Available online at www.sciencedirect.com

SCIENCE DIRECT®

Genomics 85 (2005) 16–23

GENOMICS

www.elsevier.com/locate/ygeno

A robust hybrid between genetic algorithm and support vector machine for extracting an optimal feature gene subset

Li Li[a,b], Wei Jiang[a], Xia Li[a,b,c,*], Kathy L. Moser[d], Zheng Guo[a,b,c], Lei Du[a], Qiuju Wang[e], Eric J. Topol[f], Qing Wang[f], Shaoqi Rao[a,e,f,*]

[a]Department of Bioinformatics, Harbin Medical University, Harbin 150086, People's Republic of China
[b]College of Biological Science and Technology, Tongji University, Shanghai 200092, People's Republic of China
[c]Department of Computer Science, Harbin Institute of Technology, Harbin 150080, People's Republic of China
[d]Department of Medicine, Institute of Human Genetics, University of Minnesota, Minneapolis–St. Paul, MN 55455, USA
[e]Department of Otorhinolaryngology/Head and Neck Surgery, Institute of Otolaryngology, Chinese PLA General Hospital,
Beijing 100853, People's Republic of China
[f]Department of Cardiovascular Medicine and Department of Molecular Cardiology, The Cleveland Clinic Foundation, Cleveland, OH 44195, USA

Received 25 July 2004; accepted 18 September 2004
Available online 2 November 2004

Abstract

Development of a robust and efficient approach for extracting useful information from microarray data continues to be a significant and challenging task. Microarray data are characterized by a high dimension, high signal-to-noise ratio, and high correlations between genes, but with a relatively small sample size. Current methods for dimensional reduction can further be improved for the scenario of the presence of a single (or a few) high influential gene(s) in which its effect in the feature subset would prohibit inclusion of other important genes. We have formalized a robust gene selection approach based on a hybrid between genetic algorithm and support vector machine. The major goal of this hybridization was to exploit fully their respective merits (e.g., robustness to the size of solution space and capability of handling a very large dimension of feature genes) for identification of key feature genes (or molecular signatures) for a complex biological phenotype. We have applied the approach to the microarray data of diffuse large B cell lymphoma to demonstrate its behaviors and properties for mining the high-dimension data of genome-wide gene expression profiles. The resulting classifier(s) (the optimal gene subset(s)) has achieved the highest accuracy (99%) for prediction of independent microarray samples in comparisons with marginal filters and a hybrid between genetic algorithm and K nearest neighbors.

Keywords: Feature gene selection; Genetic algorithm; Support vector machine; Microarray

Introduction

Microarray technology allows simultaneous measurements of expression levels of thousands of genes and has provided opportunities for genetic dissection of complex diseases [1,2]. Systematic analysis of gene expression profiles using data mining integrated with bioinformatics knowledge support can provide an attractive way to identify key genes for predicting the affection state of a patient and to allow the investigation of the underlying genetic pathogenesis of a complex disease at the molecular level.

The characteristics of high-throughput gene expression profile data are the high dimension (usually thousands), noise, and relatively small sample size (usually tens). Therefore, data reduction techniques with flavors of machine learning have been a focus of the methodological development in recent years. Their utilizations are not

* Corresponding authors. S. Rao is to be contacted at Department of Cardiovascular Medicine and Department of Molecular Cardiology, The Cleveland Clinic Foundation, 9500 Euclid Avenue, Cleveland, OH 44195. Fax: (216) 444 2682. X. Li, Fax: +86 451 8666 9617.

E-mail addresses: Lixia@ems.hrbmu.edu.cn (X. Li), raos@ccf.org (S. Rao).

0888-7543/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2004.09.007

simply to reduce the dimension of feature genes and to avoid the curse of dimension. More importantly, their applications can improve the performance of the resulting classifiers and exclude the interference of a large number of irrelevant genes by hunting for the feature genes relevant to a complex biological phenomenon like a disease.

Methods for data reduction, or specifically for feature gene selection in the context of microarray data analysis, can be classified into three major groups: marginal filters, wrappers, and embedded approaches [3,4]. Filtering approaches such as *t* test and nonparametric scoring [5,6] select a feature based on its marginal contribution without accounting for its interactions with other features. The selection process is separated from the classification process because a classifier is not built. Another group of methods for feature gene selection are wrappers and a hybrid of filtering and wrapping [3,7]. In a wrapping approach, the algorithm for feature gene subset selection exists as a wrapper around an induction algorithm. It conducts a search for a good subset using the induction algorithm itself as part of the function evaluating feature subsets. The induction algorithm is run on microarray data, usually partitioned into internal learning and external test sets. The feature subset with the highest evaluation is chosen as the final set on which to build a classifier. Because the feature subset selection by a wrapper is able to couple tightly with the decision mechanism of a classifier, maximal classification accuracy on a separate test set can be attained. It can be argued that the intrinsic capacity of several supervised classifiers to discard (and not include in the final model) a subset of the features (e.g., decision trees, IF-THEN decision rules) could be a third method to perform feature subset selection, known as "embedded" [8]. Nevertheless, wrappers and embedded algorithms are often not clearly distinguished, although feature searching strategies are only slightly different.

