BIOINFORMATICS

Vol. 19 no. 1 2003
Pages 37–44

Genetic algorithms applied to multi-class prediction for the analysis of gene expression data

C.H. Ooi[1] and Patrick Tan[2,*]

[1]Nanyang Technological University, School of Mechanical and Production Engineering, 50 Nanyang Avenue, Singapore 639798 and [2]Division of Cellular and Molecular Research, National Cancer Center/Defence Medical Research Institute, 11 Hospital Drive, Singapore 169610, Republic of Singapore

Received on April 17, 2002; revised on July 3, 2002; accepted on July 15, 2002

ABSTRACT

Motivation: An important challenge in the use of large-scale gene expression data for biological classification occurs when the expression dataset being analyzed involves multiple classes. Key issues that need to be addressed under such circumstances are the efficient selection of good predictive gene groups from datasets that are inherently 'noisy', and the development of new methodologies that can enhance the successful classification of these complex datasets.

Methods: We have applied genetic algorithms (GAs) to the problem of multi-class prediction. A GA-based gene selection scheme is described that automatically determines the members of a predictive gene group, as well as the optimal group size, that maximizes classification success using a maximum likelihood (MLHD) classification method.

Results: The GA/MLHD-based approach achieves higher classification accuracies than other published predictive methods on the same multi-class test dataset. It also permits substantial feature reduction in classifier genesets without compromising predictive accuracy. We propose that GA-based algorithms may represent a powerful new tool in the analysis and exploration of complex multi-class gene expression data.

Availability: Supplementary information, data sets and source codes are available at http://www.omniarray.com/bioinformatics/GA.

Contact: cmrtan@nccs.com.sg

INTRODUCTION

The increasing use of DNA microarrays or 'genechips' to generate large-scale gene expression datasets has led to several important statistical and analytical challenges. One area in which this technology is showing exciting promise is in the field of molecular diagnosis, where the phenotypic classification of a biological sample is largely based on gene expression data. Examples of such phenotypic classifications include tumor subtypes (lung, colon, etc.) and the prediction of chemotherapy response (Staunton *et al.*, 2001). A significant advantage of this new approach is that classification schemes based upon molecular data can often detect biological subtypes that have traditionally eluded more conventional approaches (Alizadeh *et al.*, 2000; Bittner *et al.*, 2000).

Several mathematical approaches have been developed to identify and select key predictive genes in an expression dataset for use in classification. Currently, most of these reports have focused on situations where the expression dataset being analyzed contains only two (binary) to three major classes (e.g. cancer versus normal tissue, response to treatment versus no response) (Alon *et al.*, 1999; Staunton *et al.*, 2001; Li *et al.*, 2001; Zhang *et al.*, 2001). A much more challenging situation occurs, however, when the expression dataset in question contains multiple classes. Under such scenarios, (especially when the number of classes exceeds five), the methodologies that suffice for binary or 3-class (Keller *et al.*, 2000) datasets may not necessarily produce comparable accuracies for larger, more complex, datasets. For example, Dudoit *et al.* (2000) compared various methods of classification on three microarray datasets: a 3-class lymphoma dataset (Alizadeh *et al.*, 2000), a binary leukemia dataset that also doubles as a 3-class dataset (Golub *et al.*, 1999) and a dataset of cell lines corresponding to nine tumor types ('NCI60') (Ross *et al.*, 2000). In that report, the best classifiers, while returning test error rates near 0% for the binary and 3-class datasets, nevertheless returned a *minimum* test error rate of 19% for the NCI60 dataset. Examples such as these indicate that there is a strong need for the development of better algorithms that can effectively analyze multiple-class expression data.

*To whom correspondence should be addressed.

© Oxford University Press 2003

One potential reason for the reduced performance accuracy observed in the multi-class scenarios may be that many currently used approaches rely upon rank-based gene selection schemes (suggested by Dudoit *et al.*, 2000). While they are good at identifying genes that are strongly correlated to the target phenotype class distinction, rank-based methods tend to ignore correlations between genes. In addition, previous methodologies also suffer from the constraint that the size of the predictor set has to be specified *a priori*. Thus, it is unclear if the number of genes used in the final predictor set is actually the optimal number to generate an accurate class prediction.

Genetic algorithms (GAs), as introduced by Goldberg (1989), are randomized search and optimization techniques that derive their working principles by analogy to evolution and natural genetics. Because they are aided by large amounts of implicit parallelism (Grefenstette and Baker, 1989), GAs are capable of searching for optimal or near-optimal solutions on complex and large spaces of possible solutions. Furthermore, GAs allow searching of these spaces of solutions by considering multiple interacting attributes simultaneously, rather than by considering one attribute at a time. Because of these advantages, GAs may represent another useful tool in the classification of biological phenotypes based on gene expression data. Previous reports have described the use of GAs for binary class prediction problems (Li *et al.*, 2001). However, despite their suitability for addressing problems involving large solution spaces, the potential for using GAs in multiple-class prediction settings has to date remained unexplored.

