BIOINFORMATICS

Vol. 17 no. 12 2001
Pages 1131–1142

Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method

Leping Li[1,*], *Clarice R. Weinberg*[1], *Thomas A. Darden*[2] *and Lee G. Pedersen*[2,3]

[1]*Biostatistics Branch,* [2]*Laboratory of Structural Biology, National Institute of Environmental Health Sciences, Research Triangle Park, NC 27709, USA, and* [3]*Department of Chemistry, University of North Carolina, Chapel Hill, NC 27599-3290, USA*

Received on September 4, 2000; revised on November 11, 2000, February 6, 2001 and June 6, 2001; accepted on June 16, 2001

ABSTRACT

Motivation: We recently introduced a multivariate approach that selects a subset of predictive genes jointly for sample classification based on expression data. We tested the algorithm on colon and leukemia data sets. As an extension to our earlier work, we systematically examine the sensitivity, reproducibility and stability of gene selection/sample classification to the choice of parameters of the algorithm.

Methods: Our approach combines a Genetic Algorithm (GA) and the k-Nearest Neighbor (KNN) method to identify genes that can *jointly* discriminate between different classes of samples (e.g. normal versus tumor). The GA/KNN method is a stochastic supervised pattern recognition method. The genes identified are subsequently used to classify independent test set samples.

Results: The GA/KNN method is capable of selecting a subset of predictive genes from a large noisy data set for sample classification. It is a multivariate approach that can capture the correlated structure in the data. We find that for a given data set gene selection is highly repeatable in independent runs using the GA/KNN method. In general, however, gene selection may be less robust than classification.

Availability: The method is available at http://dir.niehs.nih.gov/microarray/datamining

Contact: LI3@niehs.nih.gov

INTRODUCTION

Microarrays have become important tools for profiling global gene expression patterns of cells/tissues. Currently, such studies involve many thousands of genes but only a few hundred or fewer samples. A widely used technique for microarray data analysis is clustering analysis (Alon *et al.*, 1999; Ben-Dor *et al.*, 1999; Bittner *et al.*, 1999; Getz *et al.*, 2000a,b; Hartuv *et al.*, 2000). Clustering analysis groups genes that have correlated patterns of expression that can provide insight into gene–gene interactions and gene function.

We recently proposed a multi-dimensional classification method that not only selects a small fraction of genes that jointly discriminate between different classes of samples, but also assesses the relative predictive importance of all genes for sample classification (Li *et al.*, 2001). A non-parametric pattern recognition approach, the k-Nearest Neighbors (KNN), and a searching tool, a Genetic Algorithm (GA), are employed. Since typical array data consist of a large number of genes and a small number of samples, for a given data set, many subsets of genes that discriminate between different classes of sample may exist. Our strategy is to find many such subsets and then assess the relative importance of genes for sample classification by examining the frequency of membership of the genes in these near-optimal sets. The most frequently selected genes are presumed to be most relevant to sample distinction. The method (Li *et al.*, 2001) has been applied to colon cancer data (Alon *et al.*, 1999) and leukemia data (Golub *et al.*, 1999). We find that the frequency with which genes are selected by the GA/KNN method is statistically informative. When the most frequently selected genes were used for sample classification using a validation set, samples were largely classified correctly. Thus, the GA/KNN method is capable of selecting a subset of predictive genes from a large noisy data set for sample classification. Other computational methods that select a subset of genes for sample classification have also been developed (Golub *et al.*, 1999; Ben-Dor *et al.*, 2000). Detailed discussion of

*To whom correspondence should be addressed.

the differences between the GA/KNN method and other methods can be found elsewhere (Li *et al.*, 2001).

As an extension to our earlier report (Li *et al.*, 2001), we here explore three aspects of the algorithm. First, we consider its sensitivity to choice of algorithm parameters using lymphoma (Alizadeh *et al.*, 2000) and colon data (Alon *et al.*, 1999). We compare results under various choices of 'chromosome' size (e.g. 5, 10, 20, 30, 40 and 50). The patterns of gene selection and the classification reliability of the selected genes (using a validation set) are analyzed. Second, we explore reproducibility of this stochastic search strategy by repeating the gene selection procedure in two independent runs and comparing the results. Third, we examine the sensitivity of gene selection results to the assignment of samples to the training set. We do this by dividing each data set (both lymphoma and colon) into a training set and test set in three different ways, resulting in three different 'training' and 'test' sets for the same data. Each 'training' set is used to select a subset of genes. The patterns of gene selection using the three 'training' sets sampled from the same data are then compared.

