Identifying Significant Genes from Microarray Data

Han-Yu Chuang[1,+], Hongfang Liu[2,*] , Stuart Brown[3], Cameron McMunn-Coffran[4,*], Cheng-Yan Kao[1,+], and D. Frank Hsu[4,+]

1. *Department of Computer Science and Information Engineering, National Taiwan University, Taipei, 106, Taiwan*
2. *Department of Biomedical Informatics, Columbia University, New York NY, 10032, USA*
3. *Research Computing Resources, School of Medicine, New York University, New York, NY 10016, USA*
4. *Department of Computer and Information Science, Fordham University, New York, NY 10023*

r90002@csie.ntu.edu.tw, hsu@cis.fordham.edu

Abstract

Microarray technology is a recent development in experimental molecular biology which can produce quantitative expression measurements for thousands of genes in a single, cellular mRNA sample. These many gene expression measurements form a composite profile of the sample, which can be used to differentiate samples from different classes such as tissue types or treatments. However, for the gene expression profile data obtained in a specific comparison, most likely only some of the genes will be differentially expressed between the classes, while many other genes have similar expression levels. Selecting a list of informative differential genes from these data is important for microarray data analysis. In this paper, we describe a framework for selecting informative genes, called Ranking and Combination analysis (RAC), which combines various existing informative gene selection methods. We conducted experiments using three data sets and six existing feature selection methods. The results show that the RAC framework is a robust and efficient approach to identify informative gene for microarray data. The combination approach on two selecting methods almost always performed better than the less efficient individual, and in many cases, better than both. More significantly, when considering all three data sets together, the combination approach, on average, outperforms each individual feature selection method. All of these indicate that RCA might be a viable and feasible approach for the microarray gene expression analysis.

1. Introduction

Microarray technology provides biomedical researchers the ability to measure expression levels of thousands of genes simultaneously [11, 12, 23, 31, 32, 35]. These measurements quantify the hybridization of cellular mRNA to cDNA or oligonucleotide probes which are immobilized on a solid substrate. Such gene expression profiles are used to understand the molecular variations among disease related cellular processes, and also to help the development of diagnostic tools and classification platforms in cancer research [1, 2, 14, 28, 29, 30].

The major challenge of microarray data analysis is the large number of genes compared to the small number of samples in a typical experiment. For the data obtained in a specific experiment, only some of the genes will be useful to differentiate samples among different classes, while many other genes are irrelevant to this task. Those irrelevant genes not only introduce unnecessary noise to microarray data analysis, but also increase the dimensionality; which results in computational difficulties in various other tasks such as clustering, classification, or construction of relevance networks [3, 27]. To eliminate those "probable noise" genes, the identification of informative genes, is a feature selection problem which is crucial in microarray data analysis [3, 5, 20, 38]. Moreover, isolating highly informative genes may reveal some insight into the pathomechanism and indicate ways to further interpret the data.

In the pattern recognition and machine learning literature [20, 21], the feature selection problem has received much attention, where one has class-labeled data and wants to figure out which features best discriminate among associated classes. In microarray data analysis, the

* Current Address:
HongFang Liu, Department of Information System, University of Maryland at Baltimore County, Baltimore, MD 21250, USA
Cameron McMunn-Coffran, Bio-Computing Research Center, Rockefeller University, New York, NY 10021

+ Research partially supported by the Ministry of Economic Affairs, R.O.C. under the title "Development of novel technology for diagnosis and treatment of angiogenesis-related disorders" and contact #: 91-EC –17–A–19–S1–0016.

problem is to select genes that clearly differentiate the classes and to drop genes with little or no impact. Different applicable techniques for feature selection can be divided into two main categories: the numerical combination and the method of ranking and scoring [9]. The numerical combination approaches simplify the complex data via finding some new features with values that are numerical combination of values of the original variables. One of the most popular numerical combination approaches is principal component analysis (PCA) [4, 40]. PCA finds a set of orthogonal principal components, which corresponds to the directions of maximum variances, for the purpose of reducing the dimensionality of the data matrix. However these numerical combination methods cannot discover which specific informative genes are responsible for the major trends observed in the data.

On the other hand, ranking and scoring methods score the discriminability of each gene based on its own expression patterns. Two major estimations of discriminability, parametric and nonparametric, have been proposed [9]. The parametric estimation approaches assess the discriminability of genes using different statistical analysis including the t-statistic, Fisher, or Golub criterion [14, 15, 34]. Parametric estimation depends on exact expression levels and the number of replicate samples. The nonparametric estimation approaches rank samples according to their expression levels and select genes according to a certain metric based on the disorder of classes in the ranked list. Examples of nonparametric estimation include TNoM, MDMR and WEPO [3, 9, 27]. Details of these parametric and nonparametric examples will be discussed later.

In our previous studies [9], we found that there is no single method which is always the best in every study. The outcomes of different methods may differ substantially. This discordance causes difficulties in the interpretation of the data. Moreover, it is unclear which method should be applied to new unknown data sets. However, the prevailing phenomenon implies that a gene is significantly worth further analyzing, if it is identified as an informative one in most common using methods. Thus, combining meaningful results from different methods seems to a reasonable way to study.

