A top-*r* Feature Selection Algorithm for Microarray Gene Expression Data

Alok Sharma, Seiya Imoto, and Satoru Miyano

Abstract—Most of the conventional feature selection algorithms have a drawback whereby a weakly ranked gene that could perform well in terms of classification accuracy with an appropriate subset of genes will be left out of the selection. Considering this shortcoming, we propose a feature selection algorithm in gene expression data analysis of sample classifications. The proposed algorithm first divides genes into subsets, the sizes of which are relatively small (roughly of size h), then selects informative smaller subsets of genes (of size $r < h$) from a subset and merges the chosen genes with another gene subset (of size r) to update the gene subset. We repeat this process until all subsets are merged into one informative subset. We illustrate the effectiveness of the proposed algorithm by analyzing three distinct gene expression datasets. Our method shows promising classification accuracy for all the test datasets. We also show the relevance of the selected genes in terms of their biological functions.

Index Terms— Feature selection, classification accuracy, top-*r* features, DNA microarray gene expression data.

1 Introduction

Transcriptome data, which consists of several thousands of gene expression profiles obtained from multiple tissue samples, have been used to find sets of important genes for separating tissue samples into several groups with relevant biological or clinical properties. Particularly in cancer research, these kinds of methods (i.e., supervised classification) play an important role in understanding the gene regulation mechanisms of cancer heterogeneity. In the cancer classification problem, not all gene expression profiles contribute, but it is thought that several sets of genes or pathways with multiple genomic mutations determine biological or clinical properties. Therefore, for cancer classification with transcriptome data, it is crucial to discard the genes that are not important and retain the informative genes through efficient computational data analysis techniques.

Several feature selection algorithms have been developed to identify important genes [1]-[20]. Some of the methods, such as weighted naïve Bayes [6], assume feature independency. However, the features may have dependencies among themselves; genes essentially form pathways, and the expression profiles of these genes are highly correlated. Thus, this assumption of feature independency could degrade the classification performance. The feature selection methods described by Ben-Bassat [2], Golub et al. [4], Pavlidis et al. [7], Pan [10], and Mak and Kung [15] are independent of the classifier. These methods are mainly based on an individual ranking scheme; therefore, it is possible that some of the selected features are mutually redundant. This scheme is most effective for statistically independent features. Because these methods ignore the interaction with the classifier, the classification performance will not be very high. Yu and Liu [11] have proposed a classifier-independent method. However, it uses a pairwise scheme, in that the features are selected based on feature correlation. Some selection methods [1],[21] utilize a forward selection scheme. In this scheme, the best feature is selected first, and a subsequent feature is included in the subset such that the included feature improves the performance (e.g., in terms of classification) of the feature subset. The above-mentioned strategies for acquiring feature subsets could be biased towards the highest-ranking feature, as the feature with the highest performance will be selected first in the subset (or will be given highest priority). However, low-rank features, if selected in an appropriate subset, could provide better classification performance. The classifier-dependent method, reported by Inza et al. [5], could exhibit high classification performance but applies a random search, which can become computationally intensive. Ramaswamy et al. [8] have shown that the support vector machine (SVM)-based method can achieve high classification accuracy; however, large number of genes (or features) is required in a feature subset for this purpose. Therefore, no single feature selection algorithm can have superiority in all aspects. Nonetheless, most of the methods attempt to achieve high classification accuracy using a subset of genes or features.

The purpose of this paper is to design a method to find a small subset of important genes that could provide high classification accuracy. Additionally, the observed genes should have biological relevance. The proposed method should be able to identify a relatively small gene subset. Due to the small number of genes in the subset, research-

- *A. Sharma is with the Human Genome Center, Institute of Medical Science, University of Tokyo and School of Engineering & Physics, University of the South Pacific. E-mail: aloks@ims.u-tokyo.ac.jp.*
- *S. Imoto is with the Human Genome Center, Institute of Medical Science, University of Tokyo. E-mail: imoto@ims.u-tokyo.ac.jp.*
- *S. Miyano is with the Human Genome Center, Institute of Medical Science, University of Tokyo. E-mail: miyano@hgc.jp.*

Manuscript received (27-Feb-2011).

ers can conduct biological experiments for investigating biomarkers in a time-efficient and cost-effective manner. The information retrieved from these economical biological experiments can then be translated to pharmacology, which could help in the timely diagnosis of cancers. Furthermore, the combination of genes in a small subset can be easily, reliably and precisely interpreted for cancer efficacy. In the proposed method, the selected top-r features (or genes) provide promising classification accuracy. This approach for finding features is different from conventional approaches. Here, we consider finding features of low importance that could perform well, in terms of classification accuracy, if selected in an appropriate feature subset. To do this, we investigate the feature subset in the following manner.

In the proposed approach, we first partition the features into smaller blocks of size h. This block partitioning is introduced to reduce the search space. In a given block, a feature is discarded that is not performing well in terms of classification; that is, an irrelevant feature from a subset is removed at an iteration time point in the approach that causes minimum loss of information for the subset. This elimination of features from a subset is performed until all the features are ranked in a given block. This process is then performed for all the remaining blocks. Once the top-r features (where $1 < r < h$) from each of the blocks are obtained, they are compared among themselves to obtain the best feature subset. The partitioning of features into smaller blocks helps in memory management and improves computational complexity; the full search for top-r features in a block ensures the retention of the important subset of features. The proposed algorithm can also be applied to multi-class cases.

Figure 1 demonstrates the selection of informative genes from the gene subsets by the proposed algorithm. Two gene subsets are illustrated, with each containing three genes ($h = 3$). The genes in subset #1 and subset #2 are *A*, *B*, *C* and *D*, *E*, *F*, respectively. The elliptical area in the figure corresponds to the information $I(G)$, pertaining to a gene or gene subset G. It can then be seen from the figure that $I(A) > I(B) > I(C)$ and $I(D) > I(E) > I(F)$. Then, the problem is to select two genes $(r = 2)$ from the given gene subsets. First, the search algorithm finds two genes each from the subsets that have the highest information in combination. It can be seen from the figure that though gene *A* (in gene subset #1) has the highest information, it significantly overlaps with its neighboring genes. Because $I(B \cup C) > I(A \cup B)$ and $I(B \cup C) > I(A \cup C)$, removing gene A from the subset sacrifices the least amount of information. This yields a new gene subset #1a. Most of the forward selection and individual ranking schemes would select gene *A*, as it pertains to the highest information. However, the proposed scheme discards gene *A* because it is a redundant gene in combination with other genes. In gene subset #2, gene *F* does not overlap with any other genes. However, because $I(D \cup E) > I(D \cup F)$ and $I(D \cup E) > I(E \cup F)$, gene *F* will be discarded, as it contains minimal information. In this case, most of the forward selection and individual ranking schemes would follow suit. This elimination procedure yields gene subset #2a. Finally, the merger of the two gene subsets (#1a and #2a) yields gene subset #3. In this subset, two genes with the highest amounts of information will be retained. Because $I(B \cup D)$ is the maximum, the output gene subset contains genes *B* and *D*. The following can be observed from this illustration: the proposed algorithm eliminates redundant features; it removes features with minimal information; and the best gene subset is produced by evaluating all the possible combinations in a given gene subset.