METHODS

Data sets

Lymphoma data. The gene expression data (http://llmpp.nih.gov/lymphoma/) contain 4026 genes across 47 samples, of which 24 are referred to as germinal center B-like DLBCL and 23 as activated B-like DLBCL (Alizadeh *et al.*, 2000). The data were originally 'filtered' and *log*-transformed (base 2) (Alizadeh *et al.*, 2000). The data were divided into a training set (the first 34 samples) and a test set (13 samples). The training set was normalized so that the mean and Standard Deviation (SD) for each gene across the 34 training set samples are 0 and 1, respectively. The training set was used to obtain a subset of predictive genes during the training process. The genes identified were subsequently used to classify the independent (test) samples. To mimic what would be done when using the method to classify a single 'unknown', each gene in the test set was normalized by the mean and SD of the corresponding gene in the training set.

Colon data. The original colon data contain the expression levels of 2000 genes across 62 samples, of which 40 are tumor tissue and 22 normal tissue (Alon *et al.*, 1999). In this study, the five samples (N34, N36, T30, T33 and T36) identified as likely to have been contaminated (Li *et al.*, 2001) were removed. The data were then *log*-transformed. The data set was divided into a training set (the first 40 samples) and a test set (17 samples).

Overall methodology

The details of the GA/KNN method were reported elsewhere (Li *et al.*, 2001) (see also http://dir.niehs.nih.gov/microarray/datamining/). The general approach is as follows. Many subsets of genes that can jointly discriminate between different classes of samples (e.g. normal versus tumor) may be identifiable, using the 'genetic algorithm'. When many such combinations are obtained, the frequency with which genes are selected is analyzed. The relative predictive importance of the genes is then inferred: the most frequently selected genes are presumed to be most relevant to sample distinction and, therefore, are used to predict the class membership of independent samples. Further details are given below.

k-nearest neighbors

To apply KNN (see e.g. Massart *et al.*, 1988), each sample was represented by a pattern of expression that consists of d genes. Each sample was then classified according to the class memberships of its k (arbitrarily set to 3) nearest neighbors, as determined by the Euclidean distance in the d-dimensional space. If all of the 3 nearest neighbors of a sample belong to the same class, the sample is classified as that class. Otherwise, the sample is considered unclassifiable.

Genetic algorithm

To use 'brute force' to select from a large set of genes a subset of genes that can jointly discriminate between different classes of samples is not practical. For instance, the number of ways to select 50 from 2000 is approximately 10^{100}. Therefore, an intelligent technique is required to go through relatively fewer combinations to find a useful subset. Since GA has been shown to be effective in searching complex high-dimensional space (Holland, 1975; Goldberg, 1989), it was adapted here as a search tool. Examples of GA applications in chemical and biological problems can be found in Judson (1997). The details of the GA/KNN procedure are reported elsewhere (Li *et al.*, 2001). A brief description of the four components (chromosome, fitness, selection and mutation) of the GA is given below.

Each 'chromosome' (a mathematical entity, not the biological chromosomes) consists of d distinct genes that are initially randomly selected from all genes (4026 in lymphoma and 2000 in colon). A set of chromosomes (typically 100) is constructed to form a 'population' or a 'niche'. For a typical run, 10 niches are allowed to evolve in parallel. The 'fitness' (merit) of each chromosome is determined by its ability to classify the training set samples according to the KNN procedure. If the class memberships of a training set sample and its three nearest neighbors in the particular d-dimensional space defined by a chromosome agree, a score of 1 is assigned to

that sample. These agreement scores are summed across the training set and we refer to this sum divided by the number of training samples (e.g. 34 in lymphoma and 40 in colon) as the cross-validation R^2. A maximal cross-validation R^2 is 1. The single best chromosome (with the largest cross-validation R^2) from each niche is entered into the respective next generation for that niche deterministically and the remaining 99 positions are filled based on sampling that is weighted according to the relative fitness of the chromosomes in the parent generation (probabilistically).