Recently, evidence combination and data fusion have been studied in a variety of different application domains such as information retrieval [25, 37], pattern recognition [39], and molecular similarity searching [13]. There are two different ways to combine evidence. One is based on ranks, and the other is based on scores. Hsu *et al.* [18] studied the behavior and relationship between rank combination and score combination. In particular, they introduced the important concept and parameter called **rank/score graph**. They then showed that under certain condition rank combination outperforms score combination.

For this paper, we examined three parametric and three nonparametric feature selection methods for identifying informative genes: t-Test [15], Fisher [34], Golub [14], TNoM [3], Wilcoxon rank sum test [27] and WEPO [9]. We then applied rank combination to combine different feature selection methods. For each one of these methods, we first introduce some underlying theory and the process of computation. Then, we present our comparison and combination study of these methods on three different publicly available datasets: Adenocarcinoma Ac data set [26], Lymphoma Lp data set [1], and Colon Cc cancer dataset [2]. We used two measures to evaluate the performance. One measure is the classification accuracy of each feature selection, where we used support vector machines (SVM) and leave-one-out cross-validation (LOOCV) [3, 19] to obtain the accuracy. The other measure is precision, where we used those known informative genes that have been confirmed by biomedical researchers. Details of each data set and each measure will be discussed in the experiment section.

The experiments showed that the performance of combining various approaches is competitive with these compared parametric and nonparametric methods. Genes selected by our combination approach are as informative as surveyed methods in both classification and biological interpretation with the added advantages of efficiency, flexibility, and adaptability. Moreover, there is a trend worth going deep into that combination of heterogeneous methods may achieve the best performance while the heterogeneity of methods is properly defined.

The rest of this paper is organized as follows: Section 2 discusses various feature selection methods considered in the paper. Section 3 describes our combination methods. Section 4 describes the experiments (using three datasets) for comparing different feature selection methods. Section 5 presents the result. Section 6 contains conclusions and the future research directions of our study.

2. Feature Selection Methods

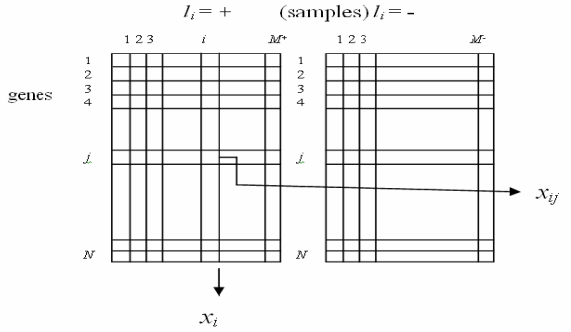

Figure 1. The illustration of *D*

Assume that a *gene expression data set D* on *N* genes for *M* mRNA samples consists of pairs $<x_i, l_i>$, for $i = 1, \ldots, M$. Each *sample* x_i can by summarized by a *N*-dimensional vector $(x_{i1}, \ldots, x_{iN})$, where x_{ij} describes the expression level of gene *j* in mRNA sample *i*. The *class label* l_i associated with x_i is either - or + (for simplicity, we focus on two label classifications). If there are missing values in the original data, we may use an EM-type algorithm [22] to impute the values so that the data, upon which we apply feature selection methods, contains no missing values. Figure 1 is the illustration of *D*.

2.1. Parametric approaches

The parametric approaches discriminately assess the genes by using different statistical criteria to estimate the degree of compactness between genes of the same class and the separation between genes of two different classes. These approaches score genes using some estimators of gene expression data, such as values of mean or standard deviation. The statistical criteria are based on the assumption that the data comes from some kind of distribution. Each parametric approach puts different weights in the variance and number of samples of the statistical criteria. The three parametric methods we discuss here -- t-Test, Fisher and Golub -- all consider a gene more informative when the corresponding score is larger.

2.1.1. t-Test (A)

The two-sample *t*-Test [15, 32] is used to determine if the means of two populations are equal. In microarray data analysis, the unpaired two-sample *t*-Test is often used since samples may be derived from different physical locations and may not have the same distribution.

In our study, t-Test gives the discriminative power of the k^{th} gene as

$$T(k) = \frac{|\mu_k^+ - \mu_k^-|}{\sqrt{\sigma_k^{+2}/(M^+ - 1) + \sigma_k^{-2}/(M^- - 1)}} \quad (1)$$

where M^+ and M^- are the sample sizes, μ_k^+ and μ_k^- are the sample means, and σ_k^{+2} and σ_k^{-2} are the sample variances of all x_{ik} with l_i = + or - , respectively.

2.1.2. Fisher (B)

Fisher criterion is a classical measure to assess the degree of separation between two classes [4, 34]. It is a t-Test-like statistic. The score for gene *k* is defined as

$$J(k) = \frac{(\mu_k^+ - \mu_k^-)^2}{\sigma_k^{+2} + \sigma_k^{-2}} \quad (2)$$

where μ_k^+ and μ_k^- are the sample means, and σ_k^{+2} and σ_k^{-2} are the sample variances of all x_{ik} with l_i = + or - .