A procedure for feature gene selection can be divided into two distinct steps: the search step and the evaluation step. Although an exhaustive search over the entire feature gene space and branch-and-bound algorithm [3,9] can lead to an optimal solution, the two approaches have rarely been used in analysis of high-dimension microarray data because of computational costs. Thus, heuristic search methods such as greedy climbing hill [10], the best first method, and genetic algorithm (GA) [11–13] have received increased attention for dimension reduction. The first two methods search for an optimum by changing the local search space, though the best first method allows backtracking along the search path. Thus, they fail to capture many important feature subsets. In either approach, once a feature is taken in (or removed out), it will never be considered again. Genetic algorithm is an adaptive search engine that emulates the natural selection process in genetics [14]. It employs a population of competing solutions-evolved over time by crossover and mutation and selection-to converge to an optimal solution. The solution space is efficiently searched in parallel and a set of solutions instead of a single solution is computed to avoid becoming trapped in a local optimum that can occur with other search techniques. In addition, its robustness to size of search space and the underlying multivariate distribution assumptions has made it a promising method for feature gene selection over a high-dimension space. Nevertheless, genetic algorithm itself is merely a searching algorithm. To apply this algorithm to microarray data analysis, several issues need to be resolved. First, a suitable fitness function has to be defined to map the biological reality [15]. Second, the number of genes contained in the optimal feature gene subsets can be large at early generations. The coupled classifier evaluating fitness of a candidate gene subset must have the capacity for handling data of a very high-dimension feature gene space but of a limited sample space. There are many potential choices for a classifier, but the majority can deal with only a limited dimension of features. Support vector machine (SVM) [16–19], which resulted from recent advances in statistical learning theory and machine learning, is an exception. Its unique advantages for treating this particular data structure, for avoiding overfitting and dimensional curse, and for nonlinear modeling have made it a popular tool in pattern recognition.

In this study, we propose a hybrid between genetic algorithm and support vector machine (termed GA-SVM) that can fully utilize the unique merits of the two data mining tools. Genetic algorithm is used as the search engine, while support vector machine is used as the classifier (the evaluator). We apply the proposed approach to the microarray data of diffuse large B cell lymphoma. This application has demonstrated its high potential as a powerful tool for mining high-dimension data such as genome-wide gene expression profiles.

Method

First, we randomly generate the fixed-length binary strings for *N* individuals to build up the initial population. Each string represents a feature subset (coding of the feature subset) and the values at each position in the string are coded as either presence or absence of a particular feature. Then, we calculate the fitness (i.e., how well a feature subset survives over the specified evaluation criteria) for each feature subset. We adopt classification accuracy as the fitness index (eval) that is evaluated using a linear SVM. Better feature subsets have a greater chance of being selected to form a new subset through a crossover or mutation. Mutation changes some of the values (thus adding or deleting features) in a subset randomly. Crossover combines different features from a pair of subsets into a new subset. The algorithm is an iterative process in which each successive generation is produced by applying genetic operators to the members of the current generation. In this manner, good subsets are "evolved" over time until the

