Hindawi Publishing Corporation
EURASIP Journal on Bioinformatics and Systems Biology
Volume 2007, Article ID 64628, 15 pages
doi:10.1155/2007/64628

Research Article

Gene Selection for Multiclass Prediction by Weighted Fisher Criterion

Jianhua Xuan,[1] Yue Wang,[1] Yibin Dong,[1] Yuanjian Feng,[1] Bin Wang,[1] Javed Khan,[2] Maria Bakay,[3] Zuyi Wang,[1,3] Lauren Pachman,[4] Sara Winokur,[5] Yi-Wen Chen,[3] Robert Clarke,[6] and Eric Hoffman[3]

[1] *Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, Arlington, VA 22203, USA*
[2] *Department of Pediatric Oncology, National Cancer Institute, Gaithersburg, MD 20877, USA*
[3] *Research Center for Genetic Medicine, Children's National Medical Center, Washington, DC 20010, USA*
[4] *Disease Pathogenesis Program, Children's Memorial Research Center, Chicago, IL 60614, USA*
[5] *Department of Biological Chemistry, University of California, Irvine, CA 92697, USA*
[6] *Lombardi Cancer Center, Georgetown University, Washington, DC 20007, USA*

Received 30 August 2006; Revised 16 December 2006; Accepted 20 March 2007

Recommended by Debashis Ghosh

Gene expression profiling has been widely used to study molecular signatures of many diseases and to develop molecular diagnostics for disease prediction. Gene selection, as an important step for improved diagnostics, screens tens of thousands of genes and identifies a small subset that discriminates between disease types. A two-step gene selection method is proposed to identify informative gene subsets for accurate classification of multiclass phenotypes. In the first step, individually discriminatory genes (IDGs) are identified by using one-dimensional weighted Fisher criterion (wFC). In the second step, jointly discriminatory genes (JDGs) are selected by sequential search methods, based on their joint class separability measured by multidimensional weighted Fisher criterion (wFC). The performance of the selected gene subsets for multiclass prediction is evaluated by artificial neural networks (ANNs) and/or support vector machines (SVMs). By applying the proposed IDG/JDG approach to two microarray studies, that is, small round blue cell tumors (SRBCTs) and muscular dystrophies (MDs), we successfully identified a much smaller yet efficient set of JDGs for diagnosing SRBCTs and MDs with high prediction accuracies (96.9% for SRBCTs and 92.3% for MDs, resp.). These experimental results demonstrated that the two-step gene selection method is able to identify a subset of highly discriminative genes for improved multiclass prediction.

1. INTRODUCTION

Molecular analysis of clinical heterogeneity in cancer diagnosis and treatment has been difficult in part because it has historically relied on specific biological insights or has focused on particular genes with known functions, rather than systematic and unbiased approaches for recognizing tumor subtypes and associated biomarkers [1–3]. The development of gene expression microarrays provides an opportunity to take a genome-wide approach to classify cancer subtypes [2] and to predict therapy outcome [3]. By surveying mRNA expression levels for thousands of genes in a single experiment, it is now possible to read the molecular signature of an individual patient's tumor. When the signature is analyzed with computer algorithms, new classes of cancer that transcend distinctions based on histological appearance alone emerge, and new insights into disease mechanisms that move beyond classification or prediction emerge [4].

Although such global views are likely to reveal previously unrecognized patterns of gene regulation and generate new hypotheses warranting further study, widespread use of microarray profiling methods is limited by the need for further technology developments, particularly computational bioinformatics tools not previously included by the instruments. One of the major challenges is the so-called "the curse of dimensionality" mainly due to small sample size (10–100 in a typical microarray study) as compared to large number of features (often $\geq$30 000 genes). Most commonly used classifiers suffer from such a "peaking phenomenon," in that too many features actually degrade the generalizable performance of a classifier [5]. The detrimental impact of small sample size effect on statistical pattern recognition has led

to a series of valuable recommendations for classifier designs [6, 7].

Feature selection has been widely used to alleviate the curse of dimensionality; the goal being to select a subset of features that assures the generalizable yet lowest classification error [5, 8]. Feature selection may be done through an exhaustive search in which all possible subsets of fixed size are examined so that a subset with the smallest classification error is selected [5, 8]. A more elegant yet efficient approach is based on sequential suboptimal search methods [9, 10]; the sequential forward floating selection (SFFS) and sequential backward floating selection (SBFS) [10] are among the most popular methods.

Several studies on gene selection for the molecular classification of diseases using gene expression profiles have been reported [2, 11–13]. For example, Golub et al. used signal-to-noise ratio (SNR) to select informative genes for the two-class prediction problem of distinguishing acute lymphoblastic leukemia (ALL) from acute myeloid leukemia (AML) [2]. Khan et al. were the first to use an ANN-classifier approach to select a subset of genes for the multiclass prediction of small round blue cell tumors (SRBCTs) [12]. A sensitivity analysis of ANN's input-output relations was applied and identified a relatively large set of genes (96 genes). Dudoit et al. extended the two-class SNR method for multiple classes using the ratio of their between-group to within-group sums of squares, which is essentially a form of one-dimensional Fisher criterion [11]. Tibshirani et al. proposed a much simpler method, namely, the "nearest shrunken centroid" method, for SRBCTs classification, where a smaller gene set (43 genes) achieved comparable classification performance [14]. The work most closely related to our approach was reported by Xiong et al. [15]. They argued that a collection of individually discriminatory genes may not be the most efficient subset, and considered the joint discriminant power of genes. Specifically, they used Fisher criterion (FC) and sequential search methods to identify the biomarkers for the diagnosis and treatment of colon cancer and breast cancer [15]. However, it has been shown that when there are more than two classes, the conventional FC that uses the squared Mahanalobis distance between the classes is suboptimal in the dimension-reduced subspace for class prediction. Specifically, large between-cluster distances are overemphasized by FC and the resulting subspace preserves the distances of already well-separated classes, causing a large overlap of neighboring classes [16, 17].

In this paper, we propose to use weighted Fisher criterion (wFC) to select a suboptimal set of genes for multiclass prediction, since wFC (suboptimally) measures the separability of clusters and approximates, most closely, the true mean Bayes prediction error [17]. The weighting function in wFC criterion is mathematically deduced in such a way that the contribution of each class pair depends on the Bayes error rate between the classes. Thus, the wFC criterion deemphasizes the contribution from well-separated classes, while emphasizing the contribution from neighboring classes. A two-step feature selection is then conducted: (1) individually discriminatory genes (IDGs) are first selected by one-dimensional wFC; and (2) sequential floating search methods are then followed to select jointly discriminatory genes (JDGs) measured by wFC. The proposed two-step procedure is applied to two data sets—(1) NCI's SRBCTs and (2) CNMC's muscular dystrophies—to demonstrate its ability in obtaining improved diagnostics for multiclass prediction.

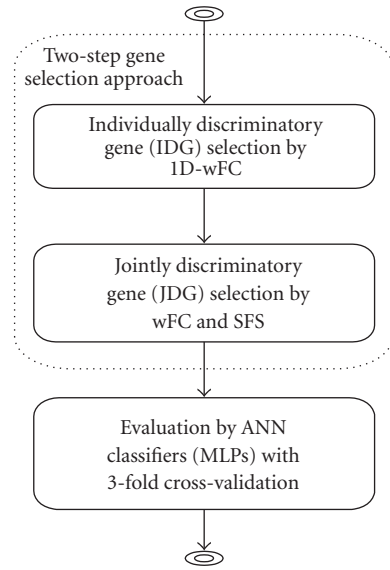

Figure 1: Block diagram of the two-step feature selection approach.

2. METHODS

In an attempt to improve class prediction, we propose a two-step feature selection approach by combining wFC and the sequential floating search (SFS) method. Figure 1 illustrates the conceptual approach. IDG selection is performed first to identify an initial gene subset of reasonable size (usually 50–200) under the wFC criterion. An SFS procedure is then conducted to refine the gene subsets of varying size according to the corresponding joint discriminant power. Finally, two popular types of classifiers, multilayer perceptrons (MLPs) and support vector machines (SVMs), are constructed to estimate the classification accuracy when using the selected gene sets, where the optimal gene subset corresponds to the smallest classification error.