Once a chromosome is chosen for transmission to the next generation, between 1 and 5 of its genes are randomly selected for mutation, with probabilities, 0.531 25, 0.25, 0.125, 0.0625, and 0.031 25, respectively. In this way, a single mutation is given the highest probability while simultaneous multiple mutations have lower probability. However, when d is small (e.g. 5 or 10), only one gene is selected for mutation with a probability of 1.0. Once the number of genes in the chromosome to be mutated has been determined, these genes are randomly selected and replaced randomly, using genes not already in the chromosome. The niche is allowed to evolve by repeating these steps until at least one chromosome achieves a targeted R^2 for classification of the training set.

A solution was considered to have been found when the cross-validation R^2 reached at least 31/34 for lymphoma data and 38/40 for colon data. With this strategy, genes of predictive importance that fail for a few samples can still be selected. Once the termination criterion is met from any of the 10 niches, the selected high-R^2 chromosome is set aside and the GA/KNN procedure is restarted. The process is terminated when 10 000 high-R^2 (not necessarily distinct) chromosomes have been obtained. Each such chromosome corresponds to a set of d genes that discriminates fairly well between different classes of samples in the training set (e.g. germinal center B-like DLBCL versus activated B-like DLBCL for lymphoma or normal versus tumor for colon). The frequency with which genes were selected is then analyzed. A set of the most frequently selected genes is then validated by using it to classify the test set samples. A flowchart for data analysis is given in Figure 1.

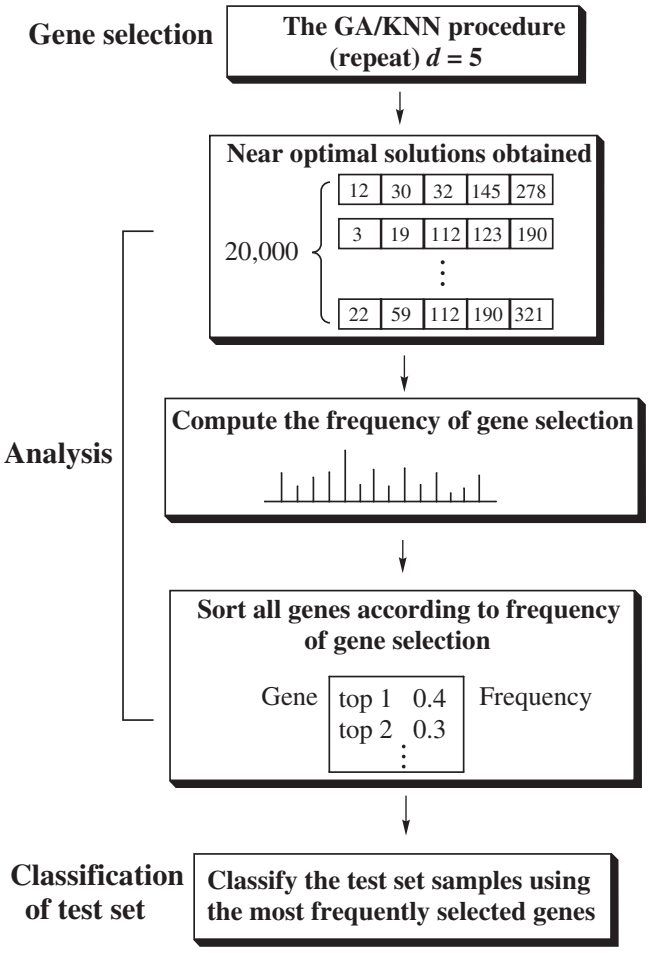

Fig. 1. A flowchart for gene selection, data analysis, and classification of test set samples. Each module is highlighted. For clarity purposes, only $d = 5$ is shown. Note that the d corresponds to the length of the 'chromosomes'. For the details of the GA/KNN procedure, see the text and reference (Li *et al.*, 2001).

Sensitivity, reproducibility, and stability studies

Choice of d. To explore the *sensitivity* of the gene selection/sample classification to the choices of 'chromosome' size (also referred to as d, see Section k-nearest neighbors), a series of d (5, 10, 20, 30, 40 and 50) was tested on both lymphoma and colon data. For each d, 10 000 high-R^2 'chromosomes' were obtained. The patterns of gene selection and classification reliability were analyzed.

Independent re-runs of the GA/KNN method on the same data. We repeated the above gene selection procedure using different random seed numbers. For each d, another 10 000 subsets of high-R^2 chromosomes were obtained. The patterns of gene selection from two independent runs were compared to assess the *reproducibility* of the procedure.