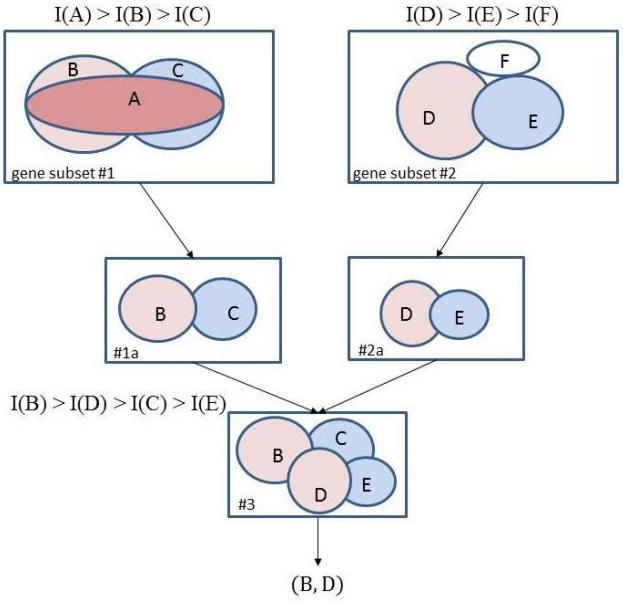


Fig. 1: An illustration of the process of gene selection using the proposed feature selection algorithm.

Three DNA microarray gene expression datasets namely, SRBCT, Prostate Tumor and MLL are used for experimentation purposes. Their performance in terms of classification accuracy using only the top-4 features is very promising.

The paper is organized as follows. Section 2 describes the proposed feature selection algorithm. Section 3 provides an illustration of the datasets used in this work. Section 4 discusses the experimentation. Section 5 discusses the sensitivity analysis of the algorithm. Section 6 describes the redundancy analysis. Section 7 concludes the paper.

2 Proposed Feature Selection Algorithm

In this section, we describe the proposed feature selection algorithm. The two main aspects of the algorithm namely, the successive feature selection and block reduction are discussed in Sections 2.1 and 2.2, respectively. The algorithm is summarized in Table 1.1, and the computational considerations are given in Section 2.3.

2.1 Successive Feature Selection

The successive feature selection (SFS) procedure has been illustrated in Figure 2. In the SFS procedure, a set of $h \leq 10$ features is processed one at a time (this value of h is taken due to memory constraints; it is experimentally found that the suitable value of h is equal to or lower

than 10). The output is the rank of features. In the successive levels, a feature is dropped one at a time, and a subset of features is obtained. Then, the classification accuracy using a classifier is evaluated, and the best subset of features is processed to the next level. There could be more than one best subset of features in a given level. For example, in Figure 2, a feature is dropped in level 1 that gives four different subsets of features. The best set in level 1 is $\{x_1, x_2, x_4\}$ which is selected for level 2. In a similar way, a feature is dropped from the best set of features of level 1 into level 2, which gives 3 different subsets of features. The best sets in level 2 are $\{x_2, x_4\}$ and $\{x_1, x_2\}$ (supposing that their classification accuracies are the same and are higher than those of other subsets), and the best set in level 3 is $\{x_2\}$. This process is terminated when all the features are ranked. In the figure, two ranked sets are obtained: namely, $R_1 = \{x_2, x_4, x_1, x_3\}$ and $R_2 = \{x_2, x_1, x_4, x_3\}$, which indicate that x_2 is the top-ranked feature and that x_3 is the bottom-ranked or least important feature. If we want to select the three top-ranked features, then the result will be $F_1 = \{x_2, x_4, x_1\}$ and $F_1 = \{x_2, x_1, x_4\}$. If the order of features is not important, then instead of writing two sets, F_1 and F_2, we can select a set of common top-3 ranked features: i.e., $F_k = F_1 \cup F_2 = \{x_1, x_2, x_4\}$.

The SFS procedure can be applied by partitioning the training data into a training set and a validation set. The training set is used to estimate the model parameters of the classifier, and the validation set is used to evaluate the classification accuracy of the feature subsets at each of the levels.

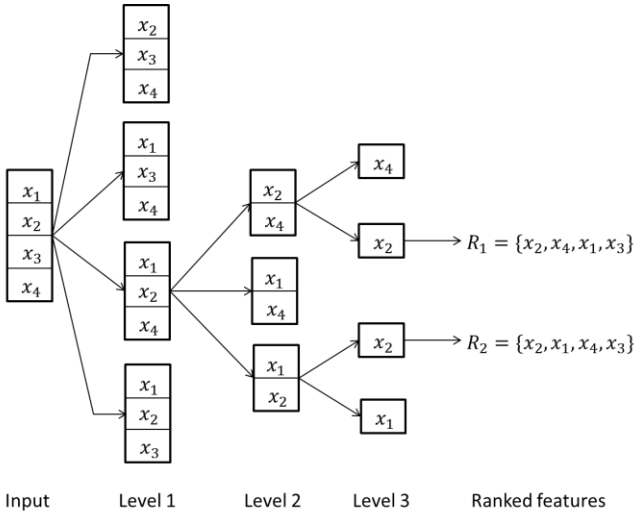


Fig. 2: Successive feature selection: an illustration.

2.2 Block Reduction

The block reduction procedure has been briefly described in Figure 3. A d-dimensional feature vector has been partitioned into m roughly equal blocks, S_j, for $j = 1 \ldots m$ of size $h \le 10$. Each block has at least r features. All the blocks have been processed through the SFS procedure one at a time, which yields top-r feature sets, F_j, for $j = 1 \ldots q$. Then, the unique features of two consecutive feature sets, F_1 and F_2, are used to find the best top-r feature set, F_b. Next, the unique features of F_b and F_3 are used to obtain the best set. This process is continued for all the q sets. The obtained best top-r feature set, F_b, from the block reduction procedure is stored for further pruning. The details of the block reduction procedure have been described in Table 1.1.

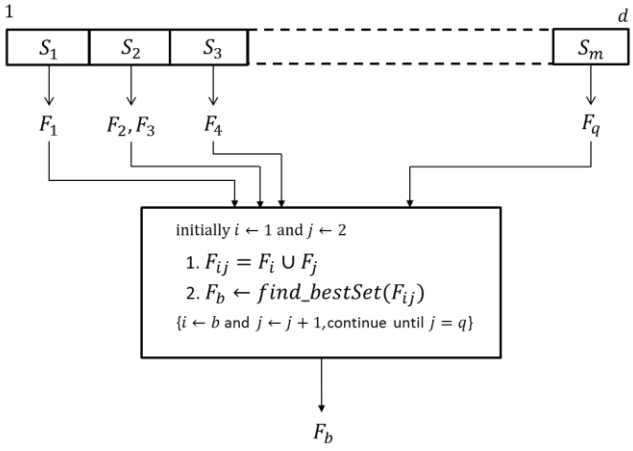


Fig. 3: Block reduction procedure to find the top-r features.