2.1. IDG selection by wFC

Gene g_i (i is an index to a particular gene ID, $i = 1, \ldots, N$) will be selected as an individually discriminatory gene (IDG) if its discriminant power across all clusters, measured by one-dimensional wFC (1D wFC),

$$J_{\mathrm{IDG}}(g_i) = \frac{\sum_{k=1}^{K_0-1} \sum_{l=k+1}^{K_0} p_k p_l \omega(\Delta_{i,kl}) \left(\mu_{i,k} - \mu_{i,l}\right)^2}{\sum_{k=1}^{K_0} p_k \sigma_{i,k}^2}, \quad (1)$$

is above an empirically determined threshold, where K_0 is the number of clusters, p_k is the priori probability of class k, and $\mu_{i,k}$ is the mean expression level of gene g_i in class k, with

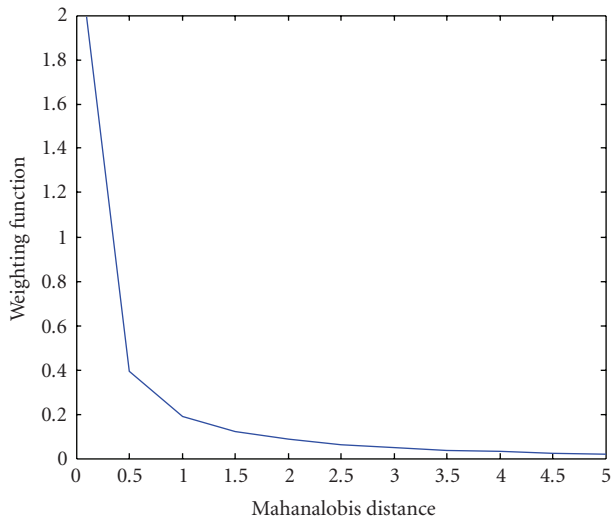

Figure 2: Weighting function to deemphasize the well-separated (distant) classes while emphasizing the neighboring (close) classes.

corresponding standard deviations $\sigma_{i,k}$. The weighting function, $\omega(\Delta_{i,kl})$, is designed to give more weight to the proximate cluster pairs. Figure 2 depicts the $\omega(\Delta_{i,kl})$ function that is defined in the following form [17]:

$$\omega(\Delta_{i,kl}) = \frac{1}{2\Delta_{i,kl}^2}\,\mathrm{erf}\left(\frac{\Delta_{i,kl}}{2\sqrt{2}}\right), \tag{2}$$

where $\Delta_{i,kl}$ is gene g_i's Mahanalobis distance between classes k and l, defined by

$$\Delta_{i,kl} = \frac{|\mu_{i,k} - \mu_{i,l}|}{\sqrt{\sum_{k=1}^{K_0} p_k \sigma_{i,k}^2}}. \tag{3}$$

Accordingly, we rank genes by $J_{\mathrm{IDG}}(g_i)$, $i = 1, \ldots, N$, and select the top M genes as the initial IDGs.

2.2. JDG selection by wFC and SFS

Class separability measured by wFC

JDGs refer to the gene subset whose joint discriminatory power is maximum among all the subsets of the same size selected from a gene pool (e.g., IDGs). The key is the consideration of the correlation (or dependence) between genes and their joint discriminatory power. We use multidimensional wFC as the measure of the class separability in JDG selection. The wFC of JDG can be defined by [17]

$$J(\mathrm{JDG}) = \sum_{k=1}^{K_0-1} \sum_{l=k+1}^{K_0} p_k p_l \omega(\Delta_{kl})\,\mathrm{trace}\left(\mathbf{S}_w^{-1}\mathbf{S}_{kl}\right), \tag{4}$$

where $\mathbf{S}_w = \sum_{k=1}^{K_0} p_k \mathbf{S}_k$ is the pooled within-cluster scatter matrix, and $\mathbf{S}_{kl} = (\mathbf{m}_k - \mathbf{m}_l)(\mathbf{m}_k - \mathbf{m}_l)^T$ is the between-cluster scatter matrix for classes k and l; $\mathbf{m}_k$ and $\mathbf{S}_k$ are the mean vector and within-class covariance matrix of class k, respectively. The weighting function, $\omega(\Delta_{kl})$, is defined as

$$\omega(\Delta_{kl}) = \frac{1}{2\Delta_{kl}^2}\,\mathrm{erf}\left(\frac{\Delta_{kl}}{2\sqrt{2}}\right), \tag{5}$$

and $\Delta_{kl} = \sqrt{(\mathbf{m}_k - \mathbf{m}_l)\mathbf{S}_w^{-1}(\mathbf{m}_k - \mathbf{m}_l)}$ is the Mahalanobis distance between classes k and l [17]. Note that when the number of samples is less than the number of genes as in many gene expression profiling studies, the pooled within-cluster scatter matrix $\mathbf{S}_w$ in wFC is likely to be singular, hence resulting in a numerical problem in calculating $\mathbf{S}_w^{-1}$. There are two possible remedies: the first one is to use pseudoinverse instead [18, 19], as originally implemented in [17]; the second one is to use singular value decomposition (SVD) method. Practically, we can set those very small singular values (say $<10^{-5}$) to a predefined small value (like 10^{-5}) for calculating $\mathbf{S}_w^{-1}$. The second method was used in our implementation of the algorithm due to its consistent performance as demonstrated in our experiments.

JDG selection by SFS methods

Optimal selection methods such as exhaustive search or the Branch-and-Bound method [20] are not practical for very high-dimensional problems such as those that include expression profiling studies. Thus, we will consider alternative suboptimal methods such as sequential search methods known as sequential backward selection (SBS) [21] and its counterpart sequential forward selection [22]. Both suffer from the so-called "nesting effect" that manifests itself explicitly as follows: (1) in case of SBS the discarded features cannot be reselected, and (2) in case of sequential forward selection the features once selected cannot be later discarded. The plus-l-minus-r method was the first method handling the nesting-effect problem [23]. According to one comparative study [6], the most effectively known suboptimal methods are the sequential floating search (SFS) methods [10]. In comparison to the plus-l-minus-r method, the "floating" search addresses the "nesting problem" without a need to specify any parameters such as l or r. The number of forward (adding)/backward (removing) steps is determined dynamically to maximize the criterion function.

We use SFS search methods to find the subset genes. As an example, the SBFS search algorithm is illustrated in Figure 3, where a floating search step, called conditionally including step (CIS), is followed after excluding a feature from the current feature set. The CIS is designed to search for the possibly "best" features from the excluded feature set. In the implementation, the CIS checks whether the updated feature set could offer any performance improvement in terms of the cost function. If improved, CIS will keep searching for the next "best" feature from the excluded feature set, otherwise it will return to exclude the next feature. The steps involved in the SBFS algorithm can be summarized as follows.

Step 1. Exclude the least significant feature from the current subset of size k. Let $k = k - 1$.

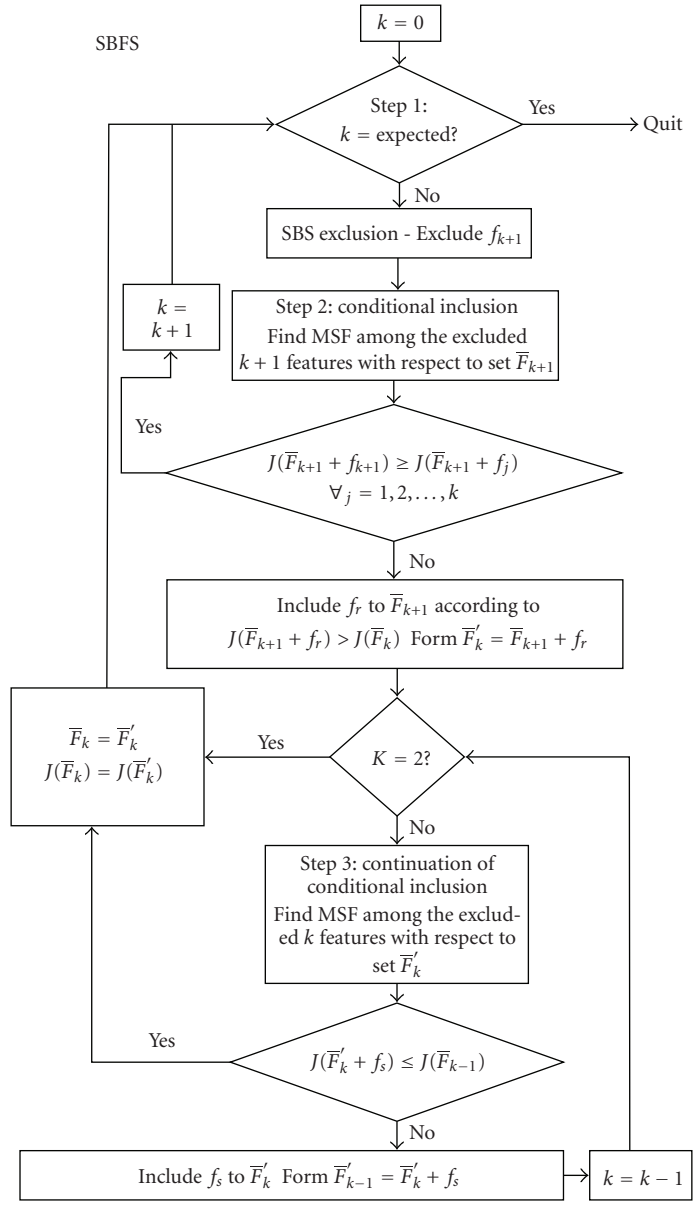

Figure 3: Block diagram of the SBFS search algorithm.

Step 2. Conditionally include the most significant feature from the excluded features.

Step 3. If the current subset is the best subset of size k found so far, let $k = k + 1$ and go to Step 2. Else return the conditionally included feature and go to Step 1.

In the above SBFS algorithm, we say that feature f_j from the set F_k is

(1) the most significant (best) feature in the set F_k if

$$J(F_k - f_j) = \min_{1 \leq i \leq k} J(F_k - f_i); \tag{6}$$

(2) the least significant (worst) feature in the set F_k if

$$J(F_k - f_j) = \max_{1 \leq i \leq k} J(F_k - f_i). \tag{7}$$

The search algorithm stops when the desired number of features is reached. A more detailed description of SBFS and SFFS algorithms can be found in [10].