2.1.3. Golub (C)

Golub and coworkers use a criterion similar to Fisher for their ALL/AML classification based on mRNA expression data [14]. The Golub score for the k^{th} gene is defined as

$$G(k) = \frac{|\mu_k^+ - \mu_k^-|}{\sigma_k^+ + \sigma_k^-} \quad (3)$$

where μ_k^+ and μ_k^- are sample means, and σ_k^+ and σ_k^- are sample standard deviations of all x_{ik} with l_i = + or - .

2.2. Nonparametric approaches

Compared to parametric approaches, nonparametric approaches use rankings obtained from expression level measurements associated with a specific gene rather than the measurements aimed at avoiding statistical instabilities. Usually, the nonparametric approaches rank samples associated with the same gene according to their expression level measurements and make punishments to the disorders that damage a perfect split of samples with different classes. The smaller the score a gene gets, the less punishment. In this paper, we survey three nonparametric methods – TNoM [3], MDMR [26] and WEPO [9] -- all consider a gene more informative when the corresponding score is smaller.

2.2.1. TNoM

Ben-Dor *et al* [3] proposed TNoM (Threshold Number of Misclassification) to score the given gene by searching for a simple decision rule corresponding to a given expression level to predict the class label of an unknown. A decision rule, sign ($ax+b$), is adapted to predict an unknown class, where *a* and *b* are parameters (Note that since only the sign of the linear expression matters, we can limit our attention to $a \in \{+1, -1\}$). TNoM looks to choose the values of *a* and *b* in order to minimize the number of errors:

$$Err(a, b \mid k) = \sum_i 1\{l_i \neq sign(a \times x_{ik} + b)\} \quad (4)$$

We can find the best values by exhaustively trying all $2(M+1)$ possible rules. (Attention is limited to threshold values that are mid-way points between actual expression values.)

The TNoM score of gene *k* is simply defined as:

$$TNoM(k) = \min_{a,b} Err(a, b \mid k) \quad (5)$$

2.2.2. MDMR (E)

The method Minimum Distance to Modal Ranking (MDMR) first ranks all the sample values of a gene and then computes the minimum distance between this ranking and a modal ranking. One example of the MDMR method is Kandal's τ distance [16].

Park *et al* [27] successfully a variant of Kandal's τ distance for this problem. It first sorts all sample pairs $<x_i, l_i>$ by x_{ik} in ascending order. At the same time, the corresponding classes (i.e., +'s and –'s) are rearranged accordingly, and the resulting class label sequence of l_i indicates the level of class disorder among the samples. A score is then defined as the minimum number of consecutive swaps needed to arrive to a perfect split sequence for the derived label sequence, where samples associated with the same class are grouped together in the

perfect split sequence. In this paper, we adopt the special MDMR method as used by Park *et al* [27]:

$$P(k) = \sum_{i;l_i=-} \sum_{j;l_j=+} h(x_{ik} - x_{jk}), h(x) = \begin{cases} 1, x > 0 \\ 0, x \le 0 \end{cases} \quad (6)$$

2.2.3. WEPO (F)

In TNoM and MDMR, genes with the same ordered expression data are regarded to have the same discriminative power. However, genes with the same TNoM or MDMR score may not have the same performance [9]. WEPO introduces z-score into the swapping ranking scheme to avoid loss of information [9].

For gene *k*, the expression levels of samples are first normalized by z-score to eliminate the problem of scaling. The z-score is:

$$z_{ik} = \frac{x_{ik} - \mu_k}{\sigma_k} \quad (7)$$

where μ_k is the sample mean and σ_k is the *mean absolute deviation* of gene *k.* Then the samples are sorted according to the normalized expression levels. The punished score of each gene is calculated by estimating the overlapped regions of two classes. The punishment is defined as:

$$WEPO(k) = \sum_{i;l_i=+} \sum_{j;l_j=-} \Psi(z_{ik} - z_{jk_i}), \Psi(z) = \begin{cases} |Z|, Z > 0 \\ 0, Z \le 0 \end{cases} \quad (8)$$

3. Combination method and Data fusion

In this section, we describe the combination approach we used. The approach, called Rank and Combination analysis (RAC), consists of two stages: the rank stage and the combination stage. The first stage (rank stage) is the process by which we rank the collection of all genes according to each of the selected features. In this regard, each of the genes in the collection is assigned a score (which can be a measurement of variance, deviation, correlation, or probability) depending on a specific feature. Sorting the collection of genes by their scores gives rise to a ranking of the genes in the dataset. The second stage (combination stage) is the process of combining the rank lists obtained from the first stage. Figure 2 illustrates the RAC architecture.

