missing values because heterogeneous data are combined; thus, we adopt a simple imputation method that all the missing values are replaced with the mean value, i.e., 0. Input genes that show high correlation to class

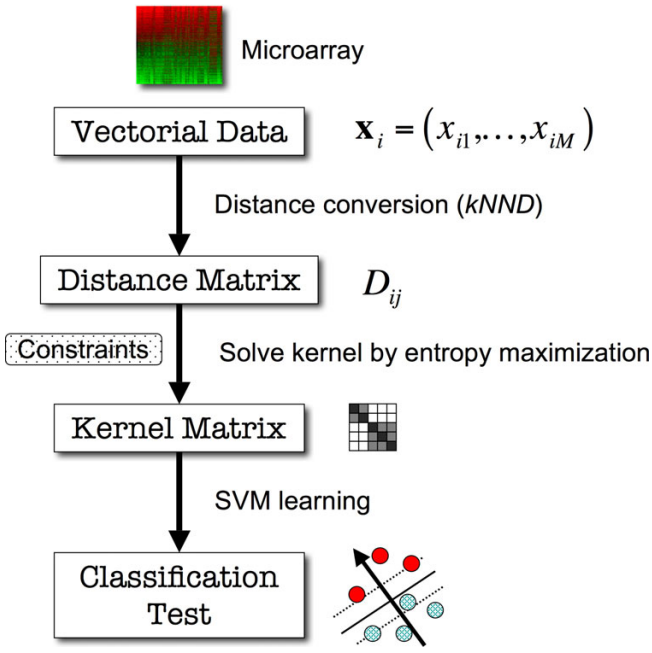

Figure 2
Schematic view of the entire process of microarray classification in the ME kernel algorithm. The input vectorial data are first converted into distance matrix to provide constraints D_{ij}. Then, entropy of a kernel matrix is maximized under the constraints, generating an optimal kernel matrix that is guaranteed to be positive semidefinite. Then, the SVM learns the classification boundary from the kernel matrix and classifies test samples.

labels, or feature genes, are selected by the standard two sample t-statistics [19] in each iteration of the leave-one-out cross-validation (LOOCV) test. The distance constraint matrices (D_{ij}) are also generated from the same feature genes. If a sample contains missing values, we again adopt a simple imputation; we replace the one-dimensional Euclidean distance $(x_{ih}- x_{jh})^2$ with 2 if x_{ih} or x_{jh} is missing. The six kernels are tested with SVMs to analyze their classification performance with various numbers of feature genes and various parameters described in Methods. The maximum accuracy among the tested parameters for each number of feature genes is recorded as the accuracy for each kernel.

Heterogeneous kidney carcinoma data

The human kidney data of normal tissues and renal clear carcinoma tissues are collected from the public gene expression database, GEO-Gene Expression Omnibus [20]. This dataset is comprised of ten platforms, two of which are spotted DNA/cDNA arrays and eight are variations of Affymetrix-type oligonucleotide arrays. To uniformly analyze the array data from different platforms, we converted as many probe names as possible to UniGene identifiers and combined all the data. The total number of UniGenes in the integrated table is as large as 54,674, all of which contain missing values in some platforms; i.e., there are no genes common to all platforms. The total number of normal and carcinoma data is 100 (62 normal and 38 carcinoma). The characteristics of each data in the composite dataset, such as platform ID, array type, number of data, and experimental comments, are shown in Table 1.

Classification analysis is performed between normal and carcinoma data. The results of the LOOCV test of 100 samples against various numbers (8–296; increasing 8 genes at each step due to computational limitations) of feature genes are plotted in Figure 3. The figure shows typical prediction curves, namely, accuracy increases with increasing number of feature genes, plateaus at some region, and decreases. Clearly, the ME kernel performs much better in all cases than the other five kernels for small numbers of feature genes (8–192). As regards accuracy, the ME kernel records maximum accuracies as high as 95.0 (89.5/98.4 sensitivity/specificity)% for 152 of feature genes. Statistically, the accuracies of the ME kernel are superior to those of the other five kernels in 64.9% of the tested points (8–296) of feature genes. This percentage increases to 95.8% when accuracies are limited only to the increase and plateau regions (8–192) of the ME kernel.

kNND denoising for AML and ALL data

Acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) data for cancer subtype classification have been reported by Golub *et al*. [1]. There are 72 samples (47 AML and 25 ALL), all of which are quite homogeneous and of good quality, and are thus suitable for artificial noise experiments. To assess the denoising ability of our ME kernel, we first replace the $\nu_{\text{add}} \times 100\%$ of original data in a gene expression profile with artificial white noise, i.e., the noise is added according to a normal distribution model with a mean of 0 and a standard deviation of twice that of each gene value distribution in the original dataset. Then, we extract 50 feature genes from the training dataset for each iteration of the LOOCV test by standard t-statistics.

As the control experiments using linear and RBF kernels, the standard singular value decomposition (SVD) denoising method is applied to reduce noise immediately after the noise is introduced. In the SVD denoising, three levels of noise removals by dfferent cumulative proportions, 85, 90, and 95%, of eigenvalues are explored. For the ME kernel, the *k*NND denoising method with the following noise level settings is applied. First, raw noise that is assumed to internally exist in the original data is arbitrarily set at $\nu_{\text{raw}} = 0.05$. Then, we define the total noise level as the sum of the raw noise and the above artificially