TABLE 1.1
Block Reduction Procedure

1. Select the r number of features to be investigated, where $1 < r < h$, and select the block size h, where $h \le 10$.
2. Decompose the training samples randomly into a training set (Tr) and a validation set (V) using a proportionality ratio p [1].
3. Partition the features of the sets (Tr and V) into m roughly equal blocks, S_j, for $j = 1 \ldots m$.
4. Apply the successive feature selection (SFS) procedure on each of S_j to get the top-r ranked feature set, F_j, and its corresponding classification accuracy, α_j, for $j = 1 \ldots q$, where $q \ge m$ and $F_j \ne F_l \forall j \ne l$.
5. Initialize $i \leftarrow 1$ and $j \leftarrow 2$.
6. Find the best feature set $F_b \leftarrow find\_bestSet(F_i, F_j)$. (see Table 1.2).
7. Terminate the process if $j = q$, or else update $i \leftarrow b$ and $j \leftarrow j + 1$, and go to Step 6.
8. If more than one set of F_b is obtained, then perform cross-validation to get one best set (for cross-validation, decompose training samples randomly n times[2] into training sets and validation sets using the proportionality ratio p and compute the average classification accuracy for all sets in F_b; select a set of F_b for which the average classification accuracy is the highest).
9. Repeat Steps 2-8 for another random decomposition of training samples. Let the new training set and validation set be defined as Tr^* and V^*. This will give a best set F_b^*.
10. Find the best set and its corresponding average classification accuracy (α_b) using F_b and F_b^*; i.e., $[F_b, \alpha_b] \leftarrow find\_set\&alpha(F_b, F_b^*)$ (see Table 1.3).
11. Repeat Steps 9-10 until α_b does not show any improvement.

[1] In our experiments, we use $p = 0.6$; i.e., 0.6 of training samples are used as a training set, and the remaining training samples are used as a validation set.

[2] In our experiments, we use $n = 20$.

TABLE 1.2
Find Best Feature Set
$F_b \leftarrow find\_bestSet(F_i, F_j)$

1. Find a set of unique features from the consecutive feature blocks: i.e., $F_{ij} = F_i \cup F_j$.
2. Apply the SFS procedure on F_{ij} as follows: $[R_k, \alpha_k] \leftarrow SFS(F_{ij})$, where R_k is the ranked set of all the features of F_{ij} and α_k is the classification accuracy of the top-r features of R_k. If n_1 ranked feature sets $R_k = \{\hat{R}_1, \hat{R}_2, \ldots \hat{R}_{n_1}\}$ are discovered, then there will be n_1 classification accuracies, $\alpha_k = \{\hat{\alpha}_1, \hat{\alpha}_2, \ldots \hat{\alpha}_{n_1}\}$, and n_1 top-r feature sets, F_k. Because α_k represents the classification accuracy of the top-r features, all the n_1 values in α_k will be identical. If there are any duplicate sets in F_k then removing them will yield $n_2 (< n_1)$ unique $F_k = \{\hat{F}_1, \hat{F}_2, \ldots \hat{F}_{n_2}\}$ sets. Note that all the sets in F_k will have the same classification accuracy.
3. Define $F = \{F_i, F_j, F_k\}$ and their corresponding classification accuracies, $\alpha = \{\alpha_i, \alpha_j, \alpha_k\}$. Find $\alpha_b = \mathbf{max}(\alpha)$.
   If only one α_b is obtained, then:
   - select $F_b \in F$

   If more than one α_b is obtained, then:
   - select $F_{tmp} \in F$ for which $\alpha_b \geq \alpha_t$, where α_b is the highest classification accuracy and (b,t) can be (i,j,k) but $b \neq t$.
   - perform cross-validation by decomposing the training samples randomly n times into training sets and validation sets using the proportionality ratio p and computing the average classification accuracy for all the sets in F_{tmp}.
   - select $F_b \in F_{tmp}$ for which the average classification accuracy is highest[3]; store the corresponding $\alpha_b \in \alpha$.

   If $\alpha_b = 100$ (maximum accuracy), then:
   - store F_b in an optimum set F_{opt}.
   - $\alpha_b \leftarrow \alpha_b - \varepsilon$, where ε is a small positive number, e.g., $\varepsilon = 0.5$ (such that the other feature set can also be compared to, in the remaining $F_j s$).
4. Return $F_b \leftarrow \{F_b, F_{opt}\}$.

TABLE 1.3
Find the Best Feature Set and Average Classification Accuracy
$F_b \leftarrow find\_set\&alpha(F_i, F_j)$

1. Find a set of unique features from the two best feature sets: i.e., $F_{bb^*} = F_b \cup F_b^*$.
2. Find the top-r ranked feature set, F_k, using the SFS procedure on F_{bb^*}.
3. Define $F = \{F_b, F_b^*, F_k\}$; decompose the training samples randomly n times into training sets and validation sets by using the proportionality ratio p, and compute the average classification accuracy for all the sets in F.
4. If the highest average classification accuracy is α_b and the corresponding set is F_b then return $[F_b, \alpha_b]$.

[3] If more than one F_b is obtained from the cross-validation procedure, then increase the value of n to obtain only one F_b. Alternatively, select the first best set that may sacrifice the classification performance.

2.3 Computational Considerations

The computational cost of the method depends upon several factors, such as the cross-validation process, the value of the parameter h and the type of classifier used. The processing time of the SFS procedure could be slow if a large value of h is considered or if several gene subsets in a given level have the same classification accuracy. The block reduction helps in memory management and computational complexity by limiting the value of h; that is, by using the SFS procedure on smaller blocks of features.

Feature selection methods based on the individual ranking scheme [4],[7],[10],[15] are the most economical for computation. The methods with a pairwise scheme (e.g., Yu and Liu [11]) require around $O(d^2)$ computations, where d is the dimensionality of the feature space. The pairwise scheme is slower to process than the individual ranking scheme. In the backward elimination scheme, a full-feature search space is used. If the dimensionality of the feature space is very large, then the search will be computationally very expensive. In the proposed feature selection method, the block reduction procedure is used to partition the d-dimensional feature set into roughly equal d/h blocks of size h and the SFS procedure is used to find the feature subset that requires between ${}^{h+1}C_2$ and $2^h - 1$ search combinations, where the term ${}^pC_q$ is the q-combination of p elements. The total number of search combinations for the method would be between ${}^{h+1}C_2 d/h$ and $(2^h - 1)d/h$.

3 Datasets Used in the Experimental Setup

Three DNA microarray gene expression datasets are utilized in this work to show the effectiveness of the proposed method. The descriptions of the datasets are given as follows:

SRBCT dataset [22]: the small round blue-cell tumor dataset consists of 83 samples, each containing 2308 genes. This is a 4-class classification problem. The tumors are Burkitt's lymphoma (BL), the Ewing family of tumors (EWS), neuroblastoma (NB) and rhabdomyosarcoma (RMS). There are 63 samples for training and 20 samples for testing. The training set consists of 8, 23, 12 and 20 samples of BL, EWS, NB and RMS, respectively. The test set consists of 3, 6, 6 and 5 samples of BL, EWS, NB and RMS, respectively.