3.1. Rank Stage and Rank Space

Given a collection of genes G = {g_1, g_2, …, g_n}, each gene is assigned a score by its measurement according to the selected features. These features can be selected parametrically or non-parametrically. After every gene is assigned a score, the genes in G are ranked. A ranking of these *n* genes is considered a permutation of these *n* elements. Since the set of all permutations of *n* elements forms a group, called the symmetric group S_n, the rank space (the set of all possible rankings of G) is equal to S_n. Moreover, by defining a metric or suitably choosing a generating set S of the group, the group S_n give rise to a graph $G(S_n,S)$, called the Cayley graph of the group S_n with generating set S [17, 18]. By harnessing the group and graph structure of $G(S_n, S)$ and by studying the rank correlation, the combination stage of the RAC approach can be modeled in a dynamic fashion [18,24].

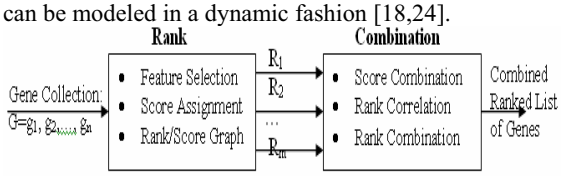

Figure 2. Rank and Combination (RAC) Architecture

3.2. Combination Stage

There are several different ways of combining the *m* rank lists that are generated according to different feature selections. Since each rank list consists of rank and score, there are both rank and score combinations. For the score combination, it would be a function (linear or non-linear) of the scores in each of the *m* feature rankings. The simplest case would be to take the weighted sum of the scores divided by *m* for each gene g_i. As for the rank combination, two schools of thought can be followed: consensus building and voting. Consensus building combines the ranks of a list by using a weighted sum from each of the component rankings. The voting method chooses a rank (e.g., maximum or minimum) from one or some of the *m* rankings.

In Figure 3, R_1, R_2 and R_3 are ranked gene lists from the rank stage using three different feature selection methods. R^* is the average linear combination of R1, R2 and R3. For more details, see Figure 4. Figure 4 depicts the procedure to obtain the combined rank list R*. First, the sum of ranks for each gene in R_1, R_2 and R_3 is averaged and represented (in step (a)). The scores $f(g_i)$'s are sorted in ascending order to form $S_f(n)$ (in step (b)) and the combined ranked list R^* is then formed (in step (c)).

n=Rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
$R_1(n)$	g_1	g_3	g_5	g_7	g_9	g_{11}	g_{13}	g_{15}	g_2	g_4	g_6	g_8	g_{10}	g_{12}	g_{14}	g_{16}
$R_2(n)$	g_2	g_5	g_8	g_{11}	g_{14}	g_1	g_4	g_7	g_{10}	g_{13}	g_{16}	g_3	g_6	g_9	g_{12}	g_{15}
$R_3(n)$	g_9	g_8	g_2	g_{12}	g_1	g_{10}	g_4	g_{11}	g_3	g_{14}	g_5	g_{13}	g_7	g_{16}	g_6	g_{15}
R*(n)	g_1	g_2	g_5	g_8	g_{11}	g_9	g_3	g_4	g_7	g_{10}	g_{13}	g_{14}	g_{12}	g_6	g_{15}	g_{16}

Figure 3. Rank Combinations R^*.

(a): $f(g_i)=\frac{1}{3}\sum_j R_j^{-1}(g_i)$																
g_i	G_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9	g_{10}	g_{11}	g_{12}	g_{13}	g_{14}	g_{15}	g_{16}
$f(g_i)$	4	4.33	7.66	8	5.33	13	8.33	5.66	6.66	9.33	6	11	9.66	10	13.33	13.66
(b): Sort $f(g_i)$ in ascending order																
n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
$S_f(n)$	4	4.33	5.33	5.66	6	6.66	7.66	8	8.33	9.33	9.66	10	11	13	13.33	13.66
(c): $R^*(n)=f^{-1}(S_f(n))$																
n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
$R^*(n)$	g_1	g_2	g_5	g_8	g_{11}	g_9	g_3	g_4	g_7	g_{10}	g_{13}	g_{14}	g_{12}	g_6	g_{15}	g_{16}

Figure 4. Procedure to Calculate R^*.

4. Experiment

We used two methods to evaluate the performance. One method measures the classification accuracy of each feature selection method, where we used support vector machines (SVM) and leave-one-out cross-validation (LOOCV) [3, 9, 19] to obtain the accuracy on a classification task. The other method uses weighted recall to measure how informative the selected genes are with respect to their established biological interpretations. We used these known informative genes that have previously been confirmed by biomedical researchers to compute weighted recall. In the following, we describe each measure in detail.

4.1. Classification Accuracy using SVM and LOOCV

One reasonable way to measure the performance of a feature selection method is to define a classification task and then to measure the classification accuracy of the task. In the experiment here, we used SVM as the classification algorithm and the leave-one-out –cross-validation (LOOCV) as the method to measure the performance [3]. Given a set with *n* class-labeled samples, the LOOCV method constructs *n* classifiers, where each classifier is trained on *n*-1 samples in the set, and is tested on the remaining sample. The classification accuracy is then the average accuracy of each classifier.