Table 1: Organization of heterogeneous kidney carcinoma dataset

Platform	Array type	#Normal/Carcinoma	Brief comments
GPL9	Spotted, DNA/cDNA	10/10	Renal clear cell carcinoma, primary tumor [30]
GPL10	Spotted, DNA/cDNA	10/10	Renal clear cell carcinoma, primary tumor [30]
GPL91	Affymetrix, oligo	14/0	Large-scale analysis of the human tran-scriptome (HG-U95A) kidney [31]; Normal human tissue expression profiling (HG-U95A) kidney [32]; Kidney transplant rejection expression profiling kidney normal donor [33]
GPL96	Affymetrix, oligo	10/9	Large-scale analysis of the human tran-scriptome (HG-U133A) kidney [34]; Renal clear cell carcinoma (HG-U133A) [35]
GPL97	Affymetrix, oligo	8/9	Renal clear cell carcinoma (HG-U133B) [35]
GPL92	Affymetrix, oligo	2/0	Normal human tissue expression profiling (HG-U95B) kidney [32]
GPL93	Affymetrix, oligo	2/0	Normal human tissue expression profiling (HG-U95C) kidney [32]
GPL94	Affymetrix, oligo	2/0	Normal human tissue expression profiling (HG-U95D) kidney [32]
GPL95	Affymetrix, oligo	2/0	Normal human tissue expression profiling (HG-U95E) kidney [32]
GPL 1074	Affymetrix, oligo	2/0	Large-scale analysis of the human tran-scriptome (GNFlb) kidney [34]

Platform IDs, array types, number of data, and experimental comments are shown.

added noise, v_{add}. For example, if 10% noise is added, the total noise level is $v_{raw} + v_{add} = 0.05 + 0.1 = 0.15$, and $(1 - 0.15)^2 \times 100 = 72.3\%$ of the nearest distance genes out of the feature gene set are considered in calculating the *k*NNDs between samples (see Methods).

We repeat the above random noise-adding test ten times and average the highest accuracies among various parameter combinations. The results are shown in Figure 4. The artificial noise added is within the range of 0–50%. Since the raw data are quite homogeneous, all kernels except linear and polynomial show the same prediction accuracy of 98.6% when no noise is added. This value decreases gradually with increasing noise levels (10–50%) for the

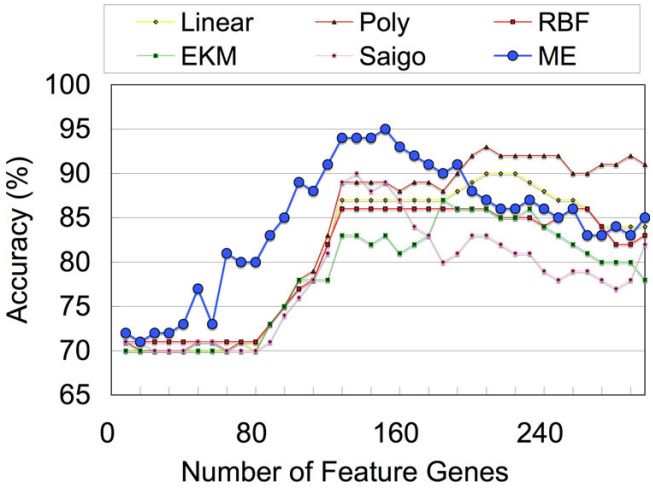

Figure 3
Classifications of heterogeneous renal carcinoma data with standard and ME kernels. In most cases, the ME kernel shows much better performance than the linear, polynomial, and RBF kernels and the two distance-based kernels for various numbers of feature genes.

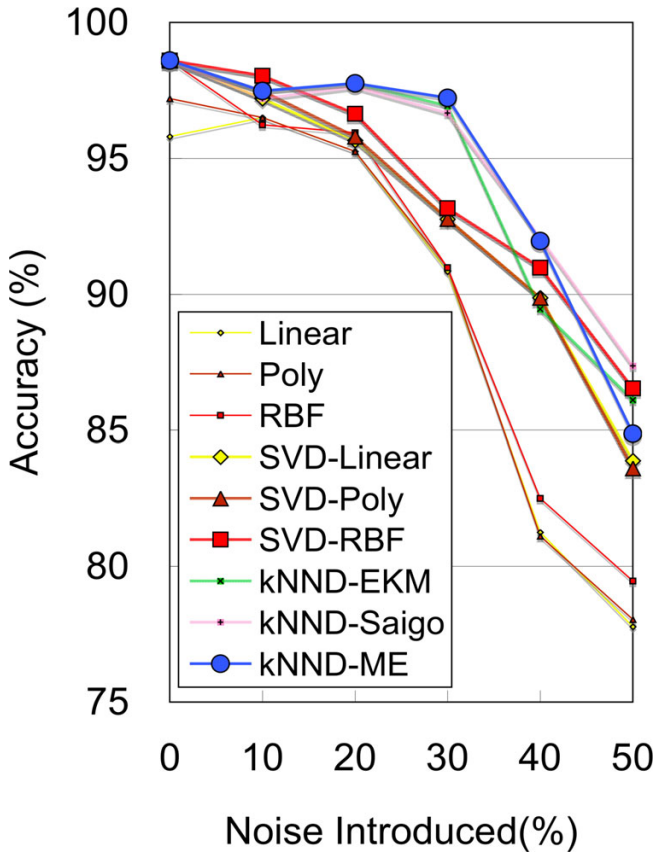

Figure 4
AML/ALL classification with artificial noise. The accuracies of standard linear and RBF kernels decrease with increasing noise levels, even with SVD denoising applied, while those of ME and other distance-based kernels with *k*NND denoising are sustained at high levels at 10–40% noise levels

vectorial kernels; for example, the accuracies of the RBF kernel decrease in the order of 96.2, 95.9, 91.0, 82.5, and 79.5%. SVD denoising boosts up these accuracies to 98.0, 96.6, 93.2, 91.0, and 86.5%, respectively. The linear and polynomial kernels also show similar accuracies to the RBF kernel when SVD denoising is used.