In this report, we use the parallelised searching capability of GAs to design a gene-selection scheme that determines the optimal set of R genes in a multi-class dataset which classifies the samples within the dataset with minimal error. Using this approach, it is possible not only to determine the specific genes that should belong to a predictor set, but also the optimal size of the predictor set, from within a pre-specified range. This approach is radically different from another GA-based method utilized by Li *et al.* (2001). Using a common test multi-class dataset, we find that the GA-based approach delivers higher levels of predictive accuracy as compared to other previously reported methods. In addition, using another multi-class dataset, we show that the GA-based approach is also capable of delivering predictive accuracies that are comparable to other methods using classifier genesets of substantially fewer features than previously required. We propose that GA-based algorithms may represent a powerful new alternative in the analysis and exploration of complex multi-class datasets.

SYSTEM AND METHODS

Dataset and Data Preprocessing

The NCI60 gene expression dataset contains the gene expression profiles of 64 cancer cell lines as measured by cDNA microarrays containing 9703 spotted cDNA sequences (Ross *et al.*, 2000), and was downloaded from http://genome-www.stanford.edu/sutech/download/nci60/dross_arrays_nci60.tgz. Following other reports (Dudoit *et al.*, 2000), the single unknown cell line and two prostate cell lines, due to their small number, were excluded from analysis, leaving a total of 61 cell lines with nine sites of origin: breast (7), central nervous system (6), colon (7), leukemia (6), melanoma (8), non-small-cell-lung-carcinoma or NSCLC (9), ovarian (6), renal (8) and reproductive (4).

During data preprocessing, spots with missing data, control, and empty spots were excluded, leaving 6167 genes. For each array, the expression data of each spot was normalized by subtracting the mean of the Cy5/Cy3 ratio of the control spots, $\mu_{control}$ from the Cy5/Cy3 ratio of each spot, and dividing the result by the standard deviation of the Cy5/Cy3 ratio of the control spots, $\sigma_{control}$. In our analysis, we used a truncated dataset containing 1000 genes with the highest standard deviation value (ranging from 0.8112 to 2.421), which are numbered from 1 to 1000. These genes are henceafter referred to by their index numbers (1 to 1000).

The second dataset ('GCM') (Ramaswamy *et al.*, 2001)) was downloaded from http://www-genome.wi.mit.edu/mpr/publications/projects/Global_Cancer_Map/, and contains the expression profiles of 198 primary tumor samples, 20 poorly differentiated adenocarcinomas and 90 normal tissues samples as measured by Affymetrix Genechips containing 16063 genes and ESTs. Only the 198 primary tumor samples are considered in this work, to make the results comparable to that reported by Ramaswamy *et al.* (2001). The 198 samples originate from 14 sites of origin: prostate (14), bladder (11), melanoma (10), uterine (10), leukemia (30), breast (12), colorectal (12), renal (11), ovarian (12), pancreatic (11), lung (12), lymphoma (22), central nervous system (20) and pleural mesothelioma (11). The data was preprocessed as above to generate a truncated 1000-gene dataset, (standard deviation ranging from 0.299 to 3.089).

Overall Methodology

The GA/MLHD classification strategy consists of two main components: (1) a GA-based gene selector and (2) a maximum likelihood (MLHD) classifier. Component (1) finds a set of R genes that is used to classify the samples, where R lies in the pre-specified range $[R_{\min}, R_{\max}]$. The actual classification process is performed using component (2). Each individual in the population thus

represents a specific gene predictor subset, and a fitness function is used to determine the classification accuracy of a predictor set.

ALGORITHM

Genetic Algorithms (GAs)

In GAs, each potential solution to a problem is represented in the form of a string, which contains encoded parameters of the solution (Holland, 1992). A string is dubbed a chromosome or an individual, while a parameter is also called an attribute. A pool of strings forms a population. Initially, a population is initialized by creating a series of random strings such that each string represents a point in solution space. A fitness function is then defined to measure the degree of goodness of a string. In essence, the fitness value associated with a string indicates the optimality of the solution that the string represents.

The selection process is based on the principle of 'survival of the fittest'. A few strings are chosen from the total population, and each chosen string is then assigned a number of copies to go into a mating pool based on its fitness value. Next, by applying genetic operators (crossover and mutation) on the strings in the mating pool, a new population of strings is formed for the next generation. The process of selection, crossover and mutation are repeated in each subsequent generation until a termination condition is satisfied. Each generation is evaluated by two parameters: the average and the maximum fitness values of the population at that generation. Commonly used termination conditions include defining the maximum number of generations, or the algorithm may be compelled to terminate when there is no significant improvement to the average or maximum fitness value of the population.