MLL dataset [23]: this dataset contains 3 classes of leukemia, namely acute lymphoblastic leukemia (ALL), myeloid/lymphoid leukemia (MLL) and acute myeloid leukemia (AML). The training set contains 57 leukemia samples (20 ALL, 17 MLL and 20 AML), whereas the test set contains 15 samples (4 ALL, 3 MLL and 8 AML). The dimension of the MLL dataset is 12582.

Prostate Tumor dataset [24]: this is a 2-class problem addressing tumor class versus normal class. It contains 52 prostate tumor samples and 50 non-tumor (or normal) samples. Each sample is described by 12600 genes. A separate test set contains 25 tumor and 9 normal samples.

The summary of datasets is given in Table 2.

TABLE 2
Summary of the Datasets used in the Experimentation

Dataset	Class	Dimension (number of genes)	Training samples	Test samples
SRBCT	4	2308	63	20
MLL	3	12582	57	15
Prostate Tumor	2	12600	102	34

4 Application to Transcriptome Data

We select the best r genes for each of the datasets. These genes are selected using the training samples only. For classification purposes, three classifiers, namely, the linear discriminant analysis (LDA) technique with nearest centroid classifier (NCC), the Bayes classifier and the nearest neighbor classifier (NNC) are used. To select the best features from the training samples, we apply a cross-validation procedure by randomly decomposing the training samples into two sets. The first set contains roughly 60% of the training samples, and the second set contains the remaining training samples. The first set is used as a training set, and the second set is used as a validation set. The cross-validation procedure has been applied at different steps in the proposed feature selection algorithm (we use $n = 20$, see Table 1). From the first iteration of the algorithm, a set of top-r genes, F_b, is obtained. The iteration continues until no improvement in the classification accuracy of the selected top-r genes is observed. To find the value of r, we apply a strategy used by Tao et al. [25]. According to this strategy, a set of experiments are first conducted on a dataset (here, SRBCT) by varying the number of genes selected to find the best classification accuracy using LDA with the NCC, the Bayes classifier and the NNC. It is observed that $r = 4$ gives a favorable performance in terms of classification accuracy (see Appendix-I). Therefore, for the other datasets, the top-4 genes are used. The gene expression data analysis (Tables 3-10 and Figure 4) is reported in this paper for LDA with the NCC. However, the classification accuracies with the Bayes classifier and the NNC are also reported for comparison purposes in Tables 4-6.

We have conducted two phases of the test. In the first phase, the original order of genes (usually ranked genes from the donors) was used. For this phase, the search for the top-4 genes using LDA with the NCC is ended at iteration number 2 for the SRBCT and Prostate Tumor datasets and is ended at iteration number 20 for the MLL dataset. The top-4 selected genes for the different datasets are depicted in Table 3.

The performance in terms of classification accuracy of the proposed feature selection method has been compared with several other feature selection methods. Tables 4-6 show comparisons to various methods. In all cases, the performance is measured on the test samples. Table 4 shows a comparison of the proposed method with several other methods on the SRBCT dataset. Table 5 shows a comparison of the proposed method with several other methods on the MLL dataset. Table 6 shows the comparison on the Prostate Tumor dataset. It can be observed from Tables 4-6 that the proposed feature selection method yields promising results using only 4 genes.

TABLE 3
Selected Top-4 Genes

SRBCT Gene symbol/ accession (numbers)	MLL Gene symbol/ accession (numbers)	Prostate Tumor Gene symbol/ accession (numbers)
CD99 (11)	41462_at (4341)	38322_at (6390)
SGCA (20)	33699_at (4982)	1513_at (11215)
WAS (62)	34306_at (7930)	863_g_at (11858)
SH3BGR (723)	35260_at (8165)	829_s_at (11871)

In the second phase, we randomly change the order of genes $n = 20$ times and apply the algorithm to select the genes. This permutation of gene order is done to test the sensitivity of the algorithm with respect to the order of genes. Then, LDA, with the NCC, is used to compute the classification accuracy. In these 20 tests, some of the genes repeatedly appear in the top-4 selected genes. For the SRBCT dataset, the classification accuracy ranged between 85% and 100%; for the MLL dataset, the classification accuracy ranged between 80% and 100%; and for the Prostate Tumor dataset, the classification accuracy ranged between 76.47% and 100%. The gene clusters with the best classification accuracy (i.e., 100%) are depicted in Table 7. The collated genes obtained from the proposed algorithm on all the tests are shown in Table 8. The classification accuracy by permuting the order is in a reasonable range. However, the accuracy can be improved by retuning the size of the feature subset. Because we want to provide a small number of important genes that can be translated to biological experiments (e.g., finding driver mutations), we prefer keeping the size of the feature subset small. It is therefore a trade-off between the size of the feature subset and the sensitivity with respect to gene order. However, the range of classification accuracy can be improved in the following manner. If N times the permutation and selection are conducted, then N gene subsets will be obtained. It is possible to retrieve $M < N$ gene subsets from these N subsets that are exhibiting better performances in terms of classification accuracy on training data (e.g., using a k-fold cross validation). These M gene subsets can be used in further computational and biological analyses.

Next, we investigated the biological significance of the genes obtained from the experiments. To do this, we collated all the genes obtained from the tests: that is, the top-4 genes obtained using the original order and the top-4 genes obtained when randomizing the order 20 times. Gene duplicates are removed and processed with FatiGo [26] and Ingenuity Pathway Analysis (IPA, http://www.Ingenuity.com) to find their functional properties. The p-value has been kept lower than 5%, and the Gene Ontology (GO) biological process (levels from 3 to 9) has been used for all the datasets. Several significant GO biological processes have been obtained. However, processes with term sizes of less than 100 have been provided in the supplementary material at the link

{http://bonsai.hgc.jp/~imoto/SuplTCBB11Sharma.pdf}. A total of 200 significant terms are obtained for the SRBCT dataset using a collection of 36 genes from the proposed algorithm. For the MLL dataset, 19 significant terms are obtained using 42 genes. For the Prostate Tumor subset, 7 significant terms are obtained using 47 genes. We use the Prostate Tumor subset as a prototype to detail functional properties. Processing the selected 47 genes of the Prostate Tumor subset gives several high-level biological functions. The top-5 high-level functions are depicted in Figure 4. In the figure, the horizontal axis shows the high-level functions, and the vertical axis shows the negative logarithm of the p-value. The most significant high-level function of attention is the cancer function. The cancer function exhibits several sub-functions. Some of its sub-functions are listed in Table 9. In Table 9, the first column depicts the tumor-related functions; the second column depicts the corresponding p-values; the third column shows the gene symbols; and the fourth column defines the number of genes used for the corresponding functions.