2.3. Classification by MLPs and SVMs

Once selected, the JDGs are fed into neural networks (in particular, multilayer perceptrons (MLPs)) and/or SVMs for performance evaluation (see Figure 4). MLPs have been successfully applied to solve a variety of nonlinear classification problems [24]. In our experiments, we use three-layer perceptrons where the computation nodes (neurons) in the hidden layer enable the network to extract meaningful features from the input patterns for better nonlinear classification. The connectivity weights are trained in a supervised manner

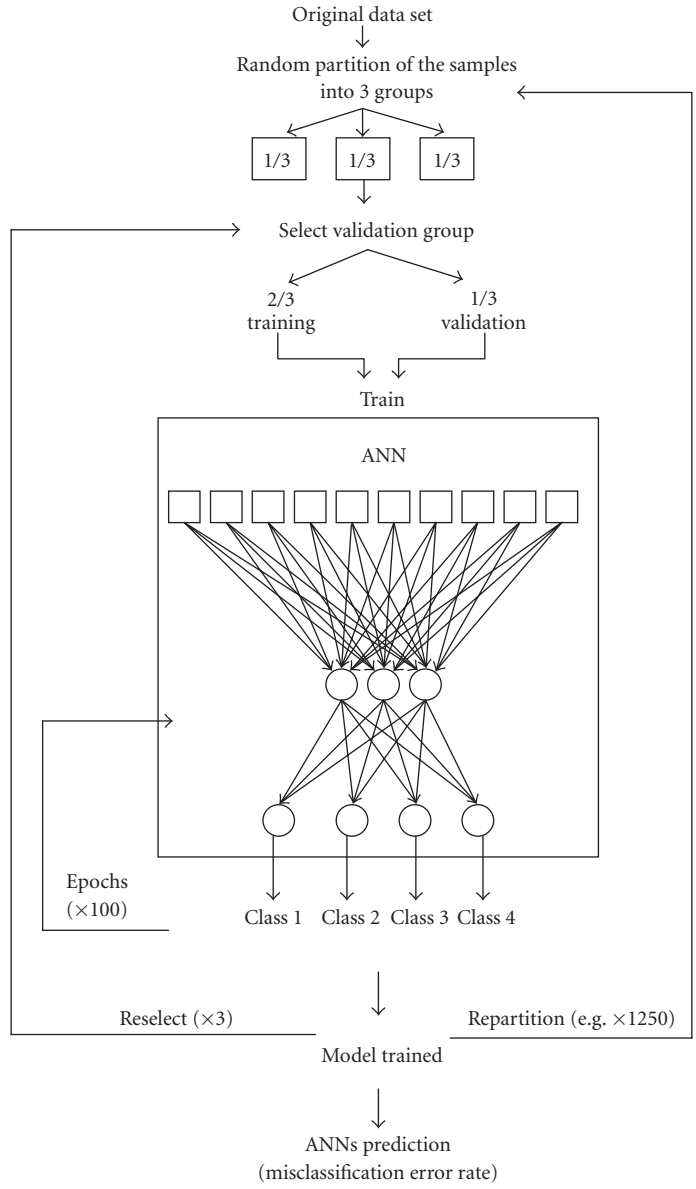

Figure 4: Evaluation by MLP Classifiers with 3-fold cross-validation.

using the error back propagation algorithm [24]. Recently, SVMs have also been applied to multiclass prediction of tumors using gene expression profiles [25, 26]. To overcome the limitation of SVM being a binary classifier in nature, Ramaswamy et al. used a one-versus-all (OVA) approach to achieve multiclass prediction. Each OVA SVM is trained to maximize the distance between a hyperplane and the closest samples to the hyperplane from the two classes in consideration. Given m classes hence m OVA SVM classifiers, a new sample takes the class of the classifier with the largest real-valued output, that is, class $= \arg\max_{i=1\cdots m} f_i$, where f_i is the real-valued output of the ith OVA SVM classifier (one of m OVA SVM classifiers). As reported in [25, 26], it seems that SVMs could provide a better generalizable performance for class prediction in high-dimensional feature space.

To estimate the accuracy of a predictor for future samples, the current set of samples was partitioned into a training set and a separate test set. The test set emulates the set of future samples for which class labels are to be predicted. Consequently, the test samples cannot be used in any way for the development of the prediction model. This method of estimating the accuracy of future prediction is the so-called split-sample method. Cross-validation is an alternative to the split-sample method of estimating prediction accuracy. Several forms of cross-validation exist including leave-one-out (LOO) and k-fold cross-validation. In this paper, we use 3-fold cross-validation (CV) and/or 10-fold CV to estimate the prediction accuracy of the classifiers; if the number of samples is large enough, we will estimate the prediction accuracy using blind test samples.

Table 1: A summary of CNMC's MD data (CNMC, 2003).

Class Number	Type of muscular dystrophy	Number of samples
1	BMD—Becker muscular dystrophy (hypomorphic for dystrophin)	5
2	DMD—Duchenne muscular dystrophy	10
3	Dysferlin—Dysferlin deficiency; also called Limb-girdle muscular dystrophy 2B (LGMD 2B)	10
4	FSHD—Fascioscapulohumeral dystrophy	14

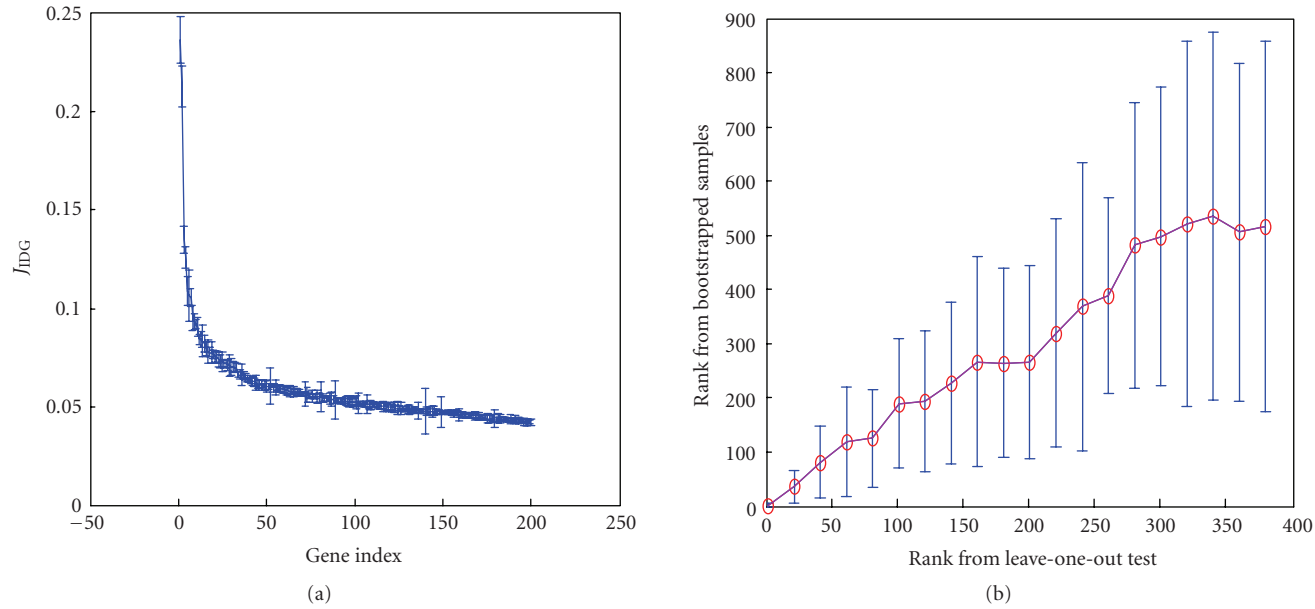

Figure 5: IDG selection with 1D wFC: (a) mean and standard deviation of 1D wFC by leave-one-out (LOO) test, and (b) a stability analysis of ranking using bootstrapped samples with 20,000 trials (error bar indicates the standard deviation of the rank from bootstrapped samples).

For 3-fold cross-validation, the classifiers are trained with 2/3 of the samples and tested on the remaining 1/3 of the samples for misclassification error calculation (see Figure 4). The procedures are repeated many shuffle times (e.g., 1000 times) of samples to split data set into a training data set and a testing data set. The overall misclassification error rate is calculated as the mean performance from all the trials.

3. RESULTS

We applied our two-step feature selection approach to two gene expression profiling studies: (1) NCI's data set of small round blue cell tumors (SRBCTs) of childhood [12], and (2) CNMC's data set of muscular dystrophies (MDs) [27]. The SRBCT data, consisting of expression measurements on 2,308 genes, were obtained from glass-slide cDNA microarrays, prepared according to the standard National Human Genome Research Institute protocol. The tumors are classified as one of four subtypes—(1) Burkitt lymphoma (BL), (2) Ewing sarcoma (EWS), (3) neuroblastoma (NB), and (4) rhabdomyosarcoma (RMS). A total of 63 training samples and 25 test samples are provided, although 5 of the latter are not SRBCTs. The CNMC's MD data were acquired from Affymetrix's GeneChip (U133A) microarrays with a total of 39 sample arrays, consisting of expression measurements on 11,252 genes [28]. The gene expression profiles were obtained using Affymetrix's MAS 5.0 probe set interpretation algorithms [29]. Samples are clinically classified as either one of the four types of muscular dystrophy. Table 1 gives a summary of the four classes in this study and the number of samples in each class.