Interestingly and surprisingly, the three *k*NND-distance-based methods show high accuracies; for example, the *k*NND-ME kernel has an accuracy of 97.8% even at 20% noise level and maintains high accuracies of 97.2 and 92.0% at 30–40% noise levels. The EKM and Saigo kernels using *k*NND-distance also show similar accuracies to the *k*NND-ME kernel. To verify our results, we extensively analyzed the same data with various parameters including many cumulative proportions in the SVD but obtained similar tendencies, confirming the superior denoising ability of the *k*NND-based method [21].

Heterogeneous oral cavity carcinoma metastasis data

We further analyze the total performance of the *k*NND-ME method with a more practical problem-heterogeneous oral cavity carcinoma metastasis data. The data consist of two GEO datasets (GSE2280 and GSE3524) from different authors [22,23]. One dataset (GSE2280) is derived from primary squamous cell carcinoma dataset of the oral cavity [22], containing 14 metastasis (samples from lymph node tissues are excluded) and eight non-metastasis samples. The other oral squamous cell carcinoma dataset (GSE3524) is comprised of nine metastasis and nine non-metastasis samples (two of stage-unknown samples are excluded) [23]. Both are from the same platform, Affymetrix HG-U133A, where 22,283 genes are analyzed. The size of each dataset is too small and not suitable for SVM classification if analyzed separately. However, combining the two datasets, we obtain as many as 23 metastasis and 17 non-metastasis samples, making it possible to carry out the classification analysis.

The results of the LOOCV test of the 40 samples against various numbers (1–100; increasing one gene at each step) of feature genes with four different kernels, namely, linear, polynomial, RBF, and ME with *k*NND denoising (*k*NND-ME), are shown in Figure 5. In the *k*NND-ME kernel, five different noise levels, $\nu = 0$ (no noise), 0.05, 0.1, 0.15, and 0.2 are evaluated. For comparison, we also classify the two datasets separately and average the accuracies (Figure 5a). The results clearly show that the *k*NND-ME kernel surpasses the other kernels in both averaged and mixed datasets. Statistically, the accuracies of the *k*NND-ME kernel are superior to those of the other three kernels in the averaged and the mixed datasets in 98% and 48%, respectively, of all the tested points. The difference in accuracy is much greater in the averaged dataset than in the mixed dataset. The mean differences between the

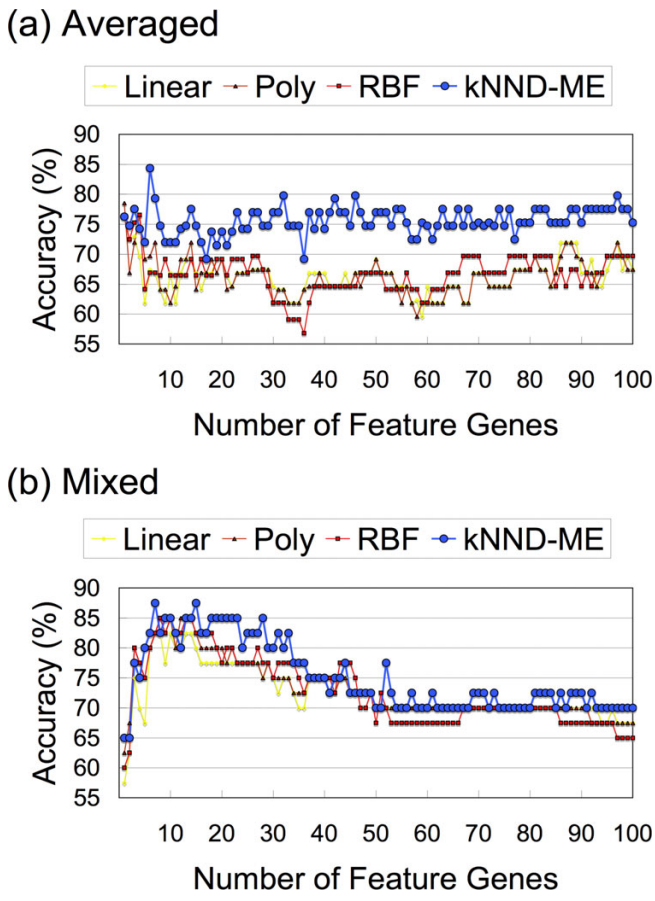

Figure 5

Oral cavity carcinoma metastasis classification. Prediction of metastasis by SVMs is performed with gene expression data of squamous cell carcinoma of the oral cavity. Classification accuracies of three kernels, i.e., linear, polynomial, RBF, and ME with *k*NND denoising, are compared. Accuracies are measured by (a) predicting each dataset separately and averaged, and (b) predicting the mixed dataset.

*k*NND-ME kernel and either of the linear, polynomial or RBF kernel with highest accuracy at each point in the averaged and the mixed datasets are 7.7% and 1.4%, respectively. The accuracies increase and plateau at around 3–30 feature genes in the mixed dataset, while no clear increase or plateau is found for the averaged dataset. The overall maximum accuracy of 87.5 (91.3/82.4 or 82.6/94.1 sensitivity/specificity)% is observed for the *k*NND-ME kernel at two points, 7 and 15 feature genes, in the mixed dataset. Those accuracies are obtained with $\nu = 0$ and 0.05 denoising parameters. The result also indicates that the *k*NND-ME kernel shows more stable and higher accuracies than the other kernels for large numbers of feature genes.

Incidentally, the top 15 feature genes that show the highest average ranks by t-statistics in the LOOCV test and that are considered to be associated with oral carcinoma

metastasis are: HFE (AF150664), FLJ12529 (NM_024811), CXorf56 (NM_022101), HEATR1 (NM_018072), MGAM (NM_004668), APOL3 (NM_014349), PYY2 (NM_021093), RBP3 (J03912), UBE2V1 (NM_003349, NM_021988, NM_022442, NM_199144, NM_199203), KCNJ15 (U73191), GLS (AB020645), ARHGEF3 (NM_019555), MDM1 (NM_020128), ZC3H13 (AL136745), and C9orf16 (NM_024112).