TABLE 4
Comparison of the Methods on the SRBCT Dataset

Methods (Feature Selection + Classification)	Number of selected genes	Classification accuracy
Information gain + Naïve Bayes [25]	150	68%
Information gain + SVM random [25]	150	95%
Information gain + SVM exhaustive [25]	150	91%
Towing rule + Naïve Bayes [25]	150	73%
Towing rule + SVM random [25]	150	95%
Towing rule + SVM exhaustive [25]	150	95%
Sum minority + Naïve Bayes [25]	150	68%
Sum minority + SVM random [25]	150	95%
Sum minority + SVM exhaustive [25]	150	91%
Max minority + Naïve Bayes [25]	150	77%
Max minority + SVM random [25]	150	91%
Max minority + SVM exhaustive [25]	150	91%
Gini index + SVM Naïve Bayes [25]	150	78%
Gini index + SVM random [25]	150	95%
Gini index + SVM exhaustive [25]	150	95%
Sum of variances + SVM Naïve Bayes [25]	150	63%
Sum of variances + SVM random [25]	150	91%
Sum of variances + SVM exhaustive [25]	150	95%
t-statistics + Naïve Bayes [25]	150	63%
t-statistics + SVM random [25]	150	91%
t-statistics + SVM exhaustive [25]	150	95%
One-dimensional SVM + SVM Naïve Bayes [25]	150	63%
One-dimensional SVM + SVM random [25]	150	91%
One-dimensional SVM + SVM exhaustive [25]	150	95%
Information gain + LDA with NCC	4	70%
Information gain + Bayes classifier	4	45%
Information gain + NNC	4	60%
Chi-squared + LDA with NCC	4	55%
Chi-squared + Bayes classifier	4	50%
Chi-squared + NNC	4	70%
Gain Ratio + LDA with NCC	4	75%
Gain Ratio + Bayes classifier	4	85%
Gain Ratio + NNC	4	85%
Proposed feature selection + LDA with NCC	4	100%
Proposed feature selection + Bayes classifier	4	90%
Proposed feature selection + NNC	4	95%

TABLE 5
Comparison of the Methods on the MLL Dataset

Methods (Feature Selection + Classification)	Number of selected genes	Classification accuracy
BScatter + SVM Linear [27]	5-5000	91.7% - 100%
BScatter + SVM Polynomial [27]	5-5000	91.7% - 98.6%
BScatter + SVM Gaussian [27]	5-5000	90.3% - 98.6%
Chi-squared + SVM Linear [27]	5-5000	90.3% - 97.2%
Chi-squared + SVM Polynomial [27]	5-5000	91.6% - 97.2%
Chi-squared + SVM Gaussian [27]	5-5000	80.6% - 97.2%
Information gain + Naïve Bayes [25]	150	67%
Information gain + SVM random [25]	150	100%
Information gain + SVM exhaustive [25]	150	93%
Towing rule + Naïve Bayes [25]	150	60%
Towing rule + SVM random [25]	150	100%
Towing rule + SVM exhaustive [25]	150	93%
Sum minority + Naïve Bayes [25]	150	67%
Sum minority + SVM random [25]	150	87%
Sum minority + SVM exhaustive [25]	150	80%
Max minority + Naïve Bayes [25]	150	73%
Max minority + SVM random [25]	150	87%
Max minority + SVM exhaustive [25]	150	80%
Gini index + SVM Naïve Bayes [25]	150	60%
Gini index + SVM random [25]	150	100%
Gini index + SVM exhaustive [25]	150	93%
Sum of variances + SVM Naïve Bayes [25]	150	60%
Sum of variances + SVM random [25]	150	100%
Sum of variances + SVM exhaustive [25]	150	93%
t-statistics + Naïve Bayes [25]	150	60%
t-statistics + SVM random [25]	150	100%
t-statistics + SVM exhaustive [25]	150	93%
One-dimensional SVM + SVM Naïve Bayes [25]	150	60%
One-dimensional SVM + SVM random [25]	150	100%
One-dimensional SVM + SVM exhaustive [25]	150	93%
Information gain + LDA with NCC	4	73%
Information gain + Bayes classifier	4	73%
Information gain + NNC	4	67%
Chi-squared + LDA with NCC	4	73%
Chi-squared + Bayes classifier	4	80%
Chi-squared + NNC	4	60%
Gain Ratio + LDA with NCC	4	80%
Gain Ratio + Bayes classifier	4	80%
Gain Ratio + NNC	4	93%
Proposed feature selection + LDA with NCC	4	100%
Proposed feature selection + Bayes classifier	4	100%
Proposed feature selection + NNC	4	93%

TABLE 6
Comparison of the Methods on the Prostate Tumor Dataset

Methods (Feature Selection + Classification)	Number of selected genes	Classification accuracy
PCLs [28]	Unknown	97%
Discretization + decision trees [29]	3071	74%
RCBT [12]	unknown	97%
SVMs [12]	unknown	79%
Signal to noise ratios + kNN [24]	4	77%
α-depended degree + decision rule [18]	1	91%
Information gain + LDA with NCC	4	74%
Information gain + Bayes classifier	4	74%
Information gain + NNC	4	74%
Chi-squared + LDA with NCC	4	74%
Chi-squared + Bayes classifier	4	74%
Chi-squared + NNC	4	74%
Gain Ratio + LDA with NCC	4	85%
Gain Ratio + Bayes classifier	4	94%
Gain Ratio + NNC	4	62%
Proposed feature selection + LDA with NCC	4	100%
Proposed feature selection + Bayes classifier	4	97%
Proposed feature selection + NNC	4	97%

TABLE 7
Top-4 Genes with the Best Classification Accuracy

SRBCT Gene symbol/accession (numbers)	MLL Gene symbol/accession (numbers)	Prostate Tumor Gene symbol/accession (numbers)
{SGCA (20), APCDD1 (51), AF1Q (120), CORO1A (48)} and {SGCA (20), IGF2 (46), GATA3 (116), HLA-DPA1 (561)}	{38604_at (3399), 34306_at (7930), 34410_s_at (8034), 39556_at (9586)}	{41480_at (4377), 37639_at (6185), 38322_at (6390), 40045_g_at (6915)}

TABLE 8
The Genes Obtained from the Proposed Algorithm on the SRBCT, MLL and Prostate Tumor Datasets by Randomly Permuting the Gene Orders

SRBCT (gene symbols/ accessions)	{CDK6, CAV1, FCGRT, CD99, CD79B, SGCA, NCOA1, IGF2, CORO1A, APCDD1, PAPPA, WAS, ELF1, RYR1, CCND1, FVT1, TNFAIP6, FNDC5, FNDC5, NEB, GATA3, AF1Q, MAPK7, MYO1B, ANTXR1, CD79A, PSMB8, ARHE, CBX1, HLA-DPA1, SH3BGR, YAP1, KCNAB2, BTK, CSDA, SLC35A1, MGC11349}
MLL (gene symbols/ accessions)	{AFFX-HUMTFRR/M11507_5_at, 31575_f_at, 32378_at, 36780_at, 37508_f_at, 38521_at, 38604_at, 41462_at, 33305_at, 33699_at, 33806_at, 34771_at, 35161_at, 36553_at, 37539_at, 37933_at, 37944_at, 38969_at, 39011_at, 39707_at, 40763_at, 40777_at, 40797_at, 41752_at, 32808_at, 33882_at, 34306_at, 34410_at, 35260_at, 35340_at, 36986_at, 37346_at, 37766_s_at, 38046_at, 39556_at, 41523_at, 32543_at, 1914_at, 1894_f_at, 1752_at, 1008_f_at, 573_at}
Prostate Tumor (gene symbols/ accessions)	{31444_s_at, 31959_at, 35116_at, 35119_at, 35430_at, 35465_at, 35928_at, 39939_at, 41480_at, 41661_at, 35164_at, 35235_at, 35641_g_at, 36928_at, 37639_at, 38322_at, 40045_g_at, 40436_g_at, 40839_at, 32815_at, 34792_at, 34840_at, 34865_at, 36666_at, 37330_at, 37720_at, 37736_at, 38026_at, 38028_at, 38057_at, 38098_at, 38125_at, 38406_f_at, 39816_g_at, 40282_s_at, 41504_s_at, 32598_at, 2041_i_at, 2035_s_at, 1922_g_at, 1513_at, 914_g_at, 863_g_at, 829_s_at, 769_s_at, 440_at, 322_at}