3.1. NCI's SRBCTs

The two-step gene selection procedure was performed on 63 expression profiles of NCI's SRBCTs to identify a subset of genes. First, IDG gene selection was performed on 63 training samples to identify the top ranked genes. The individual discriminant power of each gene was calculated by 1D wFC. To assess the bias and variance of 1D wFC measurement, we performed leave-one-out trials on the data set. In Figure 5(a), we show the mean and standard deviation of the 1D wFC measurement. Additional material (Table S1) lists the top 200 IDGs with gene names and descriptions, which is available online at our website (http://www.cbil.ece.vt.edu). In this study, we ranked the IDGs according to the mean

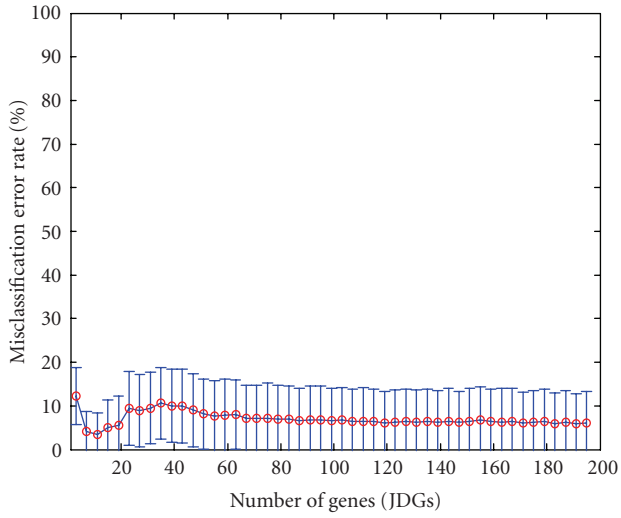

Figure 6: JDGs' prediction performance (NCI's SRBCTs): Misclassification error rates calculated by MLPs with 3-fold cross-validation; error bar indicates the standard deviation.

of the 1D wFC measurement, and obtained 200 top ranked IDGs to start the second step (i.e., the JDG selection step). Note that the number 200 was somehow chosen empirically, however, several considerations was taken into account. First, the performance of these 200 IDGs were evaluated by the classifiers to make sure that the MCER was reasonably small (indicating the genes' discriminatory power); second, in addition to using leave-one-out test to assess the bias and variance of 1D wFC measurement, a stability analysis of ranking using bootstrapped samples was also conducted to support the choice of the IDG number. In this experiment, 20 000 trials of bootstrapping were used to estimate the mean and standard deviation of the rank. In Figure 5(b), we plotted out the rank estimated from leave-one-out test versus the rank estimated from the bootstrapping trails as well as the standard deviation. Although such plot could not give us a definite cut-off value to determine the number of IDGs, the standard deviation of the rank estimated from bootstrapping trials showed relatively smaller variations for 100 top ranked IDGs than those for the genes ranked after 250. It is worth mentioning that the so-called "neighborhood analysis" described in [2] could also be used to tackle this problem with the help of a permutation test method. We have not tried this approach yet due to that according to our limited knowledge, there exist some difficulties in handing multiple classes and unbalanced samples.

JDG selection was performed to select best JDGs from the 200 IDGs. We used the SBFS method to select the best JDGs for each given number of features (in this case, the number of JDGs is from 1 to 199). The prediction performance of the JDG sets was evaluated by ANN classifiers (MLPs) using misclassification error. The MLPs comprised one hidden layer with 3 hidden nodes and the misclassification error was calculated by 3-fold cross-validation with 1,250 shuffles. Figure 6 shows the misclassification error rate (MCER) with respect to the selected JDGs, The best prediction performance was obtained when the number of JDGs was 9 in that MCER = 3.10%. Table 2 shows the image IDs, gene symbols, and gene names of the selected 9 JDGs. Figure 7 shows the expression pattern of 63 samples in the gene space of the newly selected 9 JDGs.

As a comparison, we compared the prediction performance of our 9 JDGs with that of two other approaches: (1) 96 genes selected by ANN-classifiers [12], and (2) 43 genes selected by a shrunken centroid method [14]. Among these three sets of genes, we found the following three genes to be in common: FGFR4 (ImageID:784224), FCGRT (ImageID:770394), and IGF2 (ImageID:207274). The following six genes are shared between our 9 JDGs and the 96 genes selected by ANNs: FGFR4 (ImageID:784224), FCGRT (ImageID:770394), PRKAR2B (ImageID:609663), MAP1B (ImageID:629896), IGF2 (ImageID:207274), and SELENBP1 (ImageID:80338). The following three genes are shared between our 9 JDGs with the 43 genes selected by the shrunken centroid method: FGFR4 (ImageID:784224), FCGRT (ImageID:770394), and IGF2 (ImageID:207274). We used MLPs (with one hidden layer) to evaluate the prediction performance of these three-gene sets; Table 3 shows the comparison of these three gene lists in terms of their misclassification error rates. Due to that, two different folds were used to estimate the prediction performance (3-fold for the 96 genes selected by ANN-classifiers, and 10-fold for the 43 genes selected by a shrunken centroid method), we have conducted both 3-fold CV and 10-fold CV to estimate the performance of 9 JDGs. The MCERs of 9 JDGs were 3.10% from 3-fold CV and 2.24% from 10-fold CV, respectively. From this equal-footing comparison, it seemed to suggest that our gene selection method successfully found a gene list (9 genes, a much smaller discriminant subset than that selected by either ANN or the shrunken centroid method) with excellent classification performance. It is worthy noting that, although cross-validation is a proven method for generalizable performance estimation, different folds used in CV may result in biased estimations of the true prediction performance. From our experience, leave-one-out CV or 10-fold CV tends to offer an "over promising" performance (i.e., a much lower misclassification error rate) compared to that of 3-fold CV. If the number of samples is large enough, we would suggest using 3-fold CV together with a test on an independent data set for performance estimation.

In addition to the above comparison, we have also compared our method with Dudoit's method (based on one-dimensional Fisher criterion) on gene selection for multiclass prediction [11]. Using SRBCTs data set, we selected top ranked genes according to Dudoit's method and used MLPs to evaluate the prediction performance of the gene sets with different sizes. The estimated MCERs are shown in Figure 8 for the sets with 3 to 99 genes, where several gene sets show excellent prediction performance with MCER = 0%. Among those gene sets with MCER = 0%, the smallest gene set has 22 genes that are available online at our website (Table S2; http://www.cbil.ece.vt.edu). From this result, it seemed to suggest that Dudoit's method offered a

Table 2: NCI's SRBCTs: the gene list (with 9 JDGs) identified by our two-step gene selection method.

Image ID	Gene symbol	Gene name
784224	*FGFR4*	fibroblast growth factor receptor 4
195751	*AKAP7*	A kinase (PRKA) anchor protein 7
81518	*OCRL*	apelin; peptide ligand for APJ receptor
1434905	*HOXB7*	homeo box B7
770394	*FCGRT*	Fc fragment of IgG, receptor, transporter, alpha
609663	*PRKAR2B*	protein kinase, cAMP-dependent, regulatory, type II, beta
629896	*MAP1B*	microtubule-associated protein 1B
207274	*IGF2*	Human DNA for insulin-like growth factor II (IGF-2)
80338	*SELENBP1*	selenium binding protein 1

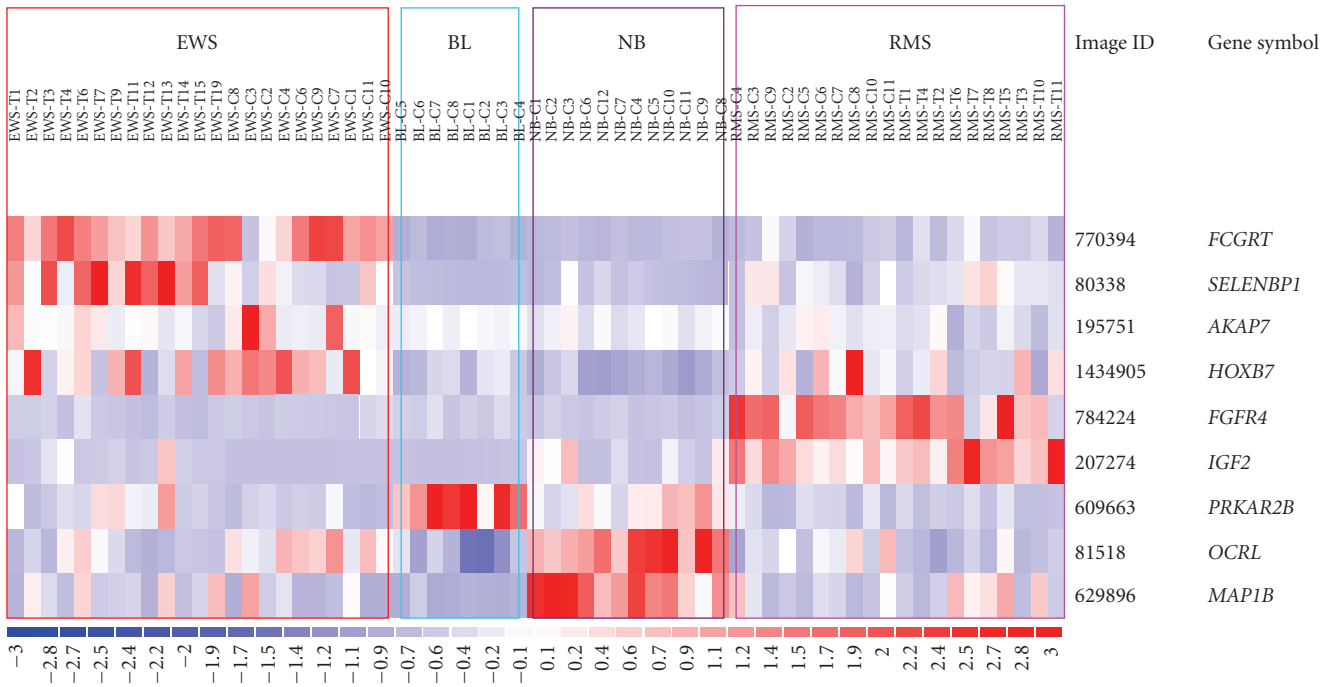

Figure 7: Expression pattern of 63 NCI's SRBCT samples in the gene space of 9 JDGs in Table 2.

slightly better prediction performance for diagnosing SRBCTs, although, again, our IDG/JDG method found a much smaller set of genes with comparable prediction performance.