STRING REPRESENTATION

The length of a chromosome (string) is $R_{\max} + 1$. For a predictive geneset consisting of $R_{\min}$ to $R_{\max}$ genes (where $R_{\min}$ and $R_{\max}$ are prespecified), the string representation would be

$$[R \quad g_1 \quad g_2 \ldots g_{R_{\max}}].$$

The first element in the string, R, denotes the size of the predictive set represented by the string, and the subsequent elements $g_1, g_2, \ldots, g_{R_{\max}}$ the indices of a subset of genes picked from the truncated 1000 gene dataset. Thus, the string connotes a set of R predictive genes indexed $g_1, g_2, \ldots, g_R$. Only the first R genes out of the $R_{\max}$ genes included in the string are used for classification.

Initialization and Evaluation

An initial population is formed by creating N random strings, where the population size N is prespecified. A random string is produced by randomly generating one integer (represented by R) in the range $[R_{\min}, R_{\max}]$ plus $R_{\max}$ integers (the gene indices) in the range $[1, 1000]$ as its attributes. Each string in the population is then evaluated using the fitness function

$$f(S_i) = 200 - (E_C + E_I) \tag{1}$$

where E_C = cross validation error rate, and E_I = independent test error rate, which are obtained by using the genes contained in string S_i as variables to classify samples with an MLHD classifier (see below).

Selection, Crossover and Mutation

Two selection methods were used to select the strings for the mating pool: (i) stochastic universal sampling (SUS) and (ii) roulette wheel selection (RWS) (described in detail by Goldberg and Deb (1991)).

Crossovers operations were performed by randomly choosing a pair of strings from the mating pool and then applying a crossover operation on the selected string pair with probability p_c. Two offspring strings are produced through the exchange of genetic information between the two parents. This probabilistic process is repeated until all parent strings in the mating pool have been considered. In the GA-based gene selector, we assess two types of crossover operations: one-point and uniform crossover (described in Haupt and Haupt (1998)). In addition, uniform mutation operations are also applied at probability p_m on each of the offspring strings produced from crossover (Spears and De Jong, 1991).

Termination

The processes of evaluation, selection, crossover and mating are repeated for G generations. After the entire run is complete, the string with the best fitness of all generations is outputted as the solution. The string with best overall fitness in a particular run may not necessarily always correspond to the 'best' string in the final or last generation, and so it is common to compare all the 'best' strings from each individual generation to one another, to determine the ultimate string with the highest overall fitness.

MLHD Classifier

To build an MLHD classifier (James, 1985), a total of M_t tumor samples are used as training samples. The remaining M_θ tumor samples are used as test samples. For the NCI60 dataset, the ratio between M_t and M_θ is 2 : 1, while for the GCM dataset, $M_t = 144$ and $M_\theta = 54$ (Ramaswamy *et al.*, 2001). The training dataset is an $R \times M_t$ matrix $X = (x_{ij})$ where element x_{ij} corresponds to the expression level of gene i in training sample j. A sample that is to be classified by the classifier is represented as a

column vector $\vec{e} = \langle e_1, e_2, \ldots, e_R \rangle^T$ where element e_i is the expression level of gene i in that sample.

Discriminant Function

The basis of the discriminant function is Bayes' rule of maximum likelihood: Assign the sample to the class with the *highest* conditional probability. In a domain of Q possible classes, assign the unknown sample to class q if

$$P\left(G_q|\vec{e}\right) > P\left(G_r|\vec{e}\right) \tag{2}$$

for all $q \neq r$ and $q, r \in \{1, 2, \ldots, Q\}$. The conditional probability $P(G_q|\vec{e})$ is the probability that the unknown sample belongs to class q given that its gene expression data vector equals $\vec{e}$.

The computation of the discriminant function for class q is based upon two parameters: the class mean vector and the common covariance matrix. For m_{tq} training samples that belong to class q, the mean of expression of gene i is

$$\mu_{q,i} = \frac{1}{m_{tq}} \sum_{k=1}^{M_t} I\left(c_k = q\right) \cdot x_{ik} \tag{3}$$

where $I(\bullet)$ is the indicator function which equals 1 if the argument inside the parentheses is true, and 0 otherwise. The class to which sample k belongs to is denoted as c_k. Putting the $\mu_{q,i}$ for all i together forms the class mean vector $\vec{\mu}_q = \langle \mu_{q,1}, \mu_{q,2} \ldots \mu_{q,R} \rangle^T$.