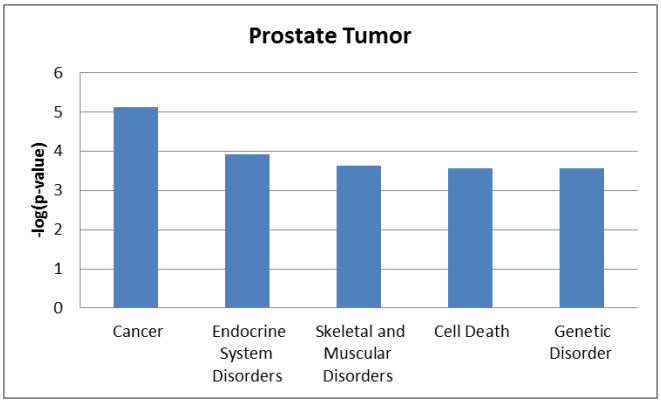


Fig. 4: Top-5 high-level biological functions for the Prostate Tumor subset using the selected genes.

We have also conducted experiments to find common genes between the proposed algorithm and the three other feature selection techniques namely information gain, Chi-squared and gain ratio. To do this, we first select top-4 genes from all the four techniques and then the resulting genes are compared. The findings are highlighted as follows: for SRBCT dataset, gene symbols GATA3 and APCDD1 are common; for MLL dataset, 35260_at is common; and, for Prostate Tumor dataset, 37639_at is common. Though some of the genes are common, the resulting gene subsets are different which led to different classification performance.

TABLE 9
Cancer Functions

Function	*p*-value	Gene symbols	#Genes
Benign tumor	7.66E-06	ABL1, AHCYL1, C18ORF1, DPT, FBLN1, GSTP1, MAF, PIK3R3, SERPINE1	9
tumor	1.14E-05	ABL1, AHCYL1, ANXA2, C18ORF1, DPT, ENO1, ERG, FBLN1, GSTP1, HPN, KLK3, MAF, P4HB, PIK3R3, PTGDS, RECK, RPL13A, SERPINE1	18
Neoplasia	2.70E-05	ABL1, AHCYL1, ANXA2, C18ORF1, COL4A6, DPT, ENO1, ERG, FBLN1, GSTP1, HPN, KLK3, MAF, P4HB, PIK3R3, PTGDS, RECK, RPL13A, RUNX1T1, SERPINB5, SERPINE1, WFS1	22
Prostatic carcinoma	2.80E-04	ANXA2, ERG, GSTP1, HPN, KLK3, PTGDS	6
Cancer	2.89E-04	ABL1, AHCYL1, ANXA2, C18ORF1, DPT, ENO1, ERG, GSTP1, HPN, KLK3, MAF, P4HB, PIK3R3, PTGDS, RECK, RPL13A, RUNX1T1, SERPINB5, SERPINE1, WFS1	20
Genital tumor	7.81E-04	ANXA2, ERG, GSTP1, HPN, KLK3, PTGDS, RECK	7
Prostate cancer	1.60E-03	ABL1, ANXA2, ERG, GSTP1, HPN, KLK3, PTGDS	7
Prostatic intraepithelial neoplasia	1.97E-03	ANXA2, KLK3, PTGDS	3
Digestive organ tumor	4.18E-03	ABL1, ANXA2, GSTP1, P4HB, RPL13A, SERPINE1	6
Gastrointestinal tumor	8.24E-03	ABL1, GSTP1, P4HB	3
Colorectal tumor	1.57E-02	ABL1, GSTP1, SERPINE1	3
Brain tumor	1.61E-02	ABL1, RECK, SERPINE1	3
Bone tumor	2.00E-02	ABL1, SERPINE1	2
Prostatic intraepithelial tumor	2.21E-02	ANXA2, PTGDS	2
Secondary tumor	2.21E-02	ABL1, SERPINE1	2

5 Sensitivity Analysis

We check the sensitivity of the proposed approach by adding Gaussian noise to the expression values. We use three different levels of Gaussian noise to contaminate the data. The generated noise levels are 1%, 5% and 10% of the standard deviation of the original expression values. The contaminated data are then analyzed again to obtain a new set of genes. We repeat the adding of Gaussian noise 10 times at each of the levels and show the 8 leading genes in Table 10.

It can be seen from the sensitivity analysis table (Table 10) that the proposed approach is able to select all the genes that were also selected in the original analysis (Table 3).

6 Redundancy Analysis

We perform redundancy analysis to investigate the redundant biological information in terms of classification accuracy on the obtained best gene subsets. To perform this analysis, we use the SRBCT dataset and the gene subsets as depicted in Tables 3 and 7. The training set and the test set of the SRBCT dataset are merged into one dataset to compute the average classification accuracy using a k-fold cross-validation procedure (where $k = 3$) on the selected gene subsets. The LDA with the NCC technique is utilized as a classifier. One gene from a given subset is removed, and the average classification accuracy is evaluated to investigate the loss of information. This removal of a gene and computation of accuracy are performed until all the genes in a given subset are evaluated. The results are depicted in Table 11.1 (gene numbers are depicted in the table; for corresponding gene symbols, see Tables 3 and 7). It can be seen from Table 11.1 that each of the genes in a given subset contains unique information. None of the genes are redundant. Table 11.2 shows the information loss by removing a gene from a given subset. It can be observed from Table 11.2 that though gene number 20 is common to all three subsets, it contributes a different level of information for the different subsets; in other words, there are other important genes that contribute to a given subset with respect to classification performance. Therefore, different subsets of genes can be used equally well for classification.