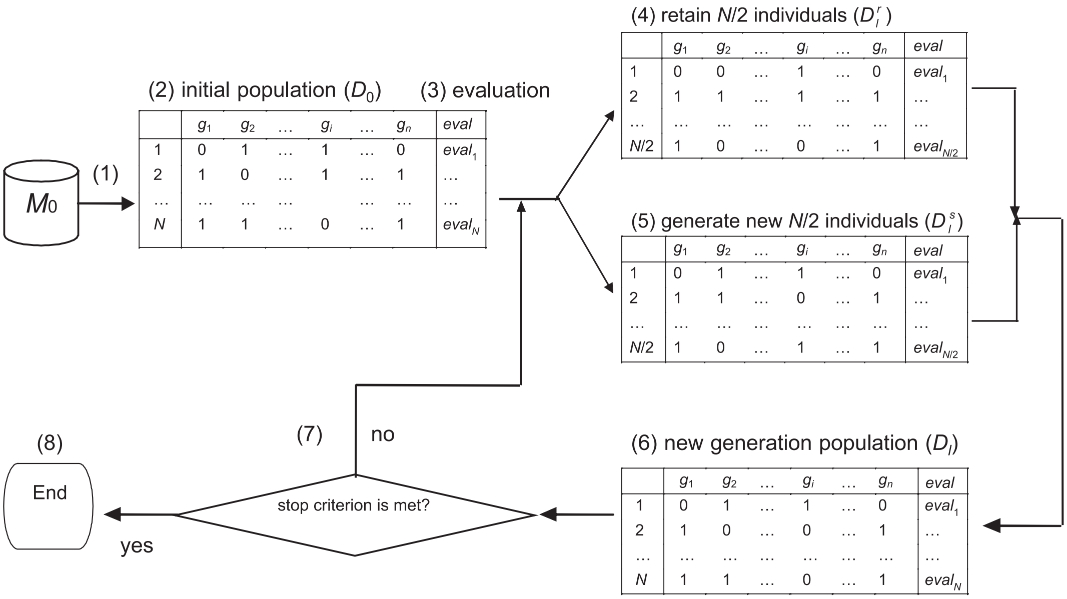

Fig. 1. GA-SVM algorithm. *M, D, G,* and eval denote gene expression profile matrix, population, gene subset, and evaluation index, respectively.

stopping criteria are met. Thus, coding feature subset, population initialization, fitness computation, genetic operation, and control parameter assignment (population size, the maximal number of generations, and the selection probability) are the major elements of the GA-SVM method. Fig. 1 shows the flow chart of the method, which is realized on a Matlab platform. The corresponding Matlab source codes are available for download, from the Web site http://www.biocc.net/ga_svm. The detailed computational procedures are given in Fig. 1 and as follows.

(1) $M_0 \leftarrow$ Read gene expression profile matrix from database; n_0 is the number of genes in M_0.

(2) $D_0 \leftarrow$ Generate N individuals (the initial population) randomly. Each individual has an n-length vector of bits of either 1 or 0. The frequencies of the two codes are set to be equal to halve the number of feature genes successively for each generation.

(3) For each individual j in D_0, determine:

$G_j \leftarrow$ a gene subset corresponding to individual j. If bit i equals 1, include g_j in the subset.

$M_j \leftarrow$ gene expression profile submatrix.

eval$_j \leftarrow$ eval(M_j) and selection probability,

$$p_j = \left(eval_j\right) / \sum_{j=1}^{N} eval_j.$$

(4) $D_l^r \leftarrow$ Retain $N/2$ individuals with the highest evaluations.

(5) $D_l^s \leftarrow$ Select individuals randomly based on their selection probabilities, apply a single crossover or multiple single-point mutations to produce new $N/2$ individuals. In detail, we select at random two individuals from the population D_0 and then perform a single crossover four times to generate eight individuals. Based on their evaluations, we select two individuals with the highest fitness to be the members in D_l^s. Mutation is performed separately as follows. An individual in D_0 is randomly selected to be the subject for mutagenesis, with a point mutation rate of 0.2 at each locus. For each batch of four mutated individuals, one individual with the best fitness is selected to be a member in D_l^s.

(6) $D_l \leftarrow D_l^r \cup D_l^s$.

(7) $D_0 \leftarrow D_l$.

(8) End $\leftarrow$ Output the optimal individual(s) based on the evaluation with eval$_j$ = eval(M_j). For each M_j, we classify the microarray samples with genes contained in individual j using a SVM. The classifier [16] is

$$\hat{y} = f(x) = \operatorname{sgn}\left\{ \sum_{i=1}^{L} a_i y_i K(x_i \cdot x) - b \right\},$$

where x is a test sample vector and x_i is the learning sample vector. L is the number of learning samples. y_i is a class indicator (for a two-class application, +1 for the first class; -1 for the second class), a_i is a nonnegative Lagrange multiplier associated with x_i and $a_i \neq 0$ for support vectors. sgn{ } is the sign function and $K(x_i \cdot x)$ is the kernel function (for linear kernel, $K(x_i \cdot x) = x_i \cdot x$, i.e., their inner product). Then, the accuracy of classification,

$$acc = \left(\sum_{t=1}^{T} I(y_t, \hat{y}_t) \right) / T,$$

where T is the number of test samples and

$$I(y_t, \hat{y}_t) = \begin{cases} 1 & \text{if} \quad y_t = \hat{y}_t \\ 0 & \text{otherwise} \end{cases}.$$

In this study, a fivefold cross-validation (CV) resampling approach is used to construct the learning and test sets. First, the two-class samples are randomly divided into five nonoverlapping subsets of roughly equal size, respectively. A random combination of the subsets for the two classes constitutes a test set and the total remaining subsets are used as the learning set. The fivefold CV resampling produces 25 pairs of learning and test sets. Individual j is evaluated by the averaged value over the 25 pairs, i.e.,

$$eval_j = \left(\sum_{k=1}^{25} acc_k \right) / 25,$$

where k is the replicate number and acc_k is the classification accuracy for the kth replicate.