We further investigate the effect of the SVD denoising when applied to the mixed dataset before learning and classification. Table 2 summarizes the results of using all the six types of kernels for raw and SVD pre-denoised data. The accuracies are averaged in each of the ten gene windows. In the SVD denoising, three levels of noise removals (85, 90, and 95% of cumulative proportions), which are the same as the AML-ALL experiment, are tested.

Although a sufficient number of genes (a total of 22,283 genes) are used for SVD denoising, the denoised dataset does not significantly improve the raw accuracies in small numbers (≤ 30) of feature genes, where the overall maximum range accuracy (84.0%) exists. SVD denoising affects only large numbers (≥ 31) of feature genes. This is probably related to the property of SVD denoising that affects the ratio of information to noise content. Further analysis is needed to understand the full property of the SVD denoising method. In summary, the maximum accuracy (87.5%) of the *k*NND-ME kernel in raw data is not improved by SVD denoising (data not shown).

Discussion

Using kidney carcinoma data, we show that the ME kernel generally gives better classification results for heterogeneous microarray datasets than the three vectorial data kernels, linear, polynomial and RBF. As an alternative approach using vectorial data kernels, it is theoretically possible to train multiple SVMs for all distinct sub-data contained in the composite dataset. However, this approach has practical diffculties in that (i) there are too many heterogeneous sub-data, (ii) some sub-data contain only a few samples, and (iii) some sub-data contain all positive (or negative) samples. The SVMs cannot be trained properly with only a few samples or data with one-sided (positive or negative) labels. In addition, if we do not know the origin (i.e., platform) of the test samples, it would be diffcult to determine which SVMs should be used for the classification. The ME kernel is much simpler yet quite flexible in this regard.

Another remarkable property of our ME kernel is that the generated kernel matrices always hold positive semidefiniteness, even when the distance matrices for input to our optimization algorithm violate the triangle inequalities. This allows one to arbitrarily choose genes from among a set of feature genes to build the distance matrices in a distance-by-distance fashion. Utilizing this property, we devised the *k*NND denoising method for the distance-based kernels, which show better performance than the linear, polynomial and RBF kernels for leukemia data, even though the data are pre-denoised by SVD. This is quite important in a situation where there are few or heterogeneous samples where SVD may not work properly for denoising because the quality of the eigenvalue decomposition depends on the number of homogeneous samples. Since the *k*NND denoising method only concerns the set of genes between sample pairs, it seems quite robust with regard to the number of samples or the degree of heterogeneity.

Furthermore, the results of kidney carcinoma and oral cavity carcinoma metastasis data in Figure 3 and Table 2 clearly show that the accuracies of the ME kernel exceed those of the other two distance-based kernels, EKM and Saigo. However, in the AML-ALL data shown in Figure 4,

Table 2: Range accuracies for the mixed oral cavity carcinoma metastasis dataset

		Range of Number of Feature Genes									
Kernel	**Noise**	**1–10**	**11 – 20**	**21 – 30**	**31 – 40**	**41 – 50**	**51 – 60**	**61 – 70**	**71–80**	**81 – 90**	**91 – 100**
Linear	raw	73.8	79.3	77.0	73.5	72.8	70.0	70.0	70.0	70.0	68.5
	SVD	66.3	76.8	74.0	74.8	75.0	76.3	77.5	77.5	77.5	77.5
Polynomial	raw	77.8	81.8	77.3	74.3	72.8	70.0	70.0	70.0	70.0	67.8
	SVD	68.0	79.5	79.0	76.3	75.0	76.0	77.5	77.5	77.5	77.5
RBF	raw	77.0	82.3	78.0	75.8	73.5	68.3	68.5	70.0	68.8	66.5
	SVD	68.0	78.5	79.0	**81.8**	**80.8**	**80.3**	79.5	77.8	78.8	79.0
EKM	raw	74.5	78.5	76.8	72.3	67.3	66.8	65.5	65.5	65.0	65.0
	SVD	68.8	80.0	79.0	**81.8**	**80.8**	**80.3**	80.0	79.5	**81.0**	80.3
Saigo	raw	77.5	81.5	80.3	73.5	70.3	70.0	69.0	68.3	68.0	67.5
	SVD	68.0	77.8	77.5	76.0	75.0	76.8	77.8	78.3	78.0	78.8
*k*NND-ME	raw	**78.5**	**84.0**	**82.8**	77.8	73.3	71.3	70.8	70.5	72.0	70.3
	SVD	69.5	78.5	79.0	79.5	77.5	79.0	**80.5**	**80.0**	80.3	**83.3**

Maximum accuracy in each range of feature gene number is shown in bold.

the ME kernel and the other two distance-based methods show similar accuracies although all of them use the same *k*NND distance data. From these observations, we can conclude that the entropy maximization process works favorably for 'heterogeneous' data and allows SVMs to find the discriminant boundaries more easily than the other two distance-based methods, EKM and Saigo.

It is also important to point out that combining similar but distinct data in the microarray analysis may enhance the diagnosis of cancer or other diseases. As shown in our example of metastasis prediction for oral squamous cell carcinoma, each dataset contains only around 20 samples, which is not suitable for training of good SVM predictors, especially in the case of the vectorial data kernel family (see Figure 5a). When the datasets are combined, however, our *k*NND-ME kernel demonstrates higher and more robust classification performance than the linear, polynomial, and RBF kernels and even the other two distance-based kernels, regardless of SVD denoising.

Conclusion

We conclude that the ME kernel-based SVM classification method will generally be useful for the analysis of promiscuous microarray data of rare specimens, e.g., minor diseases or species, that present difficulty in compiling homogeneous data in a single laboratory.

Methods

In this section, we begin with a preliminary explanation of kernel methods. Then, we describe the ME kernel in terms of its basic and advantageous properties for use with heterogeneous data.