Finally, we tested the classification capability of the MLP classifiers using the newly identified 9 JDGs on a set of 25 blinded test samples. The blinded test samples were 20 SRBCTs (6 EWS, 5 RMS, 6 NB, and 3 BL) and 5 non-SRBCTs for testing the ability of these models to reject a diagnosis. The non-SRBCTs include 2 normal muscle tissues (Tests 9 and 13), 1 undifferentiated sarcoma (Test 5), 1 osteosarcoma (Test 3) and 1 prostate carcinoma (Test 11). A sample is classified to a diagnostic group if it receives the highest vote for that group among four possible outputs, and all samples will be classified to one of the four classes. We then follow the method described in the supplementary material of [12]—by calculating the empirical probability distribution of distance between samples and their ideal output—to set a statistical cutoff for rejecting a diagnosis that a sample is classified to a given group. If a sample falls outside the 95th percentile of the probability distribution of distance, its diagnosis is rejected. As shown in Table 4, with our 9 JDGs, we can successfully classify the 20 SBRCT test samples into their categories with 100% accuracy. Then we use the 95th percentile criterion to confirm and reject the classification results. For the 5 non-SRBCT test samples, we can correctly exclude them from any of the four diagnostic categories, since they fall outside the 95th percentiles. For two of the SBRCT samples (Test 1 and Test 10), however, even though they are correctly assigned to their categories (NB and RMS, resp.), their distance from a perfect vote is greater than the expected 95th percentile distance. Therefore, we cannot confidently diagnose them by the "95th percentile" criterion. The "95th percentile" criterion also rejected the classification result of

Table 3: Misclassification error rates of MLPs using three different gene sets on SRBCTs.

Selection method	Number of genes	Misclassification error rate
ANN (Khan et al. [12]) (3-fold cross-validation)	96	4.08%
IDG/JDG selection by wFC and SFS (3-fold cross-validation)	9	3.10%
Nearest shrunken centroid (Tibshirani et al. [14]) (10-fold cross-validation)	43	3.19%
IDG/JDG selection by wFC and SFS (10-fold cross-validation)	9	2.24%

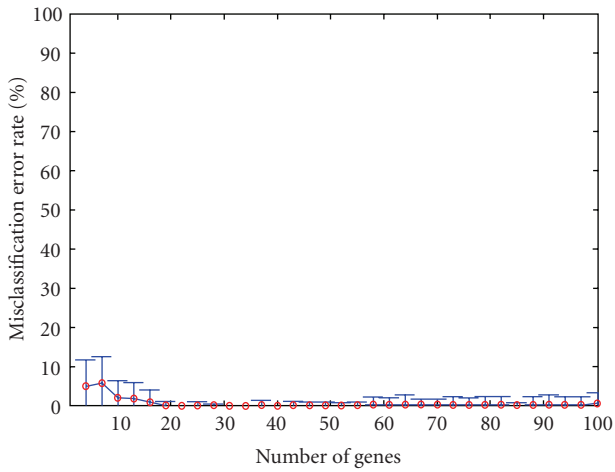

Figure 8: Prediction performance of the genes selected by Dudoit's method (NCI's SRBCTs): misclassification error rates calculated by MLPs with 3-fold cross-validation; error bar indicates the standard deviation.

two blind test SBRCT samples (Test 10 and Test 20) with ANN committee vote using NCI's 96 genes [12]. Tibshirani et al. also reported the result on the blind test data set using their shrunken centroid method. They used the discriminant scores to construct estimates of the class probabilities in a form of Gaussian linear discriminant analysis [14]. With the estimated class probabilities, all 20 known SRBCT samples in the blind test data set were correctly classified in their corresponding classes. For the 5 non-SRBCT samples, their estimated class probabilities were lower than the class probabilities of the 20 SRBCT samples, however, two of them show a relatively high class probability (> 70%) in RMS category, hence, hard to reject their diagnoses as RMS samples.

3.2. CNMC's MDs

We applied our two-step gene selection algorithm to CNMC's MD data set that consisted of 39 gene expression profiles of 4 types of muscular dystrophies (BDM, DMD, Dysferlin, and FSHD; see Table 1 for the details). In the first step, the top 100 IDGs were initially selected by 1D wFC using leave-one-out validation. Additional material (Table S3) lists the top 100 IDGs with gene names and descriptions, which can be found online at our website (http://www.cbil.ece.vt.edu). Again, the number 100 was chosen somehow empirically, however, with a preliminary study of (1) the genes' discriminatory power in terms of low MCERs and (2) a stability analysis of ranking using leave-one-out and bootstrap methods as described in Section 3.1. In the second step, different sets of JDGs were then selected by the SFFS method for the number of genes ranging from 1 to 99. Notice that we used the SFFS method in this experiment instead of the SBFS as in the SRBCT study. It has been reported in [6] that the SFFS and SBFS are of similar performance in finding suboptimal sets for class prediction. Our experimental results shown below supported this conclusion. If the targeted gene sets are likely small, the SFFS method can offer some computational advantage over the SBFS approach. The resulting sets of JDGs were fed into MLPs for performance evaluation. Figure 9 shows the calculated misclassification error rates (MCERs) achieved by the JDG sets using MLPs with 3-fold cross-validation. The minimum value of MCER was 14.8% when using 11 JDGs. We also compared the prediction performance of the JDGs with that of the IDGs. As shown in Figure 10 that the minimum value of MCER was 15.5% when using 69 IDGs. Therefore, JDGs outperformed IDGs not only with a slightly lower MCER, but also with a much smaller subset of genes. The selected 11 JDGs are listed in Table 5, and the expression pattern of these 11 JDGs in 39 gene expression profiles is shown in Figure 11.

As a comparison, we also used Dudoit's method to select top ranked genes for the MD data set, which can be found online at our website (Table S4; http://www.cbil.ece.vt.edu). In this study, we used the OVA SVM approach [26] to evaluate the prediction performance of the genes selected by Dudoit's method as well as that of IDGs and JDGs, respectively. As shown in Figure 12, the MCER of our 11 JDGs is 7.69% (std = 3.36%), which is much lower than that estimated by MLPs (i.e., MCER = 14.8%). The minimum MCER of the genes selected by Dudoit's method reaches 10.26% (std = 3.55%) using 94 genes, while reaching 5.95% (std = 3.64%) using 60 IDG genes. Therefore, for this data set, the IDG selection method (using 1D wFC) outperformed Dudoit's method (using 1D FC). The second step in our method, that is, JDG selection by wFC, can be further used to find smaller gene sets with good prediction performance. In addition to the 11-JDGs listed before, SVMs also found another set of JDGs (n = 37; a larger set compared to the 11 JDG set) with slightly better prediction performance (MCER = 6.51%; std = 3.51%).

Table 4: MLP diagnostic predictions using the 9 JDGs in Table 2 on 25 testing SRBCTs.

Sample lable	MLP committee vote				MLP classification	MLP diagnosis	Histological diagnosis
	EWS	BL	NB	RMS			
Test 1	0.028883	0.17785	**0.79642**	0.028943	NB	∼	NB-C
Test 2	**0.77272**	0.21654	0.013751	0.094902	EWS	EWS	EWS-C
Test 3	0.11429	0.090061	0.32829	**0.34814**	RMS	∼	Osteosarcoma-C
Test 4	0.000568	0.001845	0.000338	**0.99895**	RMS	RMS	ARMS-T
Test 5	0.17165	0.007505	**0.55362**	0.35478	RMS	∼	Sarcoma-C
Test 6	**0.99931**	0.000541	0.000381	0.000363	EWS	EWS	EWS-T
Test 7	0.046811	**0.93487**	0.005181	0.03	BL	BL	BL-C
Test 8	0.000629	0.030889	**0.94595**	0.029323	NB	NB	NB-C
Test 9	**0.44191**	0.2341	0.058133	0.15217	EWS	∼	Sk. Muscle
Test 10	0.25353	0.004541	0.085682	**0.84684**	RMS	∼	ERMS-T
Test 11	0.15522	0.2033	**0.30956**	0.18985	NB	∼	Prostate Ca.-C
Test 12	**0.98211**	0.012711	0.005199	0.017581	EWS	EWS	EWS-T
Test 13	0.11715	0.055632	0.32017	**0.41996**	RMS	∼	Sk. Muscle
Test 14	0.010447	0.020733	**0.97481**	0.00498	NB	NB	NB-T
Test 15	0.007201	**0.96229**	0.032875	0.001926	BL	BL	BL-C
Test 16	0.012105	0.069357	**0.92035**	0.006036	NB	NB	NB-T
Test 17	0.001885	0.038029	0.030515	**0.92519**	RMS	RMS	ARMS-T
Test 18	0.000678	**0.96813**	0.01893	0.02351	BL	BL	BL-C
Test 19	**0.99899**	4.50E-07	0.0009	0.005982	EWS	EWS	EWS-T
Test 20	**0.89893**	0.028121	0.006424	0.19883	EWS	EWS	EWS-T
Test 21	**0.94549**	0.055391	0.009161	0.037187	EWS	EWS	EWS-T
Test 22	0.004974	0.000617	0.006337	**0.99702**	RMS	RMS	ERMS-T
Test 23	0.032343	0.005809	**0.95009**	0.027456	NB	NB	NB-T
Test 24	0.033041	0.005865	0.001679	**0.98991**	RMS	RMS	ERMS-T
Test 25	0.039903	0.071522	**0.90682**	0.007421	NB	NB	NB-T

Table 5: CNMC's MDs: the gene list (with 11 JDGs) identified by our two-step gene selection method. ("+": up regulated, "−": down regulated, and "N": neither up or down regulated).