In the GA-SVM algorithm, the optimization of the feature gene subset(s) is realized via survival competitions. For each generation, we retain 50% of the high-valued individuals that will enter the next generation directly without mutations and crossovers to keep these optimal solutions unchanged. To avoid the loss of the putative important feature genes, we initially include about half of the genes in each individual so that the number of genes contained in the optimal feature gene subsets can be very large at the early generations. Then, we adopt a stepwise data reduction procedure to shrink (or to increase) the dimension of the feature subsets to achieve the minimal size and the highest classification accuracy. Stepwise data reduction is realized by halving the number of feature genes initially at each generation (i.e., coding half of bits with null) and the evolution forces (crossover and mutation). The gene expression matrices from the optimal individuals serve as the data on which a new round of iteration is performed. The data reduction process stops when the drop (or difference) in classification accuracy reaches 0.001.

Numerical application

Alizaden et al. [20] analyzed the gene expressions of malignant lymphoma, specifically the diffuse large B cell lymphoma (DLBCL), follicular lymphoma, and chronic lymphocytic leukemia, using chips with 18,000 gene transcripts. In the present study, we use the gene expression profile dataset consisting of two subtypes of DLBCL, 21 activated B-like DLBCL (AB-like DLBCL) and 21 germinal center B-like DLBCL (GCB-like DLBCL) samples, and 4026 genes marginally filtered by the authors (available at http://llmpp.nih.gov/lymphoma/data.shtml). Among the 4026 genes, 6% have missing values and are imputed by the KNN Impute algorithm [21] prior to the GA-SVM analysis. The KNN Impute algorithm uses the expression profiles of K nearest neighbors (here $K = 5$) to impute the missing values for the target gene. Therefore, M_0 is a matrix with 42 rows and 4026 columns. The number of individuals (N) of 40 and the maximal generations of 50 are empirically determined, respectively, to allow the solution space to be sufficiently searched and convergence to the best minimal subset attainable with the evolution time. An optimal feature gene subset(s) is sought for classification of the two subtypes of DLBCL.

The relationship between classification accuracy and the number of support vectors

We start with all 4026 genes and then step down to reduce the dimension of the feature genes successively for 14 iterations. The number of feature genes in the best individual at the successive generations varied as follows: 4026, 1995, 984, 504, 256, 132, 70, 41, 25, 18, 13, 7, 7, and 7 until the number of genes included was no longer changed. It should be noted that the number of feature genes for each individual at each generation is not necessarily the same because of the variation contributed by evolution forces (thus adding or deleting features). Fig. 2 shows the changes in the maximum accuracy (left plot) or number of support vectors (right plot) over the iterations. These values are averaged over 25 replicates generated by the fivefold cross-validation described previously.

As shown in Fig. 2, the classification accuracy increases with the successive reduction in the number of genes contained in the prediction set until reaching the plateau, starting from the valley of 0.9399 (no selection, all the genes included). In contrast, the number of support vectors decreases dramatically with the iterations. It is well known that the optimal hyperplane is supported by support vectors and the final classifier(s) contains only the inner product of test sample vectors and support vectors. The support vectors contain all

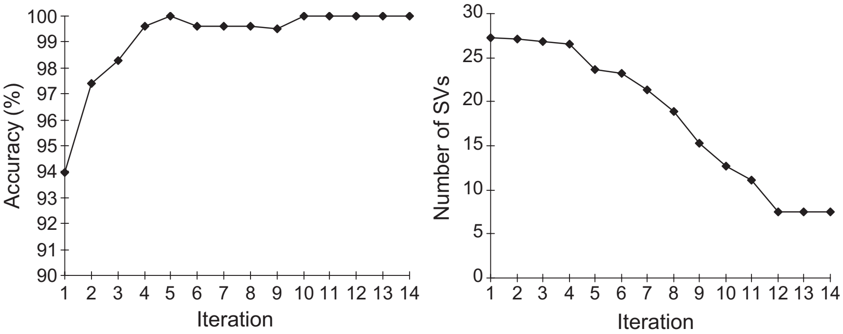

Fig. 2. Changes in the maximum accuracy of the SVM classifier (left) and changes in support vectors (right) over iterations.

the useful information for classification. In other words, classification relies only on the support vectors, which play a pivotal role in sample classification. If a set of learning data can be divided by an optimal hyperplane, the expected upper bound of the error rate for the test data is simply the averaged proportion of support vectors over the sample size for learning data [22,23]. Thus, an optimal hyperplane with the smallest number of support vectors has a wider adaptability.