Support vector machine (SVM) is a machine learning algorithm proposed by Vladimir Vapnik and his co-workers (see eg. [36]). It is based on the Structural Risk Minimization principle from statistical learning theory, and was first introduced with a paper at the COLT 1992 conference[7]. SVM can be applied to different tasks such as regression, classification, and density estimation. In the following, we concentrate on the classification implementation of SVM. For details on SVM, see for examples [6, 36].

For a binary classification task, given a training set with *n* class-labeled samples, (x_1, y_1), (x_2, y_2), ..., (x_n, y_n), where x_i is a vector consisting of feature values that represent the i^{th} sample, and $y_i \in \{-1,+1\}$ indicates the class, an SVM classifier learns a linear decision rule, which is represented using a hyperplane R^N. The label of a previously unmarked sample x is determined by which side of the hyperplane x lies. The purpose of training the SVM is to find a hyperplane that has the maximum margin to separate two classes.

In the experiment, we used a SVM software LIBSVM developed by C.C. Chang and C.J. Lin [8].

4.2. Precision of known informative genes

A different way to evaluate a feature selection method is to measure the **precision** of known informative genes that were previously confirmed to be among the top selected genes. We defined the precision as $PR(R,G)=\frac{\sum_i^{|G|}\frac{i}{|R_i|}}{|G|}$, where R is a ranking list, R_i is the top n elements in the ranking which include i genes in G, and G is a list of known informative genes.

4.3 Description of Data Sets

In the following, we summarize the three data sets used in the experiment: Ac, Lp and Cc. Data set Ac was subject to both evaluation methods, while Lp and Cc were subject to the classification method.

Ac. Adenocarcinoma data set

The expression profile associated with this data set was collected by Notterman *et al* [26]. The Notterman team obtained 18 paired colon adenocarcinoma normal tissue samples from the Cooperative Human Tissue Network. The experiment was performed with the Human 6500 GeneChip Set (Affymetrix oligonucleotide array). The data set consists of 7457 genes and 18 paired samples, in which 18 are labeled "carcinoma" and 18 are labeled "normal". Additionally, Notterman *et al.* applied 4-fold relative expression to choose informative genes and 66 genes (1.78% of those detected) had been picked with significant difference between tumor tissue and the normal samples. 11 of them were confirmed by reverse transcription-PCR (RT-PCR), which were used to measure the weighted recall of each feature selection method.

Lp. Lymphoma data set

We used a subset of the original collection of 96 expression measurements reported by Alizadeh *et al.* [1] In the original data set, 46 samples were of diffused large b-cell lymphoma (DLBCL). The remaining samples are of 8 different types of tissues. Alizadeh et al clustered the DLBCL into two classes --- Germinal centre B-like DLBCL, and Activated B-like DLBCL. In our experiment we used gene expression measurements of 5635 genes and 40 samples which are composed of 19 G C B-like DLBCL and 21 Activated B-like DLBCL.

Cc. Colon cancer data set

This collection of expression measurements from colon biopsy samples was reported by Alon *et al* [2]. The data set consists of 62 samples of colon epithelial cells from colon-cancer patients, in which 38 samples are labeled "tumor" and 20 are labeled "normal". The "tumor" biopsies were collected from tumors, and the "normal" biopsies were collected from healthy parts of the colons of the same patients. By pathological examination, the final statuses of the biopsy samples were labeled. Gene expression levels in these 62 samples were measured using high density oligonucleotide arrays. Of the genes detected in these microarrays, 2000 genes were selected based on the confidence in the measured expression levels.

4.4. Methods of Feature Selection and Combination

For each data set, we considered six feature selection methods (A) – (F) (described in Section 2) to compute scores for each gene, and then derived six rankings. Each ranking was then combined with every other ranking using RAC, where we used average linear combination R^* to combine two rankings. In total, we acquired 21 rankings (6 associated with surveyed methods and 15 associated with combinations) for each data set. We used the first evaluation method described in Section 4.1 to compute the accuracy of classification using genes that appear among the top n for $n = 5, 10, \ldots, 160$. That is, we applied SVM and LOOCV to each data set. The accuracy was averaged from $5*2^k$ to $5*2^{k+1}$, for k = 0, 1, …, 5. (see Figure 5) Additionally, we used a measure called weighted accuracy to evaluate the performance,

$$WA(R,C) = \sum_i 2^{-(i+1)} \, Accuracy \quad (C, 5*2^i)$$

where $Accuracy(C,5*2^i)$ is the accuracy of classification task C using the top $5*2^i$ genes in the ranking R (defined in Sec. 4.2). For dataset Ac, we computed the precision for each of the 21 rankings (see Figue 7).

5. Results

Figure 5 demonstrates the relation of average accuracy with regard to the size of selected top genes for the six feature selection methods. Figure 5 shows the detail of average accuracy over three datasets in relation to the size of selected top genes for 21 rankings. Figure 6 provides the detail of weighted accuracy for each data set. Figure 7 shows the precision for each feature selection and combining ranking on dataset Ac.