The class covariance matrix Σ_q is the covariance matrix of the truncated expression data for all training samples belonging to class q

$$\Sigma_q = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1R} \\ \sigma_{21} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ \sigma_{R1} & \cdots & \cdots & \sigma_{RR} \end{bmatrix} \tag{4}$$

where σ_{ij} = covariance between the gene i and the gene j in class q. The 'pooled estimate' of all class covariance matrices is the common covariance matrix (James, 1985):

$$\Sigma = \frac{1}{M_t - Q} \sum_{q=1}^{Q} \Sigma_q. \tag{5}$$

By using the common covariance matrix, the normality and the equal class probability assumptions, $P\left(G_q\right) = \frac{1}{Q}$ for all $q \in \{1, 2, \ldots, Q\}$, the linear discriminant function in the classifier becomes

$$f_q(\vec{e}) = \vec{\mu}_q^T \Sigma^{-1} \vec{e} - \frac{1}{2} \vec{\mu}_q^T \Sigma^{-1} \vec{\mu}_q. \tag{6}$$

Hence, the maximum likelihood (MLHD) classification rule is: Assign the unknown sample to Class q if

$$f_q(\vec{e}) > f_r(\vec{e}) \tag{7}$$

for all $q \neq r$ and $q, r \in \{1, 2, \ldots, Q\}$.

Evaluating a Predictor Set

Both leave-one-out cross validation and independent test are used to evaluate classifier performance. In the former, one sample is excluded from the training set, and a new MLHD classifier is rebuilt using the remaining $M_t - 1$ training samples. Thus, the new classifier is totally blind to that excluded sample. The classifier is used to classify the left-out sample, and the procedure is then repeated for all of the M_t training samples. In independent test, the MLHD classifier is built using all M_t training samples, and is used to classify M_θ independent test samples.

The cross validation error rate is $E_C = \frac{\chi_C}{M_t} \times 100$ while the independent test error rate is $E_I = \frac{\chi_I}{M_\theta} \times 100$, where χ_C and χ_I are the number of misclassified samples in the cross validation and independent tests respectively.

This evaluation method differs from that used by Dudoit *et al.* (2000), who used 150 different learning/test sets. We use only a set; hence evaluation through cross validation and independent test is needed in order to have an unbiased estimate of classifier performance.

Obtaining Predictor Genes Through Rank-Based Methods

To compare the predictor genesets discovered by the GA/MLHD classifier to genesets obtained by other rank-based strategies, we applied two rank-based methods to the truncated 1000 gene dataset. The first is the gene selection method employed by Dudoit *et al.* (2000), in which genes are ranked on the basis of the ratio of between-groups to within-groups sum of squares (BSS/WSS). The second method adapts the binary-class signal-to-noise (S2N) ratio introduced by Golub *et al.* (1999) for multi-class scenarios. Here, for Q classes, a one-versus-all (OVA) approach is used to form $2Q$ sets of top-ranked genes. For each class, one set of positively correlated genes (largest positive S2N ratio) and another set of negatively correlated genes (smallest negative S2N ratio) are formed. One top positively correlated gene and one top negatively correlated gene were selected for each class, bringing the number of chosen genes to 18, since $Q = 9$ for our dataset (see Supplementary Information for details).

RESULTS

Obtaining the Best Predictor Set for a Particular Set Size Range

We tested our GA-based gene selection methodology on a gene expression dataset containing nine classes ('NCI60'), which previous methodologies have had difficulty classifying (see the **Introduction**). Multiple runs were conducted, in which the population size, N and the maximum number of generations, G, were both set at 100. To observe how different gene selection conditions might affect

the performance characteristics of the GA-based methodology, we varied the following factors in different runs: (i) crossover method, (ii) p_c (0.7 to 1.0), (iii) p_m (0.0005 to 0.01), (iv) selection method, (v) predictor set size range $[R_{\min}, R_{\max}]$ ([5, 10], [11, 15], [16, 20], [21, 25], [26, 30]) and (vi) truncated gene dataset (1000 genes) versus an untruncated gene dataset (6167 genes).

For each predictor set size range (condition v), there are thus 96 different runs (corresponding to combination of conditions (i)-(iv)). Upon completion of a run, the best strings from each individual generation in the run are then collectively compared to find the string with the best overall fitness. For the NCI60 dataset, there appeared to be a consistent trade-off between cross validation and test error rate, so a sorting technique was introduced to obtain the most optimal predictor set that represents the best compromise. The G optimal individuals were sorted ascending by the sum of χ_C and χ_I, then by χ_I, and finally by χ_C.

Effect of GA Parameters and Reproducibility of Gene Selection

The results from the various runs for the range [11, 15] are presented in Table 1 (See Supplementary Information for results of other ranges). We found that in general SUS is superior over the RWS method, regardless of crossover strategy. For crossover methods, we found that uniform crossover produced the best predictor sets in the mid-size ranges [11, 15] and [16, 20], while one-point crossover surpassed the other in generating the best predictors sets in the extreme ranges [5, 10], [21, 25] and [26, 30] (Table 1 and Supplementary Information). The surface plots in Supplementary Information make it possible to speculate that a relatively high crossover probability ($p_c \geqslant 0.8$) and a mutation rate of 2×10^{-3} or above would be likely to produce a good predictor set. Finally, higher predictive accuracies were achieved using a truncated dataset rather than an untruncated one under otherwise identical GA parameters (Table 1), which is not surprising as the untruncated dataset reflects a much larger numerical solution space than the truncated one (about 10^{29} versus 10^{39}, or 10^{10} times larger for a 13-element predictor set). This suggests that data preprocessing using a simple standard deviation filter may effectively reduce 'noise' in the expression dataset. In conclusion, based on fitness and both cross validation and independent test error rates, we found that for the NCI60 dataset the ranges [11, 15] and [16, 20] give the highest classification accuracies.