Properties of kernel methods

Kernels are numerical expressions of similarity metrics between two samples. The basic form of kernels for two sample vectors with M dimensions $\mathbf{x}_i = (x_{i1},..., x_{iM})$ and $\mathbf{x}_j = (x_{j1},..., x_{jM})$ is represented by an inner product function such as $K(\mathbf{x}_i, \mathbf{x}_j) = \varphi(\mathbf{x}_i) \cdot \varphi(\mathbf{x}_j)$. $\varphi(\cdot)$ means an arbitrary mapping of vectors to another space with generally different dimensions called 'reproducing kernel Hilbert space (RKHS)' [7], which has many properties common to those of the Euclidean space. For technical convenience, rather than defining the mapping functions $\varphi(\cdot)$ for $\mathbf{x}_i$ and $\mathbf{x}_j$, the inner product forms, i.e., $K(\mathbf{x}_i, \mathbf{x}_j) = \varphi(\mathbf{x}_i) \cdot \varphi(\mathbf{x}_j)$ of the mappings of $\mathbf{x}_i$ and $\mathbf{x}_j$ are preferably used in practical calculation [7]. The function $K(\mathbf{x}_i, \mathbf{x}_j)$ is called a kernel.

Three standard kernels popularly used in microarray studies are the linear kernel: $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i \cdot \mathbf{x}_j$, the polynomial kernel: $K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \cdot \mathbf{x}_j + 1)^D$, and the RBF kernel: $K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\sum_{k=1}^{M}(x_{ik} - x_{jk})^2 / \sigma^2\right)$. All these kernels belong to the vectorial data kernel family that takes vectorial data for N samples as input, and we can fill all the (i, j) elements in the $N \times N$ kernel matrix with the specified kernel function.

Any kernel matrix generated from such kernel functions possesses a necessary property for SVM learning called *positive semidefiniteness* (see Appendix for details). If a kernel matrix is positive semidefinite, the mapped vectors $\varphi(\mathbf{x})$ exist in the RKHS where the triangle inequalities among the mapped vectors are conserved. Our aim is to develop kernels that are robust to heterogeneous and noisy gene expression data. To this end, we first devise a new distance metric called *k*NND (detailed later) that can fulfill our requirements. However, unfortunately, the triangle inequalities are not conserved in the metric. To construct valid kernels from such distances, we introduce the following ME kernel algorithm.

ME kernel with kNND denoising

ME kernel algorithm

The ME kernel was recently devised by Tsuda and Noble [12] to represent yeast metabolic and protein-protein interaction network (graph) structures. Unlike the standard vectorial kernels, the ME kernel does not have any predefined functions. Instead, given distance constraints, D_{ij}, between samples, we obtain the ME kernel in matrix form, $\mathbf{K}$, by basically solving the following optimization problem:

$$\max_{\mathbf{K}} H(\mathbf{K}) = -\mathrm{tr}(\mathbf{K}\log\mathbf{K})$$

subject to:

$$\mathrm{tr}(\mathbf{K}) = 1,\ ||\varphi(\mathbf{x}_i) - \varphi(\mathbf{x}_j)||^2 \le D_{ij}$$

This optimization cannot be solved analytically; hence, we have implemented an efficient numerical algorithm for optimization (for technical details, see [21]). The function $H(\mathbf{K}) = -\mathrm{tr}(\mathbf{K}\log\mathbf{K})$ is called von Neumann entropy of the kernel matrix $\mathbf{K}$ [12]. The first constraint $\mathrm{tr}(\mathbf{K}) = 1$ is necessary to avoid unlimited divergence of matrix $\mathbf{K}$. Constraints $||\varphi(\mathbf{x}_i) - \varphi(\mathbf{x}_j)||^2 \le D_{ij}$ are given as prior knowledge. For example, we can give the constraint such that a particular pair of the mapped samples in the RKHS must not be distant. Regardless of the constraint values, von Neumann entropy of a kernel is always maximized so that the kernel matrix holds positive semidefiniteness [12]. Thus, one can use different gene sets in calculating D_{ij} depending on i and j samples. The RBF kernel, in contrast, violates positive semidefiniteness if different gene sets are used to construct a kernel matrix using the above function, namely, by negating and exponentiating distance matrices D_{ij} [21]. Intuitively, as shown in Figure 1, the ME kernel is built by enlarging the geometric distances among samples when

the kernel entropy is maximized. Only related samples can stay near each other due to the constraints given as D_{ij}. Matrices **K** that resulted after fully maximizing $H(\mathbf{K})$ can be used for further kernel analysis methods such as SVM classification and kernel principal component analysis [24].

kNND denoising method for ME kernel

The main issue addressed herein is how to handle missing or noisy values that exist in a large portion of a gene expression profile consisting of heterogeneous data. To effectively eliminate such spurious values without removing the entire gene, we devised the following simple method. Assume that we have a gene expression table (i.e., M genes × N samples matrix) where a sample contains ν (×100)% of noisy genes on average. In such a case, only 1 - ν of genes in that sample contain no noise. Therefore, for any pair of samples, the ratio of common genes not containing noise is expected to be $(1 - \nu)^2$. Based on this observation, we compute the distance between two samples $\mathbf{x}_i$ and $\mathbf{x}_j$ as follows: First, we compute the one-dimensional Euclidean distances $d_h = (x_{ih} - x_{jh})^2$ for h = 1, ... , M genes. Then, we select $k = (1 - \nu)^2 \times M$ of one-dimensional Euclidean distances d_h from the nearest (smallest) ones. Finally, we take the sum of the selected d_hs as the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$. We refer to this method as *k-nearest neighbor gene distance* denoising (*k*NND denoising) method hereafter. For instance, if a sample with M = 100 feature genes contains ν = 15% of noisy values, $k = (1 - 0.15)^2 \times 100 \simeq 72$ of the nearest distance genes out of the 100 feature genes are only considered in calculating *k*NNDs between samples.