In Figure 5, we see that the performance of each method corresponds to the size of selected top genes. Each of the six feature selection methods achieved its best performance when choosing the top 10 or 20 genes. The inclusion of more genes after the top 20 in the feature selection did not improve the performance. This observation is understandable since the inclusion of more genes with low ranks is likely to introduce some noise.

From Figure 6, we can see that the performance of a combination of two approaches is better than the worst case of each individual. The trend is even clearer in Figure 7. Moreover, the combination of WEPO and TNoM resulted in higher precision than each of them. All of the three nonparametric methods achieved better performance in the precision measure than the parametric ones. In Figure 8, the curves of parametric and nonparametric means stand apart from each other on the rank/score graph. Furthermore, the distance between WEPO and TNoM/ MDMR is longer than the one between TNoM and MDMR. In other words, WEPO's scoring scheme is more different to that of TNoM than MDMR. Figure 7 and 8 demonstrate that the combination of the heterogeneous and well-performing methods outperforms each individual.

Additionally, we found there is also no "super star" method considering combinations among different measures. When averaging on three data sets in Figure 6, the combination of TNoM and Golub had the best performance. However, the categories of TNOM and Golub are different. It also indicates that combination of heterogeneous methods may achieve the best performance while the heterogeneity of methods is proper defined. When considering each data set individually, the best performance achieved on Golub, the combination of MDMR and Fisher, and the combination of Golub and TNoM for data sets Ac, Lp and Cc respectively. The power of heterogeneous combination is also observed on each data set.

6. Discussion and Conclusion

We have demonstrated that our RAC framework is a robust and efficient approach to identifying informative genes for microarray data. From Figure 5, 6, and 7, it is clear that no single feature selection method performs effectively across different data sets (and experiments) in different application domains. Results obtained in this paper using our RAC framework shows that a combinatorial approach almost always performs better than the less efficient individual, and in many cases, better than both. More significantly, when considering all three data sets together, the combination approach, on average, outperforms each individual feature selection method. All of this evidence indicates that RAC is likely to be a viable

and feasible approach for microarray gene expression analysis on any dataset.

There are several other advantages of our combination methods for identifying informative genes from microarray data:

■ **Efficiency**: sorting a list of n genes with assigned scores takes n*log$n$ steps. Moreover, combination of m rank lists should take no more than m*n*log$n$ steps. Calculation in the RAC framework becomes simple and easy to understand. Selection of efficient and effective combination would facilitate fast process and operation.

■ **Flexibility**: the RAC framework allows feature selection to use parametric, nonparametric and other means. It also allows combination methods to use both rank and score combination. Moreover, rank combination allows individuals the choice of using consensus building or voting, while score combination facilitates the options of using various linear, non-linear, or weighted combinations. Compared to other methods such as clustering association, or self-organized maps, the RAC approach is more flexible as the outputs of both stages of RAC are rank lists which include either rank and score information for the collection of genes.

■ **Adaptability**: The RAC method can be adapted to different application domains which may call for different feature selections and different combination algorithms. One of our long-term goals is to construct a RAC system which can learn from the biological environments and biological phenomena in its application domain, and then evolve to become a more intelligent expert system in that particular domain.

In this paper, we described the framework and have taken up our investigation using average linear combination of rankings. Future work will explore other ways to combine different feature ranking methods. Since the RAC framework is efficient, flexible and adaptable, we will explore other combination metrics (see D.F. Hsu and A. Palumbo [17] for performing rank combination in Cayley graphs), or combine more than two feature selection methods in a static and dynamic fashion (see H.Y. Chuang *et al*. [10]). Furthermore, we will proceed to clarify how combination would achieve best performance, such as combining heterogeneous and well-performing methods observed in this work.

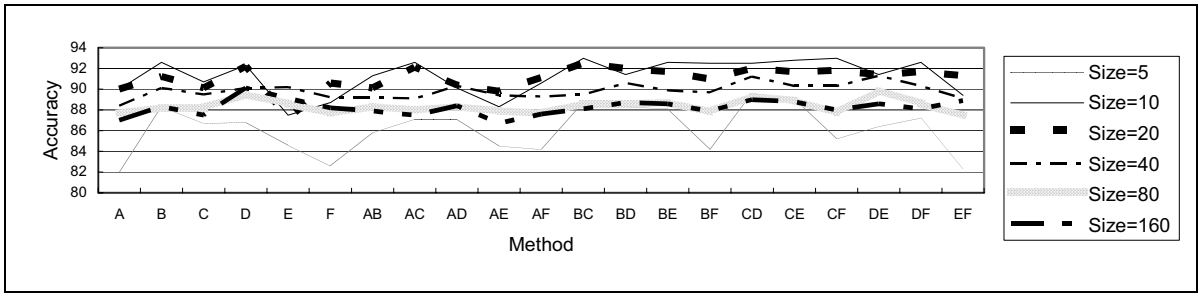

Figure 5. The accuracy of different numbers of selected top genes averaged over three data sets.