To assess the reproducibility of the GA/MLHD algorithm, we determined the frequency at which specific genes were selected to belong to an optimal predictor set across 100 independent runs with different initial starting populations, as compared to a series of 100 randomly selected genesets (Figure 1). This analysis revealed that a number of genes were consistently preferentially chosen by the GA/MLHD algorithm, suggesting that the gene selection operation executed by the algorithm is highly reproducible despite its initial 'seeding' of randomly generated individuals.

Table 1. Best predictor sets for the range $R_{\min} = 11$, $R_{\max} = 15$

Crossover	Selection	p_c	p_m	Fitness	E_C	E_I	R
Truncated data set (1000 genes)							
Uniform	SUS	1.0	0.002	180.37	14.63	5	13
One-point	SUS	0.7	0.005	172.93	17.07	10	12
Uniform	RWS	0.7	0.001	167.93	17.07	15	15
One-point	RWS	0.8	0.02	165.61	24.39	10	13
Full data set (6167 genes)							
Uniform	SUS	1.0	0.002	165.61	24.39	10	14

Comparing GA-based Predictor Sets to Predictor Sets Obtained from Other Methodologies

The best predictor set obtained using the GA-based selection scheme exhibited a cross validation error rate of 14.63% and an independent test error rate of 5% (Table 1, row 1, and see Supplementary Information for specific misclassifications). This is an improvement in accuracy as compared to other methodologies assessed by Dudoit *et al.* (2000), where the *lowest* independent test error rate was reported as 19%.

One distinguishing point of the GA-based approach is that it avoids reliance upon a rank-based gene selection scheme. To assess the significance of this feature on the actual process of gene selection, we compared the genes selected by the GA-based approach to genes selected by other commonly used predictive approaches (Table 2, see Supplementary Information for the specific identities of genes selected by the various methodologies). This analysis revealed that the rank-based methods do not identify the majority of the genes selected by the GA-based gene selector, especially when the predictor set size is similar to number of genes in the best GA predictor set (i.e. 13). The BSS/WSS rank-based method comes closest by managing to select six of the 13 GA-based predictor genes, but it is only able to achieve this when the top 100 genes are selected, and not the top 20.

The differences in the predictor sets chosen by the various gene selection schemes can be understood when one compares the expression levels of the genes in the predictor sets. In the rank-based predictor sets, it is quite apparent that several genes share similar expression patterns across the classes (Figures 2b–c, especially Figure 2b). In other words, the expression patterns of these genes are highly correlated to one another. However,

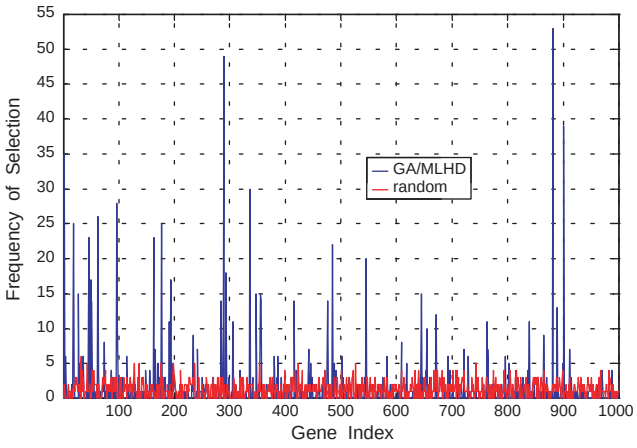

Fig. 1. Reproducibility: frequency of gene selection into 100 optimal predictor sets (fitness ranging from 170.24 to 180.37) from 100 different runs of the GA/MLHD method (blue columns) versus frequency of gene selection into 100 predictor sets through random selection (red columns). GA parameters: $R_{\min} = 11$, $R_{\max} = 15$, $p_c = 1.0$, $p_m = 0.002$, uniform crossover and SUS.

Table 2. Comparison of genes in the best predictor set to the genes found other methods, (+: found, -: not found). All gene selections were obtained from the same 1000-gene truncated dataset

GM	BW20	BW100	OV20	OV2
11	-	+	-	-
366	-	+	-	-
839	-	-	-	-
50	-	-	-	-
828	-	-	-	-
881	-	+	+	-
194	-	+	-	-
289	-	-	-	-
348	-	-	-	-
97	+	+	+	+
127	-	+	-	-
242	-	-	-	-
863	-	-	-	-
	1	6	2	1

GM: Index of predictor genes found by GA/MLHD, BW20: Found in the top 20 genes (BSS/WSS), BW100: Found in the top 100 genes (BSS/WSS), OV20: Found in the top 180 genes (OVA, top 20 from each class), OV2: Found in the top 18 genes (OVA, top 2 from each class). Last row shows the number of genes common to each method and GA/MLHD.

in the GA-based predictor sets, the presence of correlated genes is much less apparent (Figure 2a). The ability to select uncorrelated genes may be one important factor in the improved accuracy of the GA-based approach (see Discussion).