Reassignments of 'training' and 'test' sets. In the above studies, the first N samples were chosen as the training samples ($N = 34$ of 47 for lymphoma and 40 of 57 for colon). Such an assignment is referred to as the original assignment. This strategy was modified for the lymphoma data, because a systematic difference was observed between the training and test sets based on the t-statistics. Thus, the data set was randomly shuffled

before doing the 'original' assignment. In addition, we randomly chose N samples from the whole data set as the training set (referred to as the random assignment). Finally, we chose the last N as the training samples (referred to as the discrepant assignment). This yields a worst case in that the original and discrepant assignments are maximally discrepant sets (overlap of 21 and 18 samples, for lymphoma and colon, respectively). The three 'training' sets of each data set were subject to the same procedure of gene selection. The genes selected were compared to estimate the dependence of gene selection on training set composition. The classification reliability of the selected genes on the corresponding 'test' samples was also analyzed. The parameter d was set to 40 for this stability study.

RESULTS

Sensitivity

For both lymphoma and colon data, 10 000 sets of near-optimal 'chromosomes' were obtained for each d using the training set from the original assignment (see Section Methods). Each near-optimal 'chromosome' corresponds to a set of d genes that can *jointly* discriminate between different classes of samples in the training set. Results for lymphoma and colon data were similar. For simplicity, only the results of the lymphoma data are reported here. Earlier work on the colon data has been reported (Li *et al.*, 2001).

Gene selection. The statistical z score, based on normalizing the frequency with which each of the 4026 genes was selected in the 10 000 solutions, is shown in Figure 2. It appears that a few genes dominate the selection when d is small (e.g. 5). As d increases, more peaks arise and the pattern of gene selection stabilizes. Although the patterns of gene selection were similar for various choices of d, some important differences exist. Several spikes that are not evident at $d = 5$, appear at $d = 20$ and persist for higher d. The top 50 genes selected using $d = 40$ are listed in Table 1.

The genes were ranked according to the frequency of selection with the top-most gene assigned a rank of 1. The correlation between the ranks (*log*-transformed) for two choices of d is shown in Figure 3. It appears that gene selection is insensitive to choices of d between 20 and 50.

Classification of the test set samples. For each d, we classified each of the test set samples using sets of top-ranked genes, e.g. the top 1, top 2, top 3, ..., and all 4026 genes. A sample was classified as germinal center B-like DLBCL if all of its 3-nearest *training* set neighbors were germinal center B-like DLBCL, and similarly for activated B-like DLBCL. A sample remained unclassifiable if its 3-nearest *training* set neighbors were not all of one class.