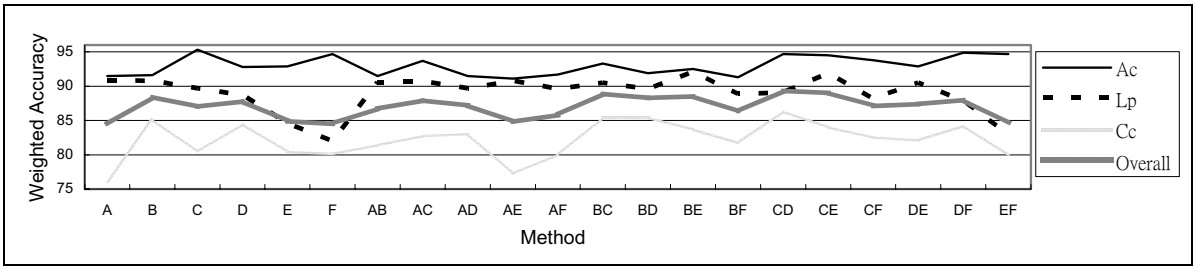

Figure 6. The weighted accuracy (%) for each feature selection and combining ranking.

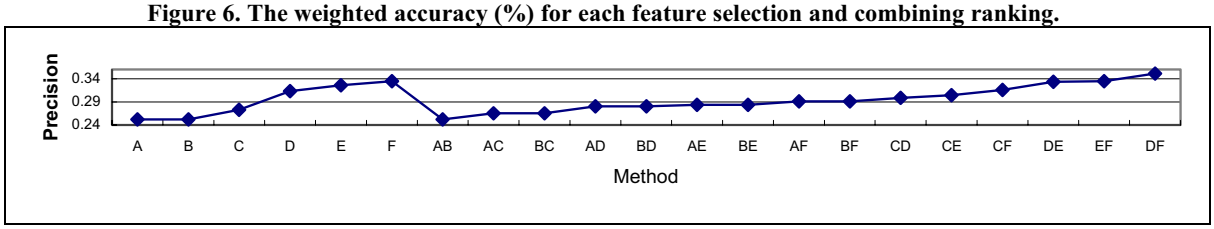

Figure 7. The precision for each feature selection and combining ranking on the Ac dataset.

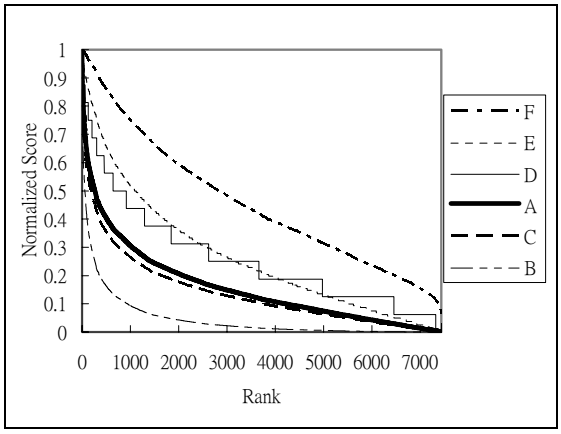

Figure 8. rank/score graph of six methods on Ac.

References

[1] A.A. Alizadeh *et al.*, "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling", *Nature*, 2000, vol. **403**, pp.503-511.

[2] U. Alon *et al.*, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays", *Proceedings of the National Academy of Sciences*, 1999, vol. **96**, pp. 6745-6750.

[3] A. Ben-Dor *et al.*, "Tissue Classification with Gene Expression Profiles", *Journal of Computational Biology*, 2000, vol. **7**, pp. 559-583.

[4] C. Bioshop, "Neural networks for pattern recognition", *Oxford University Press*, New York, 1995.

[5] A. Blum and P. Langley, "Selection of relevant features and examples in machine learning", *Artificial Intelligence*, 1997, vol. **97**, pp. 245-271.

[6] C.J.C. Burges, "A tutorial on Support Vector Machine for pattern recognition", *Data Mining and Knowledge Discovery*, 1998, vol. **2**, pp. 121-167.

[7] C. Cartes and V. Vapnik, "Support vector machines", *Machine Learning*, 1995, vol. **20**, pp. 273-297.

[8] C.C. Chang and C.J. Lin, "LIBSVM : a library for support vector machines", 2001, Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

[9] H.Y. Chuang, H.K. Tsai, Y.F. Tsai and C.Y. Kao, "Ranking genes for discriminability on microarray data", *Journal of Information Science and Engineering*, 2003, vol. **19**, pp. 953-966.

[10] H.Y. Chuang *et al.*, "Combination methods in microarray analysis", *Proceedings of I-SPAN' 04*, *IEEE CS Press,* 2004.

[11] L. DeRisi, V.R. Iyer, and P.O. Brown, "Exploring the metabolic and genetic control of gene expression on a genomic scale", *Science*, 1997, vol. **278**, pp. 680-685.

[12] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Bostein, "Clustering analysis and display of genome-wide expression patterns", *Proc. Natl. Acad. Sci*., 1998, vol. **95**, pp. 14863-14868.