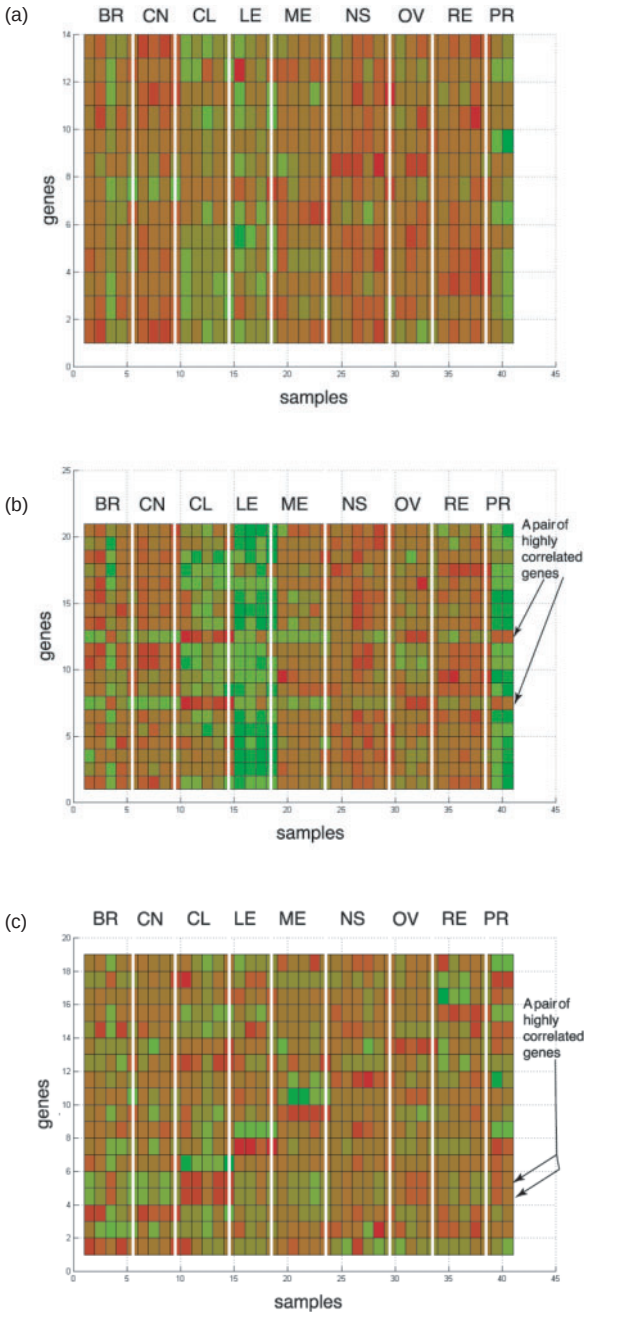

Fig. 2. Comparison of expression profiles of predictor sets obtained through different methodologies. Columns represent different class distinctions, and only training set samples are depicted. (a) Expression profile of genes selected through the GA/MLHD method (only genes for the best predictor set are shown). (b) Expression profile of 20 genes selected through the BSS/WSS ratio ranking method. (c) Expression profile of 18 genes selected through the OVA/S2N ratio ranking method. Arrows depict genes which have highly correlated expression patterns across the sample classes. Classes are labeled as follows: BR (breast), CN (central nervous system), CL (colon), LE (leukemia), ME (melanoma), NS (non-small-cell lung carcinoma), OV (ovarian), RE (renal) and PR (reproductive system).

GA/MLHD Method Permits Significant Feature Reduction While Preserving Accuracy

Recently, Ramaswamy *et al.* (2001) reported a multi-class cancer study utilizing 218 tumor samples spanning 14 classes of different sites of tumor origin, in which each sample was characterized by the expression levels of 16 063 genes ('GCM'). Using a variety of rank-based classifiers such as weighted voting and *k*-nearest neighbors (KNN), the authors reported that the best classification approach, yielding an overall 22% error rate (cross validation and independent test), consisted of using 14 separate OVA/SVM(support-vector-machine)-based binary classifiers. Significantly, however, the authors found that these optimal predictive accuracies were only achieved when the SVM classifiers were trained on *all* 16 063 genes, and that reducing the number of features in the classifier genesets compromised predictive accuracy. Since the requirement for such massive numbers of genes for accurate classification is beyond the capabilities of traditional molecular diagnostics, we wondered if the GA/MLHD classifier, by virtue of its ability to select uncorrelated features as predictor elements, might be capable of selecting smaller optimal predictive genesets without sacrificing predictive accuracy.