Comparison of GA-SVM with alternative feature selection methods for biological classification

We fix the number of feature genes in the prediction subset as 7. Hence, the subsets constructed by marginally filtering include (1) seven randomly selected genes (as a control), (2) the first seven genes ranked by a *t* test, and (3) seven genes ranked by nonparametric scoring (nonparat). For comparison, we also perform an alternative hybrid between genetic algorithm and *K* nearest neighbors (GA-*K*NN, *K* = 5). Surprisingly, GA-*K*NN has produced an optimal subset of seven genes, also, after 13 generations of evolutions. We estimate the accuracy of the five gene subsets (plus the GA-SVM selected subset: CYSLTR1, MME, D13S2489E, PIK3CG, SHMT2, Hs.348293, Hs.291994) using five classifiers: SVM, Fisher linear discriminant (Fisher) [24], logistic regression (Logist), minimal distance (MinDis), and *K*NN (*K* = 3) [25–27]. The results as shown in Fig. 3 indicate that the subset identified by GA-SVM has significantly higher performance than marginal filters. Interestingly, both GA-SVM and GA-*K*NN achieve the highest classification accuracy (99 and 95%, respectively) when the same approach is used for both feature selection and prediction.

Validation via biological evidence

For an optimal feature gene subset, not only is GA-SVM capable of distinguishing disease classes, but also the feature gene subset may correspond to a specific genetic path that can distinguish the target phenotypes (in this study, the two subtypes of DLBCL). To look for knowledge support, we carry out Gene Ontology (GO) database mining [28] (also see http://www.geneontology.org) using the well-documented gene CCR7 [29] as a reference to investigate the relationship between the feature genes selected by us and CCR7. GO annotates a gene with three components: molecular function, biological process, and cellular constitution. CCR7 is a target gene of NF-κB, which is expressed at high levels in AB-like DLBCL and low levels in GCB-like DLBCL [29]. The role of CCR7 as a biomarker to separate the two subtypes of DLBCL is well established. CCR7 is located on chromosome 17q12-q21.2 and its protein product is a member of the G-protein-coupled receptor family. This receptor is expressed in many types of lymphoid tissues and activates B and T lymphocytes (http://genome_www5.stanford.edu/cgi_bin/source/SourceSearch).

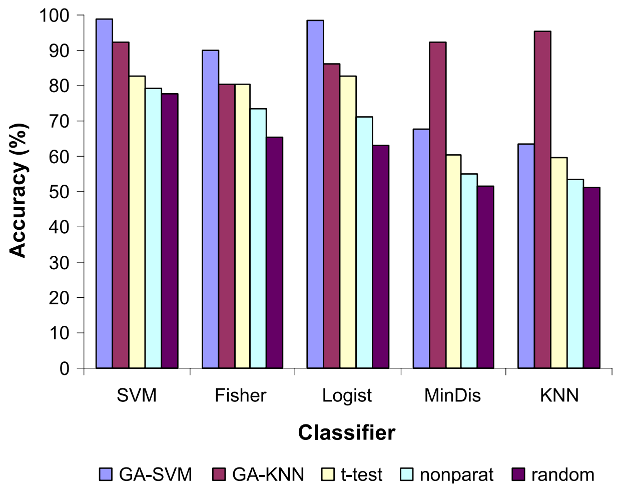

Fig. 3. Performance of biological classification for five feature selection methods (GA-SVM, GA-*K*NN, *t* test, nonparat, random).