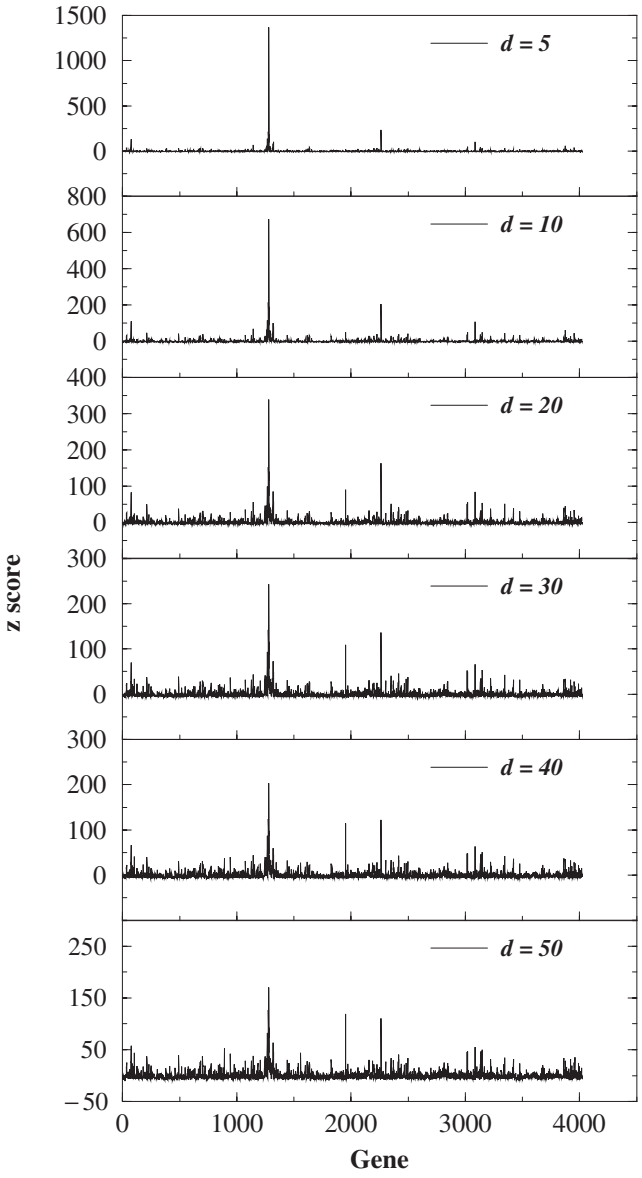

Fig. 2. The statistical z-score with which each of the 4026 genes is selected among the 10 000 solutions. Let, $Z = [S_i - E(S_i)]/\sigma$, where S_i is the number of times gene$_i$ is selected, $E(S_i)$, is the expected number of times gene$_i$ is selected, σ is the square root of the variance. Let, $A = 10\,000$, $P(\text{gene}_i) = d/4026$, the probability of gene$_i$ being selected (if random). Then, $E(S_i) = P(\text{gene}_i) \cdot A$, and $\sigma = \sqrt{\{P(\text{gene}_i) \cdot [1 - P(\text{gene}_i)] \cdot A\}}$.

The percentage of samples that were correctly classified as a function of the number of top-ranked genes used for each d for both the training and test samples is shown in Figure 4. It appears that classification is insensitive to the choices of d.

The top-most gene was able to correctly classify only

Table 1. The 50 most frequently selected genes for distinguishing germinal center B-like DLBCL and activated B-like DLBCL[a]

Student's t-statistics[b]	z-score[c]	Gene name
−9.044	203.231	Unknown UG Hs.120716 ESTs; clone = 1334 260
−7.342	124.082	Unknown UG Hs.136345 ESTs; clone = 746 300
7.675	122.267	MCL1 = myeloid cell differentiation protein; clone = 711 870
4.147	114.604	Smad4 = DPC4; clone = 774 619
−7.180	98.169	Unknown clone = 825 199
−7.468	94.136	Unknown UG Hs.169565 ESTs; clone = 825 217
−7.700	87.179	Unknown UG Hs.224323 ESTs; clone = 1338 448
−7.955	86.978	CD10 = CALLA = neprilysin = enkepalinase; clone = 200 814
−6.478	81.029	Unknown UG Hs.105261 EST; clone = 824 088
−6.994	77.399	CD10 = CALLA = neprilysin = enkepalinase; clone = 1286 850
6.442	66.611	Unknown UG Hs.169081 EST variant gene 6 TEL oncogene; clone = 1355 435
−6.635	63.788	Unknown UG Hs.140559 EST; clone = 1339 835
−6.662	60.158	Unknown clone = 1353 015
−6.516	58.545	CD10 = CALLA = neprilysin = enkepalinase; clone = 701 606
4.603	50.983	Erk3 = extracellular signal-regulated kinase 3; clone = 50 506
−5.263	48.361	PARP = poly ADP-ribose polymerase; clone = 712 849
−5.107	47.554	Unknown UG Hs.163222 ESTs; clone = 1338 044
−6.820	47.454	Unknown UG Hs.208410 EST; clone = 1353 036
−6.475	46.143	Unknown UG Hs.137038 EST; clone = 1338 981
4.766	45.840	IRAK = interleukin-1 receptor-associated kinase; clone = 345 588
−6.168	43.925	Unknown UG Hs.231798 ESTs; clone = 827 169
−6.268	43.118	Myb-related gene A = A-myb; clone = 825 476
5.677	42.715	T-cell protein-tyrosine phosphatase; clone = 665 903
−5.334	42.110	Unknown clone = 1338 436
−3.757	41.102	Unknown clone = 1354 788
4.788	40.093	Zinc finger protein 42 MZF-1; clone = 490 387
2.130	39.892	Unknown UG Hs.207506 ESTs; clone = 1338 105
−4.637	39.892	Unknown clone = 2005
−6.883	38.783	FMR2 = fragile X mental retardation 2 = putative transcription factor; clone = 1352 112
−4.148	37.371	TTG-2 = rhombotin-2; clone = 712 829
0.883	37.068	Unknown clone = 1339 086
−5.233	36.766	Unknown UG Hs.120245; clone = 1268 870
−5.047	36.564	Unknown clone = 1270 568
−5.028	35.254	Unknown UG Hs.222808 ESTs; clone = 815 273
4.835	35.052	BLC = BCA-1 = B lymphocyte chemoattractant BLC = CXC chemokine; clone = 753 794
−5.496	34.548	JAW1 = lymphoid-restricted membrane protein; clone = 417 502
5.107	33.338	T-cell protein-tyrosine phosphatase; clone = 1370 148
−6.335	33.237	Unknown clone = 1353 041
3.240	33.237	RXR-beta = retinoic acid receptor = MHC class I promoter binding protein; clone = 292 779
0.785	33.035	Unknown UG Hs.181390 casein kinase 1, gamma 2; clone = 687 112
4.073	32.632	Protein phosphatase 2A B56-beta PP2A; clone = 297 810
4.509	32.430	Microsomal glutathione S-transferase 3 MGST3; clone = 289 743
−4.690	32.027	JAW1 = lymphoid-restricted membrane protein; clone = 815 539
4.478	31.725	NERF = ets family transcription factor; clone = 768 151
3.723	31.523	Unknown UG Hs.230239 EST; clone = 1369 929
−4.668	31.321	Unknown UG Hs.27774 ESTs; clone = 1269 030
−4.651	31.019	PARP = poly ADP-ribose polymerase; clone = 200 409
4.865	30.918	Protein tyrosine phosphatase, non-receptor type 12; clone = 289 965
−4.666	30.616	BCL-6; clone = 1340 526
−5.637	30.313	Unknown UG Hs.82911 protein tyrosine phosphatase type IVA, member 2; clone = 1671 743