We applied the GA/MLHD classifier on the GCM dataset, using our experience with the NCI60 dataset as a guide to select appropriate algorithm parameters (high p_c (0.8), low p_m (0.001), population size = 30, number of generations = 120, uniform crossover, SUS, predictor set size range [1, 60]). Strikingly, we found that the GA/MLHD classifier selected an optimal classifier geneset of 32 elements producing $E_C = 20.67\%$ and $E_I = 14\%$ (overall error rate = 18%). Detailed results, together with the genes selected by the GA/MLHD classifier are depicted in Supplementary Information. Thus, in addition to being slightly more accurate than the OVA/SVM classifier (18% error for GA/MLHD versus 22% for OVA/SVM), the GA/MLHD classifier is able to achieve these levels of predictive accuracy without relying on massive numbers of genes (16 063 genes for OVA/SVM versus 32 genes for GA/MLHD, or an approximately 500-fold reduction). This data suggests that the application of the GA/MLHD classifier may be especially useful in the arena of molecular diagnostics, in which a major aim is the selection of minimal identifier genesets which nevertheless offer high predictive accuracies.

DISCUSSION

In this report, we have described a GA/MLHD-based methodology for multi-class prediction using gene expression data. We defined various parameter ranges which are likely to yield good performance accuracy, and found that optimal predictor sets produced by the GA/MLHD-based approach are more accurate than other predictor sets produced by a number of previously described methodologies. Two chief advantages of the GA/MLHD-based approach include the automatic determination of the optimal number of genes in the predictor set, as well as the ability to select uncorrelated elements as predictor set members.

The rank-based gene selection method of picking the top-scoring *R* genes to form a subset of predictive *R* genes, coupled with a weighted voting system, was first introduced by Golub *et al.* (1999) with the S2N ratio used as the score and appears to work quite well for the prediction of binary class datasets. However, expanding the weighted voting method into multi-class scenarios, as implemented by Yeang *et al.* (2001), did not produce the same satisfactory results. In another multi-class scenario, Dudoit *et al.* (2000) used the BSS/WSS ratio to select predictive genes. It is worthwhile to compare these methods to the criteria established by Hall and Smith (1998), which states that good feature subsets should contain features highly correlated with (predictive of) the class, yet *uncorrelated* with (not predictive of) each other. In other words, good predictor sets should ideally contain genes that are strongly correlated to class distinction but each of these genes should be as uncorrelated with each other as possible. In general, predictor genes obtained through rank-based selection methods typically fulfill the first criteria but not the second.

However, in the discriminant-based classifier used in our methodology, features selected to be included in a predictive set must by necessity not correlate with each other. This is because if there exists even a single correlated pair of genes in the predictor set, the classification results would be unreliable, since the covariance matrix in this case would be singular or close to singular, and hence not inversible. (Inversion of the covariance matrix is essential to calculate the discriminant function.) Therefore the predictor sets obtained through our method contain genes with no or very low correlation to each other with respect to class distinction.

There are a few previous reports that have described the use of GAs in the analysis of gene expression data. The approach proposed by Li *et al.* (2001) combines a GA with KNN rules in order to select a subset of predictive genes for phenotypic classification. We believe, however, that the GA/KNN strategy may not be optimal for multi-class scenarios for several reasons. Firstly, most of the distance metrics (Euclidean, Pearson, etc.) used to determine the *k* neighbors of a sample invariably become less sensitive as data dimensionality increases (Keller *et al.*, 2000), and samples might also be unclassifiable if no satisfactory majority vote is obtained from the *k* neighbors. Secondly, KNN requires relatively large computational memory requirements to store all training data. Thirdly, in that report the final predictor

set used for classification (of test samples) was not one of the optimal or near-optimal sets obtained directly from the GA/KNN method. Instead, individual genes were ranked based on their frequency of selection into 10 000 near-optimal sets by the said method. The *n* top-ranked genes were then picked as predictor genes. This secondary operation, which essentially constitutes a rank-based gene-selection approach, was performed so as to 'break up' the good predictor sets as directly selected by the GA/KNN approach into *different* sets. This appears to pose no problem for binary class datasets, but in multi-class datasets, where gene interactions are believed to be more complex, the same accuracy may not be obtained.

In conclusion, this report shows that highly accurate classification results can be obtained using a combination of GA-based gene selection and discriminant-based classification methods. The accuracy achieved (95% for NCI60) is better than other published methods employing the same dataset. Other advantages of the GA-based approach are that it automatically determines the optimal predictor set size and the delivery of predictive accuracies that are comparable to other methods using classifier genesets of substantially fewer features than previously required. We propose that the methodology outlined here, as well as related methods, may represent useful alternatives in the analysis and exploration of complex multi-class gene expression data.