[13] C.M.R. Ginn, P. Willett, and J. Bradshaw, "Combination of Molecular Similarity Measures Using Data Fusion", *Perspectives in Drug Discovery and Design*, 2000, vol. **20**, pp. 1-16.

[14] T. R. Golub *et al*., "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring", *Science*, 1999, vol. **286**, pp. 531-537.

[15] I. Hedenfalk *et al.*, "Gene-expression profiles in hereditary breast cancer", *New England J. Med.*, 2001, vol. **8**, pp. 344-539.

[16] T. P. Hettemansperger, "Statistical Inference based on ranks", *Wiley, New York*, 1984

[17] D. F. Hsu, and A. Palumbo, "A study of data fusion in Cayley graph $G(S_n, P_n)$", *Proceedings of I-SPAN'04, IEEE CS Press*, 2004.

[18] D.F. Hsu, J. Shapiro, and I. Taksa, "Methods of Data Fusion in Information Retreival: Rank vs. Score Combination", *DIMACS Technical Report 58*, 2002.

[19] J. Jaeger, R. Sengupta, and W.L. Ruzzo, "Improved gene selection for classification of microarrays", *Pacific Symposium on Biocomputing*, 2003, vol. **8**, pp. 53-64.

[20] R. Kohavi and G. John, "Wrapper for feature subset selection", *Artificial Intelligence*, 1979, vol. **97**, pp. 273-324.

[21] P. Langley, "Selection of relevant features in machine learning", *Proceedings of the AAAI Fall Symposium on Relevance, AAAI Press*, 1994.

[22] R.J.A. Little and D.B. Rubin, "Statistical analysis with missing data", *Wiley, New York*, 1987.

[23] D.J. Lockhart *et al.*, "Expression monitoring by hybridization to high-density oligonucleotide arrays", *Natrue Biotechnology*, 1996, vol. **14**, pp. 1675-1680.

[24] J.I. Marden, "Analysing and Modeling Rank Data", *Chapman & Hall*, 1995.

[25] K.B. Ng *et al.*, "Predicating the effectiveness of Naïve Data Fusion on the basis of system characteristics", *JASIS*, 2000, vol. **51**, pp. 1177-1189.

[26] D.A. Notterman, U. Alon, A.J. Sierk, and A.J. Levine, "Transcriptional Gene Expression Profiles of Colorectal Adenoma, Adenocarcinoma, and Normal Tissue Examined by Oligonucleotide Arrays", *Cancer Research*, 2001, vol. **61**, pp. 3124–3130.

[27] P.J. Park, M. Pagano, and M. Bonetti, "A Nonparametric Scoring Algorithm for Identifying Informative Genes from Microarray Data", *Pacific Symposium on Biocomputing*, 2001, vol. **6**, pp. 52-63.

[28] C.M. Perou *et al*., "Distinctive gene expression patterns in human mammary epithelial cells and breast cancers", *Proc. Natl. Acad. Sci.*, 1999, vol. **96**, pp. 9212-9217.

[29] R. Pollack *et al*., "Genome-wide analysis of DNA copy-number changes using cDNA microarrays", *Nature Genetics*, 1999, vol. **23**, pp. 41-46.

[30] D.T. Ross *et al*., "Systematic variation in gene expression patterns in human cancer cell lines", *Nature Genetics*, 2000, vol. **24**, pp. 227-234.

[31] M. Schena, D. Shalon, R.W. Davis, and P.O. Brown, "Quantitative monitoring of gene expression patterns with a complementary DNA microarrays", *Science*, 1995, vol. **270**, pp. 467-470.

[32] M. Schena, editor. "DNA Microarrays : A Pratical Approach", *Oxford University Press*, 1999

[33]G.W. Snedecor and W.G. Cochran, "Statistical Methods", Eighth Edition, *Iowa State University Press*, 1989.

[34] S. F. Terrence *et al.*, "Support Vector Machine Classification and Validation of Cancer Tissue Samples Using Microarray Expression Data", *Bioinformatics*, 2000, vol. **16**, pp. 906-914.

[35] "The Chipping Forecast", *Supplement to Nature Genetics*, vol. 21, 1999.

[36] V.G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response", *Proc Natl Acad Sci U S A*, 2001, vol. 98(9), pp. 5116-5121.

[37] C.C. Vogt, and G.W. Cotrell, "Fusion via a linear combination of scores", *Info. Ret*., 1999, vol. **1**, pp. 151-172.

[38] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio and V. Vapnik, "Feature selection for SVMs", *In Advances in Neural Information Processing Systems, MIT Press,* 2001, vol. 13.

[39] L. Xu, A. Krzyzak, and C.Y. Suen, "Method of Combining Multiple Classifiers and their Application to Handwriting Recognition", *IEEE Trans SMC*, 1992, vol. **22**, pp. 418-435.

[40] K.Y. Yeung and W.L. Ruzzo, "Principal component analysis for clustering gene expression data", *Bioinformatics*, 2001, vol. **17**, pp. 763-774.