[a]The genes are listed in descending order on the rank frequency obtained using the 34 training set samples from the original assignment and a gene-dimension (d) of 40 (see text for details). The gene definition is adapted from http://www.llmpp.nih.gov/lymphoma. A complete list of the 4026 genes is available on http://dir.niehs.nih.gov/microarray/datamining.

[b]Student's two-sample t-test was performed on the 34 training set samples that were normalized to a mean and SD of 0 and 1, respectively. Missing values in the data were ignored. A positive value indicated that the gene is upregulated in activated B-like DLBCL compared to germinal center B-like DLBCL and conversely. The t-value at which P is 0.1, 0.01, and 0.001 are 1.684, 2.704, and 3.551, respectively.

[c]See Figure 2 legend for detail.

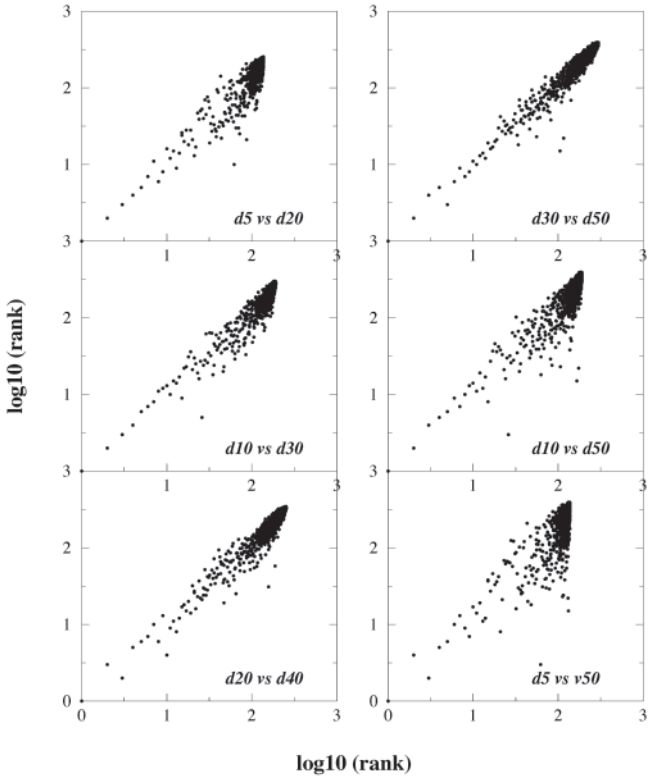

Fig. 3. The correlation between the ranks of genes ($\log_{10}$-transformed) for two choices of d. The genes were ranked according to the frequency of selection in the 10 000 solutions. The top-most gene is assigned a rank of 1 (0 after $\log_{10}$-transformation).