We annotate the identified feature genes CYSLTR1, MME, D13S2489E, PIK3CG, SHMT2, Hs.348293, and Hs.291994 and the previously documented gene CCR7 to nodes of GO. Six genes have functional annotations (except for Hs.348293 and Hs.291994). Table 1 shows the results from the database mining. Feature genes CYSLTR1, PIK3CG, and CCR7 are involved in an identical biological process (GO:0007186, G-protein-coupled receptor protein signaling pathway). CYSKTR1 and CCR7 also play a role in cytosolic calcium ion concentration elevation (GO:0007204). Both D13S2489E and CCR7 are relevant to the antimicrobial humoral response (GO:0006960). CYSLTR1, MME, D13S2489E, and CCR7 are located in the same subcellular location (integral to plasma membrane, see GO:0005887). Thus, these genes may be involved in a common molecular pathway or have a function similar to that of CCR7 [30]. Surprisingly, Hs.291994, which is currently not annotated in GO, is located on the same chromosome as SHMT2 and D13S2489E. We suspect that they may share some functional relationship. These results support the hypothesis that DLBCL is a result of a disruption of NF-κB activity or the G-protein−coupled receptor signal transduction pathways. There is differential expression of the genes involved in the two pathways between AB-like DLBCL and GCB-like DLBCL. In addition, all the feature genes identified in this investigation were from the germinal center B cell. So it is not surprising that Alizadeh et al. [20] could discover two subtypes of DLBCL using the germinal center B cell signatures.

Discussion

Microarray data analysis warrants special strategies because gene expression datasets have several important characteristics, including relatively few samples with a large dimension of feature gene space and a high signal-to−noise

Table 1
Functions of GA-SVM feature genes and CCR7

Gene name (UniGene ID)	Biological process	Cellular component	Molecular function
CYSLTR1 (Hs.124401)	GO:0007204 Cytosolic calcium ion concentration elevation GO:0007186 G-protein-coupled receptor protein signaling pathway	GO:0005887 Integral to plasma membrane GO:0005624 Membrane fraction	GO:0004974 Leukotriene receptor activity
MME (Hs.1298)	GO:0006508 Proteolysis and peptidolysis GO:0007048 Oncogenesis GO:0007267 Cell–cell signal	GO:0005887 Integral to plasma membrane	Hydrolase activity GO:0008237 Metallopeptidase activity Zinc ion binding
D13S2489E (Hs.74085)	GO:0006960 Antimicrobial humoral response (sensu Invertebrata) GO:7165 Signal transduction	GO:0005887 Integral to plasma membrane GO:0016021 Integral to membrane	GO:0004872 Receptor activity
PIK3CG (Hs.32942)	GO:0007186 G-protein-coupled receptor protein signaling pathway	1-Phosphatidylinositol 3-kinase complex	Phosphatidylinositol 3-kinase activity GO:0016740 Transferase activity
SHMT2 (Hs.75069)	One-carbon compound metabolism GO:0006544 Glycine metabolism GO:0006563 ScSerine metabolism	GO:0005739 Mitochondrion	GO:0016740 Transferase activity GO:0004372 Glycine hydroxymethyltransferase activity
CCR7 (Hs.1652)	GO:0007186 G-protein-coupled receptor protein signaling pathway GO:0006960 Antimicrobial humoral response (sensu Invertebrata) GO:0007204 Cytosolic calcium ion concentration elevation GO:0006954 Inflammatory response GO:0006935 Chemotaxis	GO:0005887 Integral to plasma membrane	GO:0016493 C–C chemokine receptor activity GO:0019735 Antimicrobial humoral response (sensu Vertebrata) GO:0001584 Rhodopsin-like receptor activity

ratio. Therefore, attempts to analyze the entire microarray data without dimensional reduction may not be appropriate. Microarrays provide massive parallel information and analysis of the data is a nondeterministic polynomial hard problem [31]. A promising direction of research for analysis of microarray data is to perform a gene selection procedure for shrinking the feature gene space and achieving a high generalization performance for biological classification and subtype discovery. Current methods for the feature gene selection can be divided into three groups: marginal filtering, sequential selection, and strictly multivariate modeling. Marginal filtering [5,7] has computational advantages of speed and simplicity, but it cannot utilize the information hidden in gene interactions. Theoretically, distribution-based multivariate modeling is the most powerful and can take into account the multivariate correlation structures of the high-dimension gene expression profiles simultaneously. Nevertheless, its computational complexities confine its application to a very limited dimension. Sequential selection approaches lie in between the above two methods and enjoy extensive application for feature selection, particularly in machine learning and pattern recognition. Recently, hidden Markov models have been increasingly applied to analysis of temporal gene expression data such as for yeast cell cycling [32,33] and can be considered as the fourth group of methods for feature gene selection (and feature gene networking) because additional time serial correlations within and between feature genes have to be taken into account properly.