7 Conclusion

In this paper, we propose a feature selection algorithm for classification problem using transcriptome data. In many feature selection algorithms (e.g. individual ranking and forward selection schemes), the gene selection is biased towards the highest ranking feature. However, low-rank genes, if appropriately selected in a subset, can exhibit better classification performance. The proposed algorithm explores this phenomenon and provides a way to investigate important genes. It is observed that the algorithm finds a small gene subset that provides high classification accuracy on several DNA microarray gene expression datasets. These subsets contain top-r genes. The small number of (r) genes would help to conduct biological experiments for investigating biomarkers in a time-efficient and cost-effective manner. The information retrieved from these economical biological experiments can then be translated to pharmacology, which could also help in the timely diagnosis of cancers. Furthermore, the combination of genes in a small subset can be easily, reliably and precisely interpreted for cancer efficacy. The proposed algorithm is also compared to several other feature selection methods and promising results are obtained. The biological significance of the obtained gene subset was also highlighted by identifying its functional properties. Moreover, the sensitivity analysis is conducted to observe the robustness of the algorithm in a noisy environment. For this purpose, the DNA microarray data were contaminated by different level of noise and the algorithm was carried out to find the genes. It is observed that the algorithm was able to select all the genes that were also selected in the original noise free environment. Furthermore, redundancy analysis is conducted to explore the importance of individual genes in a given gene subset or in other words, redundancy in gene subsets. It is observed that the genes are not redundant in a gene subset and therefore, different subsets of genes can be used equally well for classification. The following points are highlighted which can be addressed for future work:

- To permute the data N times to obtain $M < N$ gene subsets for classification and biological analyses.
- To investigate the value of selected number of genes, r, based on a specific data topology.
- To develop a ranker for pre-processing data so that the order of data has minimal or no effect on the gene selection and classification performance. However, this could provide gene subsets biased towards the ranker used, but could make the processing faster.

Appendix

In this appendix section, we experimentally compute a sufficient value of r on SRBCT dataset. In order to determine the value of r, we first varied r between 2 and 5, and for each of the value we compute the training classification accuracy and test classification accuracy using three classifiers: LDA with nearest centroid classifier (NCC), Bayes classifier and nearest neighbor classifier (NNC). The results are depicted in Table A1. It can be observed from the table that the variation of performance (among classification accuracies obtained by different classifiers) is small when $r = 4$. Also at $r = 4$, the average classification accuracy is highest. Therefore, we select $r = 4$.

Acknowledgment

We thank the Human Genome Center at the University of Tokyo for providing supercomputing resources. We also thank the two Reviewers and the Editor for their constructive comments which appreciably improved the presentation quality of the paper.

TABLE 10
Sensitivity Analysis: Selection of the Genes After Adding Gaussian Noise on Expression Values for the SRBCT Dataset

Std. dev. = 1%		Std. dev. = 5%		Std. dev. = 10%	
Frequency	Gene symbol (number)	Frequency	Gene symbol (number)	Frequency	Gene symbol (number)
10	SH3BGR (723)*	10	SGCA (20)*	9	SH3BGR (723)*
10	WAS (62)*	8	SH3BGR (723)*	7	CD99 (11)*
10	SGCA (20)*	8	CD99 (11)*	5	SGCA (20)*
10	CD99 (11)*	8	WAS (62)*	3	WAS (62)*
		2	GATA3 (116)	3	FCGRT (4)
		1	CD79A (364)	2	PCOLCE (981)
		1	SLPI (530)	2	ALDH7A1 (101)
		1	MGC11349 (1537)	2	ZNF358 (539)

*Genes were also present in the original analysis.

TABLE 11.1
Redundancy Analysis by Removing One Gene at a Time from the Selected Three Subsets and Applying a k-fold Cross-Validation ($k = 3$) to Compute the Classification Accuracy on the SRBCT Dataset

Subset #1	%	Subset #2	%	Subset #3	%
[11,20,62,723]	100.00	[20,51,120,48]	100.00	[20,46,116,561]	100.00
[11,20,62]	85.90	[20,51,120]	76.92	[20,46,116]	83.33
[11,20,723]	71.79	[20,51,48]	88.46	[20,46,561]	82.05
[11,62,723]	74.36	[20,120,48]	83.33	[20,116,561]	93.59
[20,62,723]	85.90	[51,120,48]	85.90	[46,116,561]	93.59

TABLE 11.2
Information Loss in Terms of Classification Accuracy when Removing a Gene from the Subset

Gene in subset #1	%	Gene in subset #2	%	Gene in subset #3	%
20	-25.64	20	-14.10	20	-6.41
62	-28.21	51	-16.67	46	-6.41
723	-14.10	120	-11.54	116	-17.95
11	-14.10	48	-23.08	561	-16.67

TABLE A1
Training Classification Accuracy (TrCA) and Test Classification Accuracy (TeCA) Using Different Values of r on SRBCT Dataset

value r	LDA with NCC	Bayes classifier	NNC	Average
2	TrCA: 78.2%	TrCA: 87.2%	TrCA: 90.7%	TrCA: 85.4%
	TeCA: 55.0%	TeCA: 95.0%	TeCA: 60.0%	TeCA: 70.0%
3	TrCA: 95.9%	TrCA: 89.6%	TrCA: 99.6%	TrCA: 95.0%
	TeCA: 100.0%	TeCA: 80.0%	TeCA: 95.0%	TeCA: 91.7%
4	TrCA: 99.6%	TrCA: 93.7%	TrCA: 100.0%	TrCA: 97.8%
	TeCA: 100.0%	TeCA: 90.0%	TeCA: 95.0%	TeCA: 95.0%
5	TrCA: 100.0%	TrCA: 91.1%	TrCA: 100%	TrCA: 97.0%
	TeCA: 85.0%	TeCA: 85.0%	TeCA: 90%*	TeCA: 86.7%

* Eight gene subsets are obtained at 100% TrCA. Their mean TeCA is 90%.

REFERENCES

[1] J. Kittler, Pattern Recognition and Signal Processing, Chapter Feature Set Search Algorithms Sijthoff and Noordhoff, Alphen aan den Rijn, Netherlands, pp. 41–60. 1978.

[2] M. Ben-Bassat, "Pattern recognition and reduction of dimensionality," In Krishnaiah,P. and Kanal,L., (eds.) Handbook of Statistics II, vol. 1. North-Holland, Amsterdam. pp. 773–791, 1982.

[3] W. Siedelecky and J. Sklansky, "On automatic feature selection", *International Journal of Pattern Recognition*, vol. 2, pp. 197-220, 1998.

[4] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield and E.S. Lander, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, pp. 531-537, 1999.

[5] I. Inza, P. Larrañaga, R. Etxeberria and B. Sierra, "Feature subset selection by Bayesian networks based Optimization," *Artificial Intelligence*, vol. 123, pp. 157–184, 2000.

[6] R.O. Duda, P.E. Hart and D.G. Stork, Pattern Classification, Wiley, New York, 2000.

[7] P. Pavlidis, J. Weston, J. Cai and W.N. Grundy, "Gene functional classification from heterogeneous data," *International Conference on Computational Biology*, pp. 249-255, 2001

[8] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C.H. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J.P. Mesirov, T. Poggio, W. Gerald, M. Loda, E.S. Lander and T.R. Golub, "Multiclass cancer diagnosis using tumor gene expression signatures," *Proc. Natl. Acad. Sci.*, vol. 98, no. 26, pp. 15149-15154, 2001.

[9] I. Guyon, J. Weston, S. Barnhill and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, pp. 46 389–422, 2002

[10] W. Pan, "A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments," *Bioinformatics*, vol. 18, pp. 546-554, 2002.