Acknowledgements

We thank Professor Lim Mong King and Professor Hui Kam Man for their support and encouragement. This work was funded in part by an NCC Core Grant to P.T.

REFERENCES

Alizadeh,A.A., Eisen,M.B., Davis,R.E., Ma,C., Lossos,I.S., Rosenwald,A., Boldrick,J.C., Sabet,H., Tran,T., Yu,X. *et al.* (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503–511.

Alon,U., Barkai,N., Notterman,D.A., Gish,K., Ybarra,S., Mack,D. and Levine,A.J. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *PNAS*, **96**, 6745–6750.

Baker,J.E. (1987) Reducing Bias and Inefficiency in the Selection Algorithm. In Grefenstette,J.J. (ed.), *Proceedings of the Second International Conference on Genetic Algorithms and their Application*. Lawrence Erlbaum Associates, Hillsdale, New Jersey, pp. 14–21.

Bittner,M., Meltzer,P., Chen,Y., Jiang,Y., Seftor,E., Hendrix,M., Radmacher,M., Simon,R., Yakhini,Z., Ben-Dor,A. *et al.* (2000) Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, **406**, 536–540.

Dudoit,S., Fridlyand,J. and Speed,T. (2000) Comparison of discrimination methods for the classification of tumors using gene expression data. *JASA* (in press). (*Berkeley Stat. Dept. Technical Report* #576).

Goldberg,D.E. (1989) *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley, New York.

Goldberg,D.E. and Deb,K. (1991) A comparative analysis of selection schemes used in genetic algorithms. In Rawlins,G. (ed.), *Foundations of Genetic Algorithms*. Morgan Kaufmann, Berlin, pp. 69–93.

Golub,T.R., Slonim,D.K., Tamayo,P., Huard,C., Gaasenbeek,M., Mesirov,J.P., Coller,H., Loh,M.L., Downing,J.R., Caligiuri,M.A. *et al.* (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Grefenstette,J.J. and Baker,J.E. (1989) How Genetic Algorithms Work: A Critical Look at Implicit Parallelism. In Schaffer,J.D. (ed.), *Proceedings of the Third International Conference on Genetic Algorithms*. Morgan Kaufmann, San Mateo, CA, pp. 20–27.

Hall,M.A. and Smith,L.A. (1998) Practical feature subset selection for machine learning. In McDonald,C. (ed.), *Proceedings of Australasian Computer Science Conference*. Springer, Singapore, pp. 181–191.

Haupt,R.L. and Haupt,S.E. (1998) *Practical Genetic Algorithms*. Wiley, New York.

Holland,J. (1992) *Adaptation in Natural and Artificial Systems*, 2nd edition, MIT Press, Cambridge, Massachusetts.

James,M. (1985) *Classification Algorithms*. Wiley, New York.

Keller,A.D., Schummer,M., Hood,L. and Ruzzo,W.L. (2000) Bayesian Classification of DNA Array Expression Data. Technical Report UW-CSE-(2000)-08-01, Department of Computer Science and Engineering, University of Washington, Seattle.

Li,L., Weinberg,C.R., Darden,T.A. and Pedersen,L.G. (2001) Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. *Bioinformatics*, **17**, 1131–1142.

Ramaswamy,S., Tamayo,P., Rifkin,R., Mukherjee,S., Yeang,C.H., Angelo,M., Ladd,C., Reich,M., Latulippe,E., Mesirov,J.P. *et al.* (2001) Multi-class cancer diagnosis using tumor gene expression signatures. *PNAS*, **98**, 15149–15154.

Ross,D.T., Scherf,U., Eisen,M.B., Perou,C.M., Rees,C., Spellman,P., Iyer,V., Jeffrey,S.S., Van de Rijn,M., Waltham,M. *et al.* (2000) Systematic Variation in Gene Expression Patterns in Human Cancer Cell Lines. *Nature Genet.*, **24**, 227–235.

Spears,W.M. and De Jong,K.A. (1991) On the virtues of uniform crossover. In *Proceedings of the Fourth International Conference on Genetic Algorithms*. Morgan Kaufmann, La Jolla, CA, pp. 230–236.

Staunton,J.E., Slonim,D.K., Coller,H.A., Tamayo,P., Angelo,M.J., Park,J., Scherf,U., Lee,J.K., Reinhold,W.O., Weinstein,J.N. *et al.* (2001) Chemosensitivity prediction by transcriptional profiling. *PNAS*, **98**, 10787–10792.

Yeang,C.H., Ramaswamy,S., Tamayo,P., Mukherjee,S., Rifkin,R.M., Angelo,M., Reich,M., Lander,E.S., Mesirov,J.P. and Golub,T.R. (2001) Molecular classification of multiple tumor types. *Bioinformatics*, **17**, S316–S322.

Zhang,H., Yu,C.Y., Singer,B. and Xiong,M. (2001) Recursive partitioning for tumor classification with gene expression microarray data. *PNAS*, **98**, 6730–6735.