Many important biological processes (e.g., cellular differentiation during development, aging, disease etiology) are most likely controlled by underlying complex gene-gene interactions and heterogeneous molecular pathways. However, in our own experiences and that of others [34,35], a very good distinction between classes can be achieved by a single influential gene so that the classification error rate is difficult to reduce further by including other important genes. This is the well-known “local optimum” problem that occurs in most search techniques that are sequential in nature. This is one of the reasons the more robust method that couples genetic algorithm and support vector machine is proposed here. Although several authors have explored this analysis strategy and derivatives for biological applications [15,26,36], we are the first to formalize explicitly its biological applications systematically, including a robust fitness evaluation criterion and a stepwise feature subset dimensional reduction technique to achieve the minimal size and the highest classification accuracy simultaneously. We have verified that five genes in the optimal subset of seven genes are significantly differentially expressed between two subtypes of DLBCL. On the other hand, this study suggests that marginal differential expression is not a necessary condition for a gene to be included in the best prediction subset and their interactions with other genes can lead to a better performance for classification [35,37].

SVM is used as the fitness evaluator because of its unique advantage in dealing with high solution space at the early generations of feature gene searching. It is generally believed that the optimal hyperplane with the smallest number of support vectors has a wider adaptability and the highest classification performance [16]. Li et al. [13,26] considered a hybrid between GA and *K*NN, of which *K*NN was used as the evaluator. There are several differences between their hybrid and our proposed method. First, it did not perform the genetic operator crossover, which is efficient for generating diverse solutions and increasing the speed to converge on a better solution(s). Second, to minimize the loss of the putative important feature genes, we search for a much wider solution space starting with half the number of the available genes (4026 genes). The optimal and minimal subset was obtained by a stepwise dimension-reduction technique. In contrast, in several studies, the solution space (the length of the “chromosome”) was arbitrarily determined prior to learning. Third, for a *K*NN, a suitable *K* value has to be defined.

Current applications of microarrays focus on classification or discovery of biological types, and the basic strategy is to search for a single subset that leads to the best prediction of biological types, for example, tumor versus normal tissues. However, with the simultaneous profiling of most, if not all, of the genes of an organism, more challenging but important biological tasks like hunting for biological relevant genes and thereby building the corresponding gene networks can be performed. In a separate paper, we developed a novel approach to hunting disease-relevant genes using an ensemble approach with a recursive partition tree as the feature gene search engine [37]. Alternatively, we may consider the competitor-genetic algorithm-as the search engine. In particular, its parallel searching strategy that avoids becoming trapped in a local optimum and its robustness to size of searching space and the underlying multivariate assumptions would render it a very promising method for more complex biological tasks as well. Our results show that GA coupled with SVM warrants further consideration of application to a variety of biological investigations using microarray approaches.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grants 30170515 and 30370798), the National High Tech Development Project, the Chinese 863 Program (Grants 2003AA2Z2051 and 2002AA2Z2052), and the 211 Project, the Tenth “Five-Year” Plan, Harbin Medical University.

References

[1] M. Schena, D. Shalon, R.W. Davis, P.O. Brown, Quantitative monitoring of gene expression patterns with a complementary DNA microarray, Science 270 (1995) 467–470.

[2] U. Alon, et al., Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, Proc. Natl. Acad. Sci. USA 96 (1999) 6745–6750.

[3] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artif. Intell. 97 (1997) 273–324.

[4] I. Tsamardinos, C.F. Aliferis, Towards principled feature selection: relevance, filters and wrappers, Ninth International Workshop on Artificial Intelligence and Statistics, Key West, FL, 2003.

[5] P.J. Park, M. Pagano, M. Bonetti, A nonparametric scoring algorithm for identifying informative genes from microarray data, Pac. Symp. Biocomput. (2001) 52–63.

[6] L. Li, X. Li, Z. Guo, Efficiency of two filters for feature gene selection, Life Sci. Res. 7 (2003) 372–396. In Chinese.

[7] E.P. Xing, M.I. Jordan, R.M. Karp, Feature selection for high-dimensional genomic microarray data, Machine Learning: Proceedings of the Eighteenth International Conference, San Francisco, Morgan Kaufmann, San Mateo, CA, 2001.

[8] A.L. Blum, P. Langley, Selection of relevant features and examples in machine learning, Artif. Intell. 97 (1997) 245–271.

[9] M.J. Brusco, An enhanced branch-and-bound algorithm for a partitioning problem, Br. J. Math. Stat. Psychol. 56 (2003) 83–92.

[10] R.C. Holte, Combinatorial auctions, knapsack problems, and hill-climbing search, in: E. Stroulia, S. Matwin (Eds.), Proceedings of the 14th Biennial Conference of the Canadian Society on Computational Studies of Intelligence: Advances in Artificial Intelligence, Springer, Ottawa, ON, 2001, pp. 57–66.