Multiplying the above *k*NNDs by constant G for N samples, an $N \times N$ distance constraint matrix (D_{ij}) is generated. Note that since the samples use different gene sets in *k*NND metric, positive semidefiniteness will not hold when directly imported to the RBF kernel function. Instead, however, when our ME kernel algorithm is applied and those *k*NNDs are used *as constraint*, the resulting kernel matrix holds positive semidefiniteness as well as reflects similarities between samples. Subsequently, we will train SVMs for the optimized 'ME' kernel with sample labels.

Other distance-based kernels

For comparison, we tested two approaches to obtain a kernel matrix from distance matrix D_{ij}. Both approaches are originally devised for conversion of a non-positive-semidefinite similarity matrix **S** into a kernel matrix. The first approach is to take $\mathbf{S}^T\mathbf{S}$ as a new kernel matrix. The kernel is sometimes called *empirical kernel mapping (EKM)* [16]. The second approach is to subtract the smallest negative eigenvalue of the similarity matrix **S** from its diagonal. We call it the *Saigo kernel* [17]. We obtain a similarity matrix from distance matrix D_{ij} via $S_{ij} = \exp(-D_{ij}/\sigma^2)$.

Singular value decomposition of vectorial kernels

As an alternative and conventional approach to noise reduction, singular value decomposition (SVD) is often used in many analytical studies including microarray analysis [25,26]. We use this method to denoise microarrays for comparison. Intuitively, SVD reduces dimensions of data to only informative ones, thus denoising values with regard to non-informative dimensions. More formally, genes with N expression values that are centered by means are reduced to the major q principal components by solving the eigenvectors of M genes × N samples matrix $\mathbf{A}_{M \times N}$:

$$\mathbf{A}_{M\times N} = \mathbf{U}_{M\times N} \cdot \mathbf{W}_{N\times N} \cdot \mathbf{V}^{>}_{N\times N},$$

where $\mathbf{U}_{M \times N}$ is $M \times N$ column-orthogonal matrix (columns are called left singular vectors) and $\mathbf{V}^{>}_{N\times N}$ is $N \times N$ orthogonal matrix (rows are called right singular vectors). The matrix $\mathbf{W}_{N \times N}$ is $N \times N$ diagonal matrix and $1/(N - 1)\mathbf{W}^2$ equals eigenvalues of the uncentered covariance matrix of **A**. We choose the largest q eigenvalues and replace other diagonal elements of **W** with 0, creating the $\mathbf{W}_q$ matrix. Finally, we obtain the denoised matrix, $\mathbf{A}_q$, with $\mathbf{A}_q = \mathbf{U} \cdot \mathbf{W}_q \cdot \mathbf{V}^T$. The vectorial data kernels are subsequently computed from the denoised matrix, $\mathbf{A}_q$.

For actual analysis, software called SVDMAN developed by Wall *et al*. [27] is used.

Support vector machines

The SVMs [8] are well used for the classification of samples on the basis of their input (gene expression) values [2,28]. The basic form of SVMs is a binary classifier having a hyper-plane that distinguishes two distributions of M-dimensional vectors or samples from different classes. The hyper-plane $\mathbf{w} \cdot \varphi(\mathbf{x}_i) + b$ is obtained by solving the following optimization problem:

$$\min_{\mathbf{w},b,\xi_i} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i$$

subject to

$$\gamma_i(\mathbf{w} \cdot \varphi(\mathbf{x}_i) + b) \geq 1 - \xi_i,\ \xi_i \geq 0$$

for all N samples where γ_i and C are the class label of i-th sample and a constant parameter, respectively. Optimization of this problem yields the hyper-plane that maximizes the margin between the two classes. This is called

the SVM learning algorithm. The above optimization problem can be transformed into an equivalent form of the other equations in which only the kernel function $K(\mathbf{x}_i, \mathbf{x}_j) = \varphi(\mathbf{x}_i) \cdot \varphi(\mathbf{x}_j)$ appears and mapped vectors $\varphi(\mathbf{x}_j)$ are not explicitly described [7]. Hence, the SVM learning algorithm needs only kernel matrix **K** and mapped vectors $\varphi(\mathbf{x}_j)$ are not necessary.

Leave-one-out cross validation

The classification accuracies for the evaluation of the ME kernel against other methods are estimated by a standard leave-one-out cross-validation (LOOCV) procedure where each sample is alternatively excluded from the *N* data and the SVM trained with the remaining *N* - 1 samples predicts the excluded one. All accuracies reported in this paper are calculated with the following formula:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

where *TP*, *FP*, *TN* and *FN* are true positive, false positive, true negative, and false negative frequencies, respectively, in the classification.

Parameter selection

Since classification accuracies are dependent on parameters in the kernel-SVM method, we tested various parameter values to obtain the best performance possible. For all the six (linear, polynomial, RBF, EKM, Saigo, and ME) kernels tested here, seven SVM parameters, $C = \{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\}$, are tested. For the polynomial kernel, $D = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ are tested. For the RBF, EKM, and Saigo kernels, $\sigma = \{10^{-10}, 10^{-9}, 10^{-8}, 10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$ are tested. In the ME kernel, we used only one parameter *G* that magnifies the distance constraints D_{ij} to adjust the trade-off between over-learning and generalization of classification models (for details, see [21]). The parameter *G* has to be chosen carefully. When $G \to 0$, typically $\mathbf{K} \to \mathbf{1}\mathbf{1}^\top/N$. When $D_{ij} > 2/N$ for $\forall i, \forall j$, $\mathbf{K} \to \mathbf{I}/N$. The two are somewhat extreme cases. However, if the value of *G* is positive but too small, SVM cannot find the hyper-plane separating the positive class from the negative one clearly. If the value of *G* is too large, it leads to the so-called diagonal dominant problem [16]. We tested the parameter in the range of $G = \{2^{-5}, 2^{-4}, 2^{-3}, 2^{-2}, 2^{-1}, 1, 2, 2^2, 2^3, 2^4, 2^5\}$. Note that the number of parameter combinations in the ME kernel is equal to those in the RBF, EKM and Saigo kernels in this study.