Probe set	Gene symbol	Gene name	BMD	DMD	LDMD2B	FSHD	Pathophysiology
222280_at	—	CDNA clone IMAGE:6602785, partial cds	+	N	+++	−	Unknown
212488_at	*COL5A1*	collagen, type V, alpha 1	++	+++	+	N	Regeneration
200735_x_at	*NACA*	nascent-polypeptide-associated complex alpha polypeptide	+	+	+	−	Unknown
211734_s_at	*FCER1A*	Fc fragment of IgE, high affinity I, receptor for; alpha polypeptide	++	+++	+	N	Inflammation Degeneration
209156_s_at	*COL6A2*	collagen, type VI, alpha 2	++	+++	+	N	Fibrosis
208695_s_at	*RPL39*	ribosomal protein L39	+	+	+	−−	Regeneration
205730_s_at	*ABLIM3*	actin binding LIM protein family, member 3	++	+	+	−−	Regeneration
202966_at	*CAPN6*	Calpain 6	++	+++	+	−	Regeneration
205422_s_at	*ITGBL1*	integrin, beta-like 1 (with EGF-like repeat domains)	+	++	+	N	Regeneration
202409_at	*EST (IGF2 3′)*	Hypothetical protein off 3' UTR of IGF2	++	+++	++	N	Regeneration
213048_s_at	*I-2PP2A*	Phosphatase 2A inhibitor I2PP2A	+++	+	++	−	Proliferation

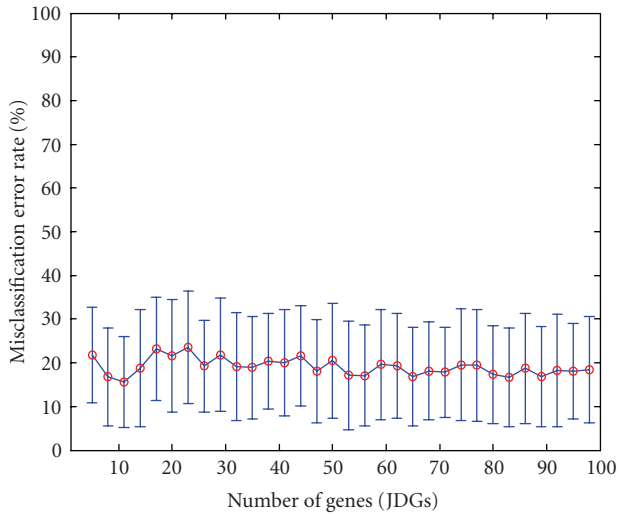

Figure 9: JDGs' prediction performance (CNMC's MDs): misclassification error rates calculated by MLPs with 3-fold cross-validation; error bar indicates the standard deviation.

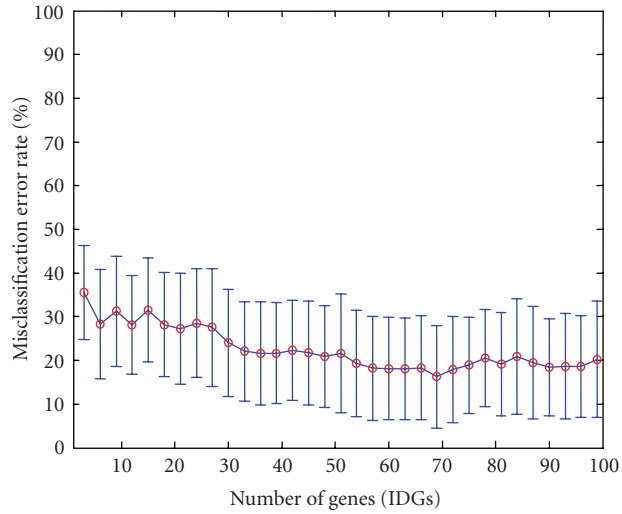

Figure 10: IDGs' prediction performance (CNMC's MDs): misclassification error rates calculated by MLPs with 3-fold cross-validation; error bar indicates the standard deviation.

A subset of the Fascioscapulohumeral muscular dystrophy (FSHD) biopsies has been recently published [32], and different smaller microarrays for Duchenne muscular dystrophy have been published (U95A-E, MuscleChip) [27, 33, 34]. The differential diagnosis of these four disorders is typically not difficult. FSHD is an adult disorder showing dominant inheritance of a subtelomeric deletion of chromosome 4q sequence that is detected using DNA tests [32]. This is a repetitive sequence, and it is not clear what the biochemical consequences of this deletion are. Duchenne muscular dystrophy is a childhood disease showing "frame-shift" mutations of the dystrophin gene, and loss of function of the protein [35, 36]. Becker muscular dystrophy is typically an adult disease, with "in-frame" deletions of the dystrophin gene leading to production of abnormal, yet semifunctional dystrophin protein. Dystrophin is a component of the plasma membrane cytoskeleton in muscle cells (myofibers), where it stabilizes the membrane. The fourth studied group, dysferlin deficiency, is also called Limb-girdle muscular dystrophy 2B (LGMD 2B). This is a recessively inherited adult disease that shows clinical symptoms very similar to Becker muscular dystrophy, but is due to loss of a trans membrane protein, dysferlin, that seems involved in vesicle traffic. Patients with complete dysferlin-deficiency by immunoblot typically have mutations of the corresponding gene.

While it is promising to diagnose these four disorders using molecular signatures, the molecular pathophysiology downstream of the primary defect is very poorly understood. Each of these disorders shows some enigmatic clinical and histological features, yet these have been difficult to tie to the primary gene and protein problem. Thus, this data set and the analysis of this using the two-step IDG/JDG selection methods described here provide potential new insights into the molecular pathophysiology of the muscular dystrophies. In considering the 11 diagnostic genes, we looked at what was known about the function of each of these genes, and then began to build a model for molecular pathophysiology based upon the IDG/JDG analyses.

Eight of the eleven diagnostic genes appear to reflect the degree of severity of the "dystrophic process" in muscle, namely, myofiber degeneration, regeneration, and fibrosis. For example, Calpain 6 is a calcium sensitive protease that is known to be involved in fusion of myoblasts into syncytial myotubes. Consistent with this role, query of a 27 time point in vivo muscle regeneration series shows very low expression in normal nonregenerating muscle, but strong induction of the gene at around day 4 of regeneration, at the time point when myoblast fusion takes place (Figure 13(a)) [30, 31, 37]. Duchenne muscular dystrophy is the most clinically and histologically severe of the four groups studied, and it showed the greatest expression of Calpain 6, while the other dystrophies showed less expression likely consistent with the relative amount of regeneration in the muscle (Figure 13(b) and Table 5). Seven additional probe sets showed similar patterns, including collagen V (fibrosis), FCER1A (mast cells), a ribosomal gene (L39), an actin binding protein (ABLIM3), an integrin (ITGBL1), and a form of IGF2. All show staged induction during muscle regeneration or fibrosis in other models [31].

Two probe sets showed unique patterns that likely drove the performance of the weighted FC in this analysis, but both are poorly characterized proteins, namely, the Phosphatase 2A inhibitor I2PP2A gene (highly expressed in BMD), and a cDNA clone (222280_at) that appears diagnostic of LGMD2B (dysferlin deficiency) (Table 5). The I2PP2A gene may be involved in the more successful regeneration of muscle seen in many BMD patients compared to the other dystrophies. The