[11] C.R. Houck, J.A. Joines, M.G. Kay, J.R. Wilson, Empirical investigation of the benefits of partial Lamarckianism, Evol. Comput. 5 (1997) 31–60.

[12] F.M. Stefanini, A. Camussi, The reduction of large molecular profiles to informative components using a genetic algorithm, Bioinformatics 16 (2000) 923–931.

[13] L. Li, T.A. Darden, C.R. Weinberg, A.J. Levine, L.G. Pedersen, Gene assessment and sample classification for gene expression data using a genetic algorithm/k-nearest neighbor method, Comb. Chem. High Throughput Screen. 4 (2001) 727–739.

[14] M. Hall, Correlation-Based Feature Selection for Machine Learning, Department of Computer Science, University of Waikato, Hamilton, New Zealand, 1998.

[15] S.J. Cho, M.A. Hermsmeier, Genetic algorithm guided selection: variable selection and subset selection, J. Chem. Inf. Comput. Sci. 42 (2002) 927–936.

[16] M.P. Brown, et al., Knowledge-based analysis of microarray gene expression data by using support vector machines, Proc. Natl. Acad. Sci. USA 97 (2000) 262–267.

[17] C.Z. Cai, W.L. Wang, L.Z. Sun, Y.Z. Chen, Protein function classification via support vector machine approach, Math. Biosci. 185 (2003) 111–122.

[18] S. Hua, Z. Sun, A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach, J. Mol. Biol. 308 (2001) 397–407.

[19] T.S. Furey, et al., Support vector machine classification and validation of cancer tissue samples using microarray expression data, Bioinformatics 16 (2000) 906–914.

[20] A.A. Alizadeh, et al., Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, Nature 403 (2000) 503–511.

[21] O. Troyanskaya, et al., Missing value estimation methods for DNA microarrays, Bioinformatics 17 (2001) 520–525.

[22] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines, Cambridge University Press, Cambridge, UK, 2000.

[23] Z. Guo, X. Li, S. Rao, Analysis of Medical Data: an Introduction to Bioinformatics, Harbin Pub., Harbin, China, 2001.

[24] E.S. Gilbert, The effect of unequal variance–covariance matrices on Fisher's linear discriminant function, Biometrics 25 (1969) 505–515.

[25] T.R. Golub, et al., Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science 286 (1999) 531–537.

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

[27] T.M. Cover, P.E. Hart, Nearest neighbor pattern classification, IEEE Trans. Inform. Theor. 13 (1967) 21–27.

[28] M. Ashburner, et al., Gene ontology: tool for the unification of biology, The Gene Ontology Consortium, Nat. Genet. 25 (2000) 25–29.

[29] R.E. Davis, K.D. Brown, U. Siebenlist, L.M. Staudt, Constitutive nuclear factor kappaB activity is required for survival of activated B cell-like diffuse large B cell lymphoma cells, J. Exp. Med. 194 (2001) 1861–1874.

[30] R.F. Murphy, M.V. Boland, M. Velliste, Towards a systematics for protein subcellular location: quantitative description of protein localization patterns and automated analysis of fluorescence microscope images, Proc. Int. Conf. Intell. Syst. Mol. Biol. 8 (2000) 251–259.

[31] H.B. Burke, Discovering patterns in microarray data, Mol. Diagn. 5 (2000) 349–357.

[32] X. Ji, J. Li-Ling, Z. Sun, Mining gene expression data using a novel approach based on hidden Markov models, FEBS Lett. 542 (2003) 125–131.

[33] A. Schliep, A. Schonhuth, C. Steinhoff, Using hidden Markov models to analyze gene expression time course data, Bioinformatics 19 (Suppl 1) (2003) i255–i263.

[34] A. Szabo, et al., Multivariate exploratory tools for microarray data analysis, Biostatistics 4 (2003) 555–567.

[35] A. Chilingaryan, N. Gevorgyan, A. Vardanyan, D. Jones, A. Szabo, Multivariate approach for selecting sets of differentially expressed genes, Math. Biosci. 176 (2002) 59–69.

[36] X.-W. Chen, Gene selection for cancer classification using bootstrapped genetic algorithms and support vector machines, Proceedings of the Computational Systems Bioinformatics (CSB 2003), IEEE, Stanford, CA, 2003, pp. 504–505.

[37] X. Li, S. Rao, Y. Wang, B. Gong, Gene mining: a novel and powerful ensemble decision approach to hunting for disease genes using microarray expression profiling, Nucleic Acids Res. 32 (2004) 2685–2694.