Authors' contributions

WF conceived of the study, carried out the analyses and drafted the manuscript. TK participated in the design of the study, wrote the program and helped to draft the manuscript. Both authors have read and approved the final manuscript.

Appendix: positive semidefiniteness of kernels

Formally, the positive semidefiniteness of a kernel matrix **K**, which guarantees the existence of mapping functions $\varphi(\cdot)$ for sample vectors, $\mathbf{x}_i$ and $\mathbf{x}_j$, is defined as follows: A symmetric matrix **K** is said to be *positive semidefinite* if **K** holds

$$\sum_{i,j} c_i c_j K_{ij} \geq 0$$

for any real numbers, c_i and c_j.

Practically, a symmetric matrix **K** is positive semidefinite if and only if a mapped feature vector can be assigned to each sample such that each element of the matrix satisfies $K_{ij} = \varphi(\mathbf{x}_i) \cdot \varphi(\mathbf{x}_j)$, where $\varphi(\mathbf{x}i)$ and $\varphi(\mathbf{x}_j)$ are the mapped feature vectors of *i*-th sample and *j*-th sample, respectively [29]. For example, a symmetric matrix

$$\mathbf{K} = \begin{bmatrix} 16 & 4 & 12 \\ 4 & 5 & 9 \\ 12 & 9 & 18 \end{bmatrix}$$

is positive semidefinite since we can assign mapped feature vectors as:

$\varphi(\mathbf{x}_1) = (0, 4)\top$, $\varphi(\mathbf{x}_2) = (2, 1)\top$, $\varphi(\mathbf{x}_3) = (3, 3)^{\mathrm{T}}$,

which hold $K_{ij} = \varphi(\mathbf{x}_i) \cdot \varphi(\mathbf{x}_j)$ indeed. A symmetric matrix,

$$\mathbf{K} = \begin{bmatrix} 16 & 4 & 12 \\ 4 & 5 & 9 \\ 12 & 9 & 5 \end{bmatrix}$$

is not positive semidefinite since there exist no mapped feature vectors satisfying $K_{ij} = \varphi(\mathbf{x}_i) \cdot \varphi(\mathbf{x}_j)$.

Acknowledgements

We thank Ms. Shiori Tomoda, Dr. Satoko Takizawa, and Dr. Hideo Akiyama for oral carcinoma metastasis data preparation. We also thank Dr. Koji Tsuda and Dr. Taishin Kin for careful reading of this manuscript and useful discussions. We finally add to thank to anonymous reviewers for helpful comments and suggestions, especially for pointing out the absence of definition for the distance between missing genes, which has brought significant improvements to the results. This work was partially supported by a Grant-in-Aid for Young Scientists (B), number 18700287, from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

References

1. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES: **Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring.** *Science* 1999, **286(5439):**531-7.
2. Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang C, Angelo M, Ladd C, Reich M, Latulippe E, Mesirov JP, Poggio T, Gerald W, Loda