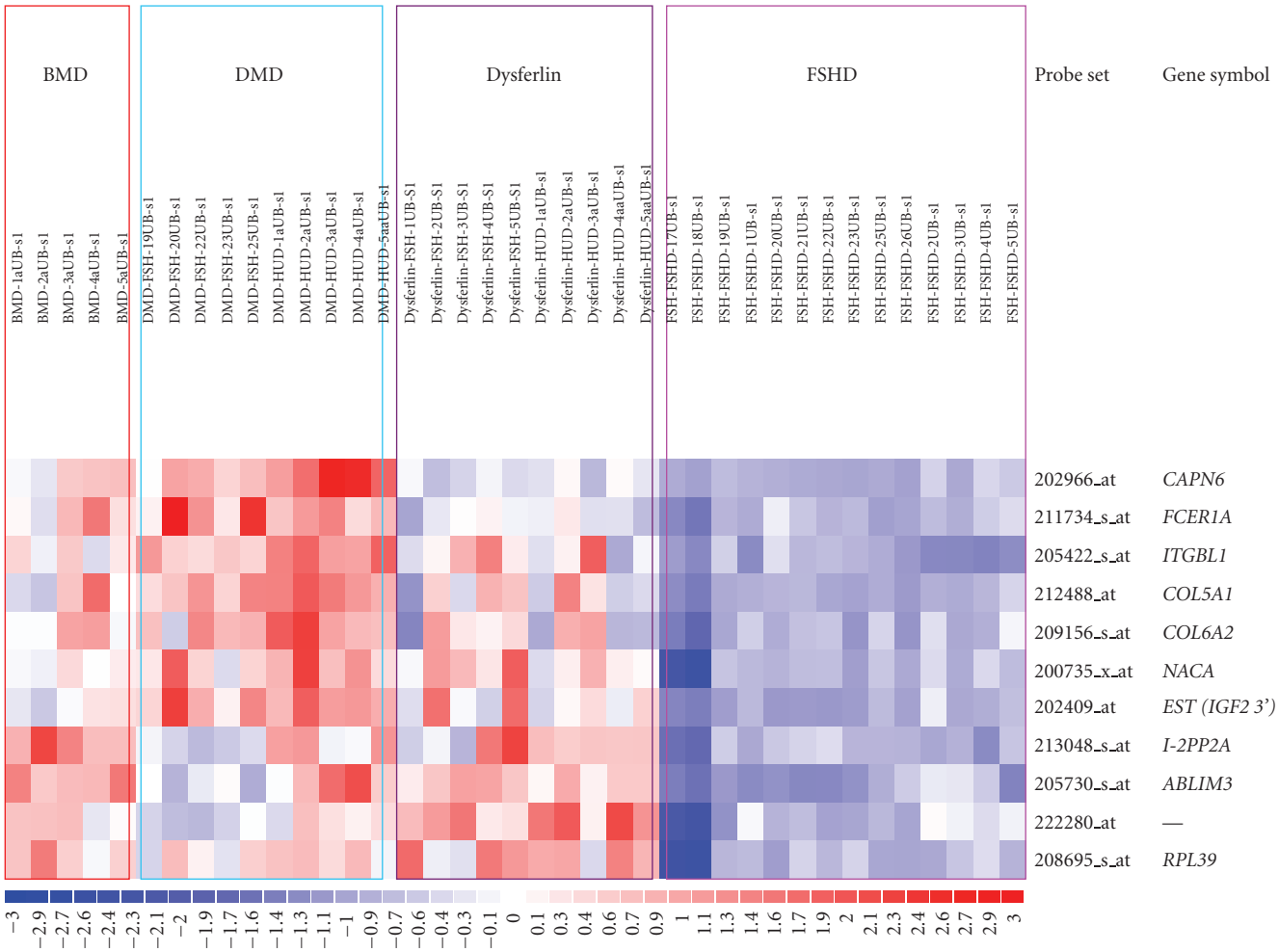

Figure 11: Expression pattern of 39 CNMC's MD samples in the gene space of 11 JDGs in Table 5.

function of the anonymous cDNA clone (222280_at) will be important to determine, as it may be involved in vesicle fusion functions, as is dysferlin.

4. CONCLUSION

We have applied a gene selection approach to multiclass prediction using gene expression profiles. The two-step method starts with individually discriminatory gene (IDG) selection using 1D wFC to reduce initially data dimensionality to a manageable size. The jointly discriminatory genes (JDGs) are then selected by a sequential search method (SFS) to further reduce the dimensionality to a smaller size. The approach has been applied to two microarray studies: (1) small round blue cell tumors (SRBCTs) of childhood, and (2) muscular dystrophies (MDs). The performance of the selected gene lists was evaluated by ANN classifiers (MLPs) and/or SVMs, which demonstrated that high and generalizable prediction performance can be achieved for diagnosing SBCTs and MDs when gene selection is properly done.

Microarray analysis is a widely used technology for studying gene expression on a global scale. However, the technology is presently not used as a routine diagnostic tool. One difficulty in using these high-throughput arrays for clinical practice would be costly, due to the fact that synthesizing the necessary polymerase chain reaction (PCR) primers for such a large number of genes increases production costs drastically [38, 39]. As demonstrated in this paper and the others (e.g., [40]), usually the measurements of a small set of genes are adequate to build a classifier that distinguishes one disease subtype from another. Therefore, for diagnosis it is not necessary to screen gene expression on a whole genome basis, but instead customized microarrays (i.e., diagnostic microarrays) with considerably less genes can be used. As an example, researchers have recently made it possible to convert a breast cancer microarray signature into a high-throughput diagnostic test [41]. We believe that our gene selection approach providing a much smaller set of genes would be an effective tool to help further reduce the costs of diagnostic microarrays for clinical applications.

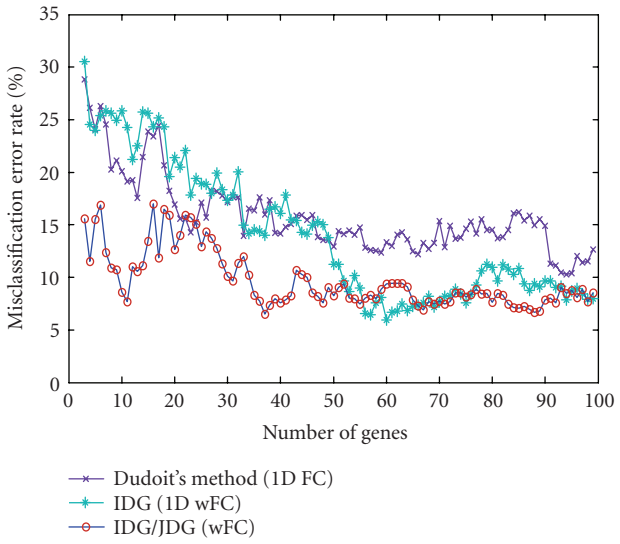

FIGURE 12: Prediction performance comparison: (1) using the genes selected by Dudoit's method, (2) using IDGs, and (3) using JDGs (CNMC's MDs); misclassification error rates calculated by SVMs with 3-fold cross-validation.

From the experimental results shown in Figures 6 and 8, we have observed an interesting behavior of MLPs with JDGs as inputs: with an increase in feature size (i.e., the number of JDGs), the generalization error first decreases, then increases, has a maximum at $p = N/\alpha$ (where N is the number of samples and α is usually found between 2 and 10) and then decreases again. This somewhat "strange" behavior was termed as "peaking phenomenon" [42] or "scissor effect" [43–45]. The behavior was theoretically investigated by Raudys and Duin and found top be the result of "small size effect" on the classifiers using peudo-Fisher linear discriminant (PFLD) [45]. In [43, 44], it was also shown that under certain conditions, ANNs (in particular, nonlinear single-layer perceptron (SLP)) can realize the PFLD. We believe that our microarray study results with MLPs provided a strong experimental support to the theoretical analysis of the "peaking phenomenon." It is worth noting that some possible approaches can be utilized to improve the generalization performance when feature size is much larger than the sample size [42]. One of the approaches is to use support vector machines (SVMs) that offer a systematic procedure for reducing the number of samples in the training set to define the classifiers [46]. Another possible approach is to use subspace methods in which each class is approximated by subspace of the feature space [47]. As pointed out in [42], the subspace methods can provide us a way to control the complexity of classifiers directly, while, in the SVM the number of support vectors, and thereby the classifier complexity, is a result of complexity constraining and cannot be preset in general. In future, we will investigate the subspace classifiers in conjunction with a large number of JDGs for improving further the generalization performance in diagnosing muscular dystrophies.

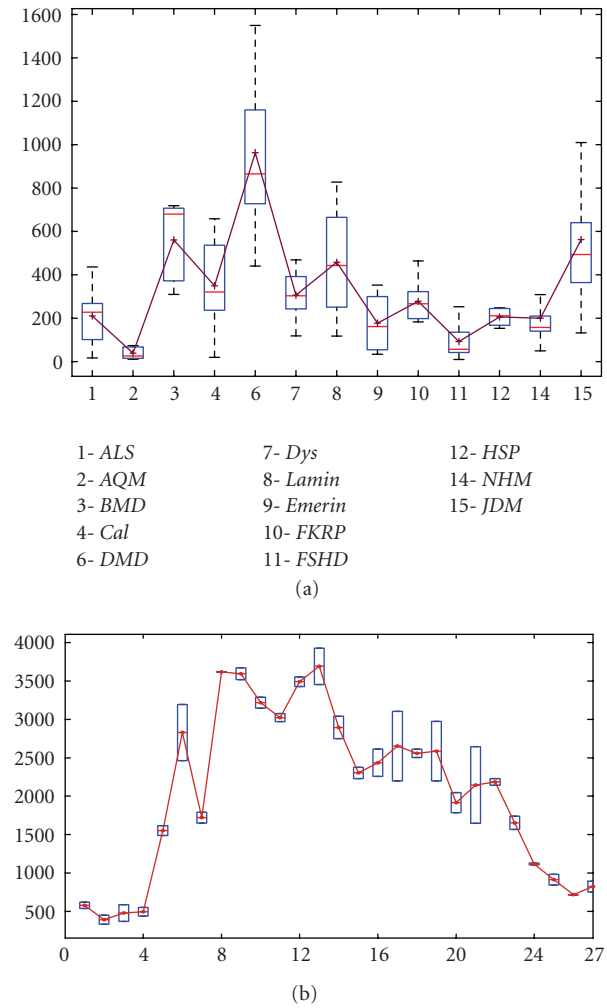

FIGURE 13: Diagnostic genes of the muscular dystrophies are often involved in muscle regeneration. (a) A query of the Calpain 6 diagnostic gene selected by wFC (see Table 5) against a 13-group muscular dystrophy data set (see http://pepr.cnmcresearch.org). Calpain 6 is the highest in DMD, as summarized in Table 5. (b) The query of Calpain 6 in a 27-time point muscle regeneration series in mice (see http://pepr.cnmcresearch.org; Zhao et al. [30], Zhao and Hoffman [31]). Calpain 6 shows very low expression at time 0 (normal muscle) with strong induction at time points at days 4–8, consistent with its role in myoblast fusion to myotubes (regeneration).