[11] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *J. Machine Learning Research*, vol. 5, pp. 1205–1224, 2004.

[12] G. Cong, K.-L. Tan, A.K.H. Tung, X. Xu, "Mining top-k covering rule groups for gene expression data," *In the ACM SIGMOD International Conference on Management of Data*, pp. 670-681, 2005.

[13] P. Jafari and F. Azuaje, "An assessment of recently published gene expression data analyses: reporting experimental design and statistical factors," *BMC Med. Inform. Decision Making*, 6, 27, 2006.

[14] H. Mamitsuka, "Selecting features in microarray classification using ROC curves," *Pattern Recognition*, vol. 39, pp. 2393–2404, 2006.

[15] M.W. Mark and S.Y. Kung, "A solution to the curse of dimensionality problem in pairwise scoring techniques," *International Conference on Neural Information Processing*, pp. 314-323, 2006.

[16] C. Zhang, X. Lu, and X. Zhang, "Significance of gene ranking for classification of microarray samples," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 3, no. 3, pp. 312–320, 2006.

[17] Y. Saeys, I. Inza and P. Larranaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[18] X. Wang and O. Gotoh, "Cancer classification using single genes", *Genome Informatics*, vol. 23, pp. 179-188, 2009.

[19] X. Lu, A. Gamst and R. Xu, "RDcurve: A non-parametric method to evaluate the stability of ranking procedures," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 7, no. 4, pp. 719-726, 2010.

[20] X.Cui, H. Zhao and J. Wilson, "Optimized ranking and selection methods for feature selection with application in microarray experiments," *Journal of Biopharmaceutical statistics*, vol. 20, no. 2, pp. 223-239, 2010.

[21] Y.-Q.Zhang and J.C. Rajapakse (as editors), "Machine learning in bioinformatics", Wiley Publication, 2009.

[22] J. Khan, J.S. Wei, M. Ringner, L.H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C.R. Antonescu, C. Peterson, and P.S. Meltzer, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural network," *Nature Medicine*, vol. 7, pp. 673-679, 2001.

[23] S.A.Armstrong, J.E. Staunton, L.B. Silverman, R. Pieters, M.L. den Boer, M.D. Minden, S.E. Sallan, E.S. Lander, T.R.Golub, and S.J. Korsemeyer, "MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia," *Nature Genetics*, vol. 30, pp. 41-47, 2002.

[24] D. Singh P.G. Febbo, K. Ross, D.G. Jackson, J. Manola, C. Ladd, P. Tamayo, A.A. Renshaw, A.V. D'Amico, J.P. Richie, E.S. Lander, M. Loda, P.W. Kantoff, T.R. Golub, W.R. Sellers, "Gene Expression Correlates of Clinical Prostate Cancer Behavior," *Cancer Cell*, vol. 1, pp 203-209, 2002.

[25] L. Tao, C. Zhang and M. Ogihara, "A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression," *Bioinformatics*, vol, 20, no. 14, pp. 2429-2437, 2004.

[26] Al-Shahrour, F., Díaz-Uriarte, R. and Dopazo, J., FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. Bioinformatics, 20, pp. 578-580, 2004.

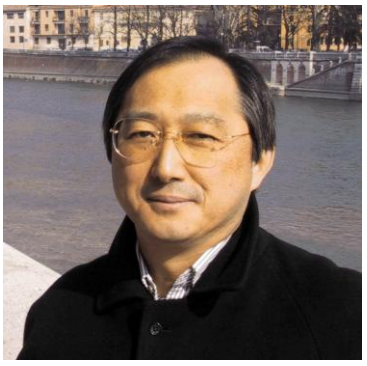
[27] H. Chai and C. Domeniconi, "An evaluation of gene selection methods for multi-class microarray data classification," *Proceedings of the Second European Workshop on Data Mining and Text Mining in Bioinformatics*, pp. 3-10, 2004.

[28] J. Li, and L. Wong, "Using rules to analyse bio-medical data: a comparison between C4.5 and PCL," *Advances in Web-Age Information Management*, Berlin/Heidelberg: Springer, pp. 254-265, 2003.

[29] A.C. Tan and D. Gilbert, "Ensemble machine learning on gene expression data for cancer classification," *Applied Bioinformatics*, vol. 2, 3 Suppl, pp. S75-83, 2003.

Alok Sharma received the BTech degree from the University of the South Pacific (USP), Suva, Fiji, in 2000 and the MEng degree, with an academic excellence award, and the PhD degree in the area of pattern recognition from Griffith University, Brisbane, Australia, in 2001 and 2006, respectively. He is currently a research fellow at the University of Tokyo. He is also with the Signal Processing Laboratory, Griffith University and the University of the South Pacific. He participated in various projects carried out in conjunction with Motorola (Sydney), Auslog Pty. Ltd. (Brisbane), CRC Micro Technology (Brisbane), and the French Embassy (Suva). His research interests include pattern recognition, computer security, and human cancer classification. He reviewed several articles from journals like IEEE Trans. NN, IEEE Trans. SMC, Part A: SH, IEEE Journal on STSP, IEEE Trans. KDE, IEEE Tans. EC, Computers & Security, Pattern Recognition, etc. He is a member of IEEE.

Seiya Imoto is currently an Associate Professor of Human Genome Center, Institute of Medical Science, University of Tokyo. He received the B.S., M.S., and Ph.D. degrees in mathematics from Kyushu University, Japan, in 1996, 1998 and 2001, respectively. His current research interests cover statistical analysis of high dimensional data by Bayesian approach, biomedical information analysis, microarray gene expression data analysis, gene network estimation and analysis, data assimilation in biological networks and computational drug target discovery.



Satoru Miyano is a Professor of Human Genome Center, Institute of Medical Science, University of Tokyo. He received the B.S., M.S. and Ph.D. degrees all in mathematics from Kyushu University, Japan, in 1977, 1979 and 1984, respectively. His research group is developing computational methods for inferring gene networks from microarray gene expression data and other biological data, e.g., protein-protein interactions, promoter sequences. The group also developed a software tool, Cell Illustrator, for modeling and simulation of various biological systems. Currently, his research group is intensively working for developing the molecular network model of lung cancer by time-course gene expression and proteome data. With these technical achievements, his research direction is now heading toward a creation of Systems Pharmacology.

He is Associate Editor of PLoS Computational Biology; IEEE/ACM Transactions on Computational Biology and Bioinformatics; and, Health Informatics. He was Associate Editor of Bioinformatics during 2002-2006 and 2007-2009. He is Editor of Journal of Bioinformatics and Computational Biology; Lecture Notes in Bioinformatics; Advances in Bioinformatics; Journal of Biomedicine and Biotechnology; International Journal of Bioinformatics Research and Applications (IJBRA); Immunome Research; Theoretical Computer Science; Transactions on Petri Nets and Other Models of Concurrency (ToPNoC); and, New Generation Computing. He is Editor-in-Chief of Genome Informatics. He is recipient of IBM Science Award (1994) and Sakai Special Award (1994).