M, Lander ES, Golub TR: **Multiclass cancer diagnosis using tumor gene expression signatures.** *Proc Natl Acad Sci USA* 2001, **98(26):**15149-54.
3. Staunton JE, Slonim DK, Coller HA, Tamayo P, Angelo MJ, Park J, Scherf U, Lee JK, Reinhold WO, Weinstein JN, Mesirov JP, Lander ES, Golub TR: **Chemosensitivity prediction by transcriptional profiling.** *Proc Natl Acad Sci USA* 2001, **98(19):**10787-92.
4. Okutsu J, Tsunoda T, Kaneta Y, Katagiri T, Kitahara O, Zembutsu H, Yanagawa R, Miyawaki S, Kuriyama K, Kubota N, Kimura Y, Kubo K, Yagasaki F, Higa T, Taguchi H, Tobita T, Akiyama H, Takeshita A, Wang YH, Motoji T, Ohno R, Nakamura Y: **Prediction of chemosensitivity for patients with acute myeloid leukemia, according to expression levels of 28 genes selected by genome-wide complementary DNA microarray analysis.** *Mol Cancer Ther* 2002, **1(12):**1035-42.
5. van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AAM, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH: **Gene expression profiling predicts clinical outcome of breast cancer.** *Nature* 2002, **415(6871):**530-6.
6. Liu H, Li J, Wong L: **Use of extreme patient samples for outcome prediction from gene expression data.** *Bioinformatics* 2005, **21(16):**3377-84.
7. Cristianini N, Shawe-Taylor J: *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods* Cambridge, UK:Cambridge University Press; 2000.
8. Vapnik V: *Statistical Learning Theory* New York, NY, USA: J. Wiley & Sons; 1998.
9. Lodhi H, Saunders C, Shawe-Taylor J, Cristianini N, Watkins C: **Text classification using string kernels.** *The Journal of Machine Learning Research* 2002, **2:**419-44.
10. Tsuda K, Kin T, Asai K: **Marginalized kernels for biological sequences.** *Bioinformatics* 2002, **18(Suppl 1):**S268-75.
11. Kondor R, Lafferty J: **Diffusion kernels on graphs and other discrete structures.** In *Proc 19th Intl Conf on Machine Learning (ICML) [ICML 2002]* Edited by: Sammut C, Hoffmann AG. San Francisco, CA, USA: Morgan Kaufmann; 2002:315-22.
12. Tsuda K, Noble WS: **Learning kernels from biological networks by maximizing entropy.** *Bioinformatics* 2004, **20(Suppl 1):**i326-33.
13. Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, Barrette T, Pandey A, Chinnaiyan AM: **Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression.** *Proc Natl Acad Sci USA* 2004, **101(25):**9309-14.
14. Warnat P, Eils R, Brors B: **Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes.** *BMC Bioinformatics* 2005, **6(265):**.
15. Nilsson B, Andersson A, Johansson M, Fioretos T: **Cross-platform classification in microarray-based leukemia diagnostics.** *Haematologica* 2006, **91(6):**821-4.
16. Scholköpf B, Weston J, Eskin E, Leslie C, Noble WS: **A Kernel Approach for Learning From Almost Orthogonal Patterns.** In *Proceedings of ECML 2002i, 13th European Conference on Machine Learning* Helsinki, Finland: Springer; 2002:511-28.
17. Saigo H, Vert JP, Ueda N, Akutsu T: **Protein homology detection using string alignment kernels.** *Bioinformatics* 2004, **20(11):**1682-9.
18. **Supplemental datasets in this paper** [http://cellmontage.cbrc.jp/~wataru/ME/]
19. Rosner B: *Fundamentals of Biostatistics* 5th edition. Pacific Grove, CA, USA Duxbury; 2000.
20. Barrett T, Suzek TO, Troup DB, Wilhite SE, Ngau WC, Ledoux P, Rudnev D, Lash AE, Fujibuchi W, Edgar R: **NCBI GEO: mining millions of expression profiles-database and tools.** *Nucleic Acids Res* 2005:D562-6.
21. Kato T, Fujibuchi W, Asai K: **Kernel Analysis for Noisy Microarray Data.** *AIST Technical Report* 2006 [http://www.cb.k.u-tokyo.ac.jp/asailab/kato/pdf/t-kato-cbrctr2006a.pdf]. (AIST02-J00001-8)
22. O'Donnell RK, Kupferman M, Wei SJ, Singhal S, Weber R, Jr BO, Cheng Y, Putt M, Feldman M, Ziober B, Muschel RJ: **Gene expression signature predicts lymphatic metastasis in squamous cell carcinoma of the oral cavity.** *Oncogene* 2005, **24(7):**1244-51.
23. Torunera GA, Ulgera C, Alkana M, Galanted AT, Rinaggioe J, Wilkf R, Tiang B, Soteropoulosa P, Hameedh MR, Schwalba MN, Dermody JJ: **Association between gene expression profile and tumor invasion in oral squamous cell carcinoma.** *Cancer Genet Cytogenet* 2004, **154:**27-35.
24. Liu Z, Chen D, Bensmail H: **Gene expression data classification with kernel principal component analysis.** *J Biomed Biotechnol* 2005, **2005(2):**155-9.
25. Alter O, Brown PO, Botstein D: **Singular value decomposition for genome-wide expression data processing and modeling.** *Proc Natl Acad Sci USA* 2000, **97(18):**10101-6.
26. Liu L, Hawkins DM, Ghosh S, Young SS: **Robust singular value decomposition analysis of microarray data.** *Proc Natl Acad Sci USA* 2003, **100(23):**13167-72.
27. Wall ME, Dyck PA, Brettin TS: **SVDMAN-singular value decomposition analysis of microarray data.** *Bioinformatics* 2001, **17(6):**566-8.
28. Tothill RW, Kowalczyk A, Rischin D, Bousioutas A, Haviv I, van Laar RK, Waring PM, Zalcberg J, Ward R, B AV, Sutherland RL, Henshall SM, Fong K, Pollack JR, Bowtell DDL, Holloway AJ: **An Expression-Based Site of Origin Diagnostic Method Designed for Clinical Application to Cancer of Unknown Origin.** *Cancer Res* 2005, **65(10):**4031-40.
29. Schölkopf B, Smola AJ: *Learning with Kernels* Cambridge, MA, USA MIT Press; 2001.
30. Boer JM, Huber WK, Sültmann H, Wilmer F, von Heydebreck A, Haas S, Korn B, Gunawan B, Vente A, Füzesi L, Vingron M, Poustka A: **Identification and Classification of Differentially Expressed Genes in Renal Cell Carcinoma by Expression Profiling on a Global Human 31,500-Element cDNA Array.** *Genome Res* 2001, **11(11):**1861-70.
31. Su AI, Cookedagger MP, Chingdagger KA, Hakakdagger Y, Walkerdagger JR, Wiltshiredagger T, Orthdagger AP, VegaDagger RG, SapinosoDagger LM, Moqrich A, Patapoutian A, HamptonDagger GM, Schultz PG, Hogenesch JB: **Large-scale analysis of the human and mouse transcriptomes.** *Proc Natl Acad Sci USA* 2002, **99(7):**4465-70.
32. Yanai I, Benjamin H, Shmoish M, Chalifa-Caspi V, Shklar M, Ophir R, Bar-Even A, Horn-Saban S, Safran M, Domany E, Lancet D, Shmueli O: **Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification.** *Bioinformatics* 2005, **21(5):**650-9.
33. Flechnera SM, Kurianb SM, Headc SR, Sharpb SM, Whisenantc TC, Zhangd J, Chismarc JD, Horvathe S, Mondalac T, Gilmartinc T, Cooka DJ, Kayd SA, Walkerd JR, Salomon DR: **Kidney Transplant Rejection and Tissue Injury by Gene Profiling of Biopsies and Peripheral Blood Lymphocyte.** *Am J Transplant* 2004, **4(9):**1475-89.
34. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB: **A gene atlas of the mouse and human protein-encoding transcriptomes.** *Proc Natl Acad Sci USA* 2004, **101(16):**6062-7.
35. Lenburg ME, Liou LS, Gerry NP, Frampton GM, Cohen HT, Christman MF: **Previously unidentified changes in renal cell carcinoma gene expression identified by parametric analysis of microarray data.** *BMC Cancer* 2003, **3(31):**.