ACKNOWLEDGMENTS

This work was supported by NIH Grants (CA109872, CA096483, EB000830, and NS29525-13A); and a DOD/CDMRP Grant (BC030280). We thank R. Lee and Y. Zhu at Georgetown University for their helpful and insightful discussions. The authors also thank the anonymous reviewers for their invaluable inputs that lead to several important improvements.

REFERENCES

[1] M. Bittner, P. Meltzer, Y. Chen, et al., "Molecular classification of cutaneous malignant melanoma by gene expression profiling," *Nature*, vol. 406, no. 6795, pp. 536–540, 2000.

[2] T. R. Golub, D. K. Slonim, P. Tamayo, et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, 1999.

[3] M. A. Shipp, K. N. Ross, P. Tamayo, et al., "Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning," *Nature Medicine*, vol. 8, no. 1, pp. 68–74, 2002.

[4] L. Liotta and E. Petricoin, "Molecular profiling of human cancer," *Nature Reviews Genetics*, vol. 1, no. 1, pp. 48–56, 2000.

[5] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: a review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 4–37, 2000.

[6] A. K. Jain and D. Zongker, "Feature selection: evaluation, application, and small sample performance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 2, pp. 153–158, 1997.

[7] S. J. Raudys and A. K. Jain, "Small sample size effects in statistical pattern recognition: recommendations for practitioners," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 3, pp. 252–264, 1991.

[8] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, Boston, Mass, USA, 2nd edition, 1990.

[9] P. A. Devijver and J. Kittler, *Pattern Recognition: A Statistical Approach*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1982.

[10] P. Pudil, J. Novovicova, and J. Kittler, "Floating search methods in feature selection," *Pattern Recognition Letters*, vol. 15, no. 11, pp. 1119–1125, 1994.

[11] S. Dudoit, J. Fridlyand, and T. P. Speed, "Comparison of discrimination methods for the classification of tumors using gene expression data," *Journal of the American Statistical Association*, vol. 97, no. 457, pp. 77–87, 2002.

[12] J. Khan, J. S. Wei, M. Ringnér, et al., "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," *Nature Medicine*, vol. 7, no. 6, pp. 673–679, 2001.

[13] T. Li, C. Zhang, and M. Ogihara, "A comparative study of feature selection and multiclass classfication methods for tissue classification based on gene expression," *Bioinformatics*, vol. 20, no. 15, pp. 2429–2437, 2004.

[14] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu, "Diagnosis of multiple cancer types by shrunken centroids of gene expression," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 10, pp. 6567–6572, 2002.

[15] M. Xiong, X. Fang, and J. Zhao, "Biomarker identification by feature wrappers," *Genome Research*, vol. 11, no. 11, pp. 1878–1887, 2001.

[16] M. Loog, *Approximate Pairwise Accuracy Criteria for Multiclass Linear Dimension Reduction: Generalisations of the Fisher Criterion*, Delft University Press, Delft, The Netherlands, 1999.

[17] M. Loog, R. P. W. Duin, and R. Haeb-Umbach, "Multiclass linear dimension reduction by weighted pairwise Fisher criteria," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 7, pp. 762–766, 2001.

[18] J. C. Koop, "Generalized inverse of a singular matrix," *Nature*, vol. 200, p. 716, 1963.

[19] W. M. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes: The Art of Scientific Computing*, Cambridge University Press, New York, NY, USA, 1986.

[20] P. M. Narendra and K. Fukunaga, "A branch and bound algorithm for feature subset selection," *IEEE Transactions on Computers*, vol. 26, no. 9, pp. 917–922, 1977.

[21] T. Marill and D. M. Green, "On the effectiveness of receptors in cognition system," *IEEE Transactions on Information Theory*, vol. 9, pp. 11–17, 1963.

[22] A. W. Whitney, "A direct method of nonparametric measurement selection," *IEEE Transactions on Computers*, vol. 20, no. 9, pp. 1100–1103, 1971.

[23] S. D. Stearns, "On selecting features for pattern classifiers," in *Proceedings of the 3rd International Conference on Pattern Recognition*, pp. 71–75, Coronado, Calif, USA, November 1976.

[24] S. Haykin, *Neural Networks: A Comprehensive Foundation*, Prentice-Hall, Upper Saddle River, NJ, USA, 2nd edition, 1999.

[25] Y. Lee and C.-K. Lee, "Classification of multiple cancer types by multicategory support vector machines using gene expression data," *Bioinformatics*, vol. 19, no. 9, pp. 1132–1139, 2003.

[26] S. Ramaswamy, P. Tamayo, R. Rifkin, et al., "Multiclass cancer diagnosis using tumor gene expression signatures," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 26, pp. 15149–15154, 2001.

[27] M. Bakay, Y.-W. Chen, R. Borup, P. Zhao, K. Nagaraju, and E. Hoffman, "Sources of variability and effect of experimental approach on expression profiling data interpretation," *BMC Bioinformatics*, vol. 3, no. 1, p. 4, 2002.

[28] M. Bakay, Z. Wang, G. Melcon, et al., "Nuclear envelope dystrophies show a transcriptional fingerprint suggesting disruption of Rb-MyoD pathways in muscle regeneration," *Brain*, vol. 129, no. 4, pp. 996–1013, 2006.

[29] Affymetrix Technical Note, "Statistical algorithms description document," Affymetrix, 2002, http://www.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf.

[30] P. Zhao, J. Seo, Z. Wang, Y. Wang, B. Shneiderman, and E. Hoffman, "In vivo filtering of in vitro expression data reveals MyoD targets," *Comptes Rendus - Biologies*, vol. 326, no. 10-11, pp. 1049–1065, 2003.

[31] P. Zhao and E. Hoffman, "Embryonic myogenesis pathways in muscle regeneration," *Developmental Dynamics*, vol. 229, no. 2, pp. 380–392, 2004.

[32] S. Winokur, Y.-W. Chen, P. S. Masny, et al., "Expression profiling of FSHD muscle supports a defect in specific stages of myogenic differentiation," *Human Molecular Genetics*, vol. 12, no. 22, pp. 2895–2907, 2003.

[33] M. Bakay, P. Zhao, J. Chen, and E. Hoffman, "A web-accessible complete transcriptome of normal human and DMD muscle," *Neuromuscular Disorders*, vol. 12, supplement 1, pp. S125–S141, 2002.

[34] Y.-W. Chen, P. Zhao, R. Borup, and E. Hoffman, "Expression profiling in the muscular dystrophies: identification of novel aspects of molecular pathophysiology," *Journal of Cell Biology*, vol. 151, no. 6, pp. 1321–1336, 2000.

[35] E. Hoffman, R. H. Brown Jr., and L. M. Kunkel, "Dystrophin: the protein product of the Duchenne muscular dystrophy locus," *Cell*, vol. 51, no. 6, pp. 919–928, 1987.

[36] M. Koening, E. Hoffman, C. J. Bertelson, A. P. Monaco, C. Feener, and L. M. Kunkel, "Complete cloning of the Duchenne muscular dystrophy (DMD) cDNA and preliminary genomic organization of the DMD gene in normal and affected individuals," *Cell*, vol. 50, no. 3, pp. 509–517, 1987.

[37] P. Zhao, S. Iezzi, E. Carver, et al., "Slug is a novel downstream target of MyoD. Temporal profiling in muscle regeneration," *Journal of Biological Chemistry*, vol. 277, no. 33, pp. 30091–30101, 2002.

[38] R. J. Fernandes and S. S. Skiena, "Microarray synthesis through multiple-use PCR primer design," *Bioinformatics*, vol. 18, supplement 1, pp. S128–S135, 2002.

[39] J. Jaeger, D. Weichenhan, B. Ivandic, and R. Spang, "Early diagnostic marker panel determination for microarray based clinical studies," *Statistical Applications in Genetics and Molecular Biology*, vol. 4, no. 1, article 9, 2005.

[40] W. Li, "How many genes are needed for early detection of breast cancer, based on gene expression patterns in peripheral blood cells?" *Breast Cancer Research*, vol. 7, no. 5, p. E5, 2005.

[41] A. M. Glas, A. Floore, L. J. M. J. Delahaye, et al., "Converting a breast cancer microarray signature into a high-throughput diagnostic test," *BMC Genomics*, vol. 7, p. 278, 2006.

[42] R. P. W. Duin, "Classifiers in almost empty spaces," in *Proceedings of the 15th International Conference on Pattern Recognition (ICPR '00)*, vol. 2, pp. 1–7, Barcelona, Spain, September 2000.

[43] S. J. Raudys, "Evolution and generalization of a single neurone—I: single-layer perceptron as seven statistical classifiers," *Neural Networks*, vol. 11, no. 2, pp. 283–296, 1998.

[44] S. J. Raudys, "Evolution and generalization of a single neurone—II: complexity of statistical classifiers and sample size considerations," *Neural Networks*, vol. 11, no. 2, pp. 297–313, 1998.

[45] S. J. Raudys and R. P. W. Duin, "Expected classification error of the Fisher linear classifier with pseudo-inverse covariance matrix," *Pattern Recognition Letters*, vol. 19, no. 5-6, pp. 385–392, 1998.

[46] V. N. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, New York, NY, USA, 1998.

[47] E. Oja, *Subspace Methods of Pattern Recognition*, John Wiley & Sons, New York, NY, USA, 1984.