a few samples. The classification improved as more genes were included. At the maximum, 11 of the 13 test set samples were correctly classified for most d (Figure 4). A decrease in classification is mostly due to the increase in the number of samples that were unclassifiable using a consensus rule. When classified using instead a majority rule (2 out of 3), a wider and relatively smooth maximum emerged (not shown). For most d, the maximum corresponded to 100% success. However, the window for the maximum was relatively small. For instance, the numbers of top ranked genes that were needed to achieve the maximum were 52–55, 68–79 and 122–131 for $d = 40$. Adjacent to the window, a germinal center B-like DLBCL sample (DLCL-0020) was consistently misclassified as activated B-like DLBCL by the KNN. When the training and test sets were combined and clustered (Eisen *et al.*, 1998) using the top 50 genes ($d = 40$), all germinal center B-like DLBCL were on one branch of the tree whereas all activated B-like DLBCL were on the other except DLCL-0009 (Figure 5).

As a comparison, when the 380 genes that were originally selected using hierarchical clustering analyses (Alizadeh *et al.*, 2000) were subject to the same procedure of classification (KNN, $k = 3$, consensus rule, each training sample is classified using 3 nearest training neighbors and each test sample using 3 nearest training neighbors), 7 of the 34 training set samples and 4 of the 13 test set samples were unclassified though none was incorrectly classified. When a majority rule ($k = 3$) was used, all test set sam-

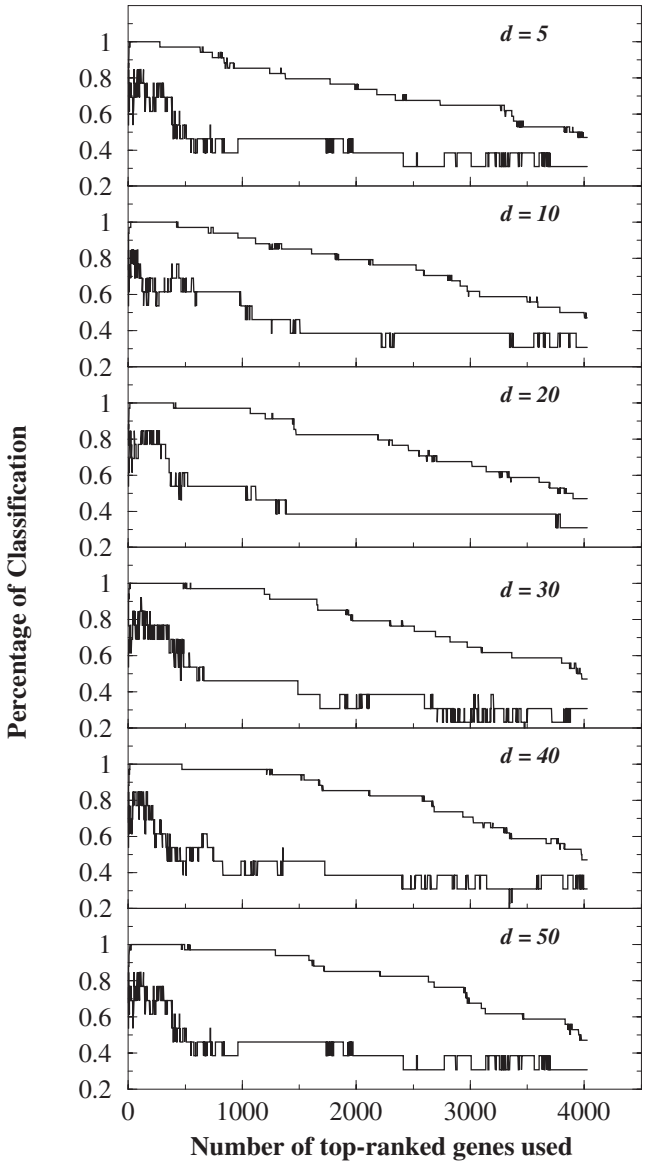

Fig. 4. The percentage of the training set (top line) and test set (bottom line) samples that were correctly classified as a function of the number of the top-ranked genes used for each d. The genes were ranked according to the frequency of selection for each particular d. Both the training and test set samples were classified using 3 nearest *training* set neighbors. All classifications were carried out using a consensus rule (3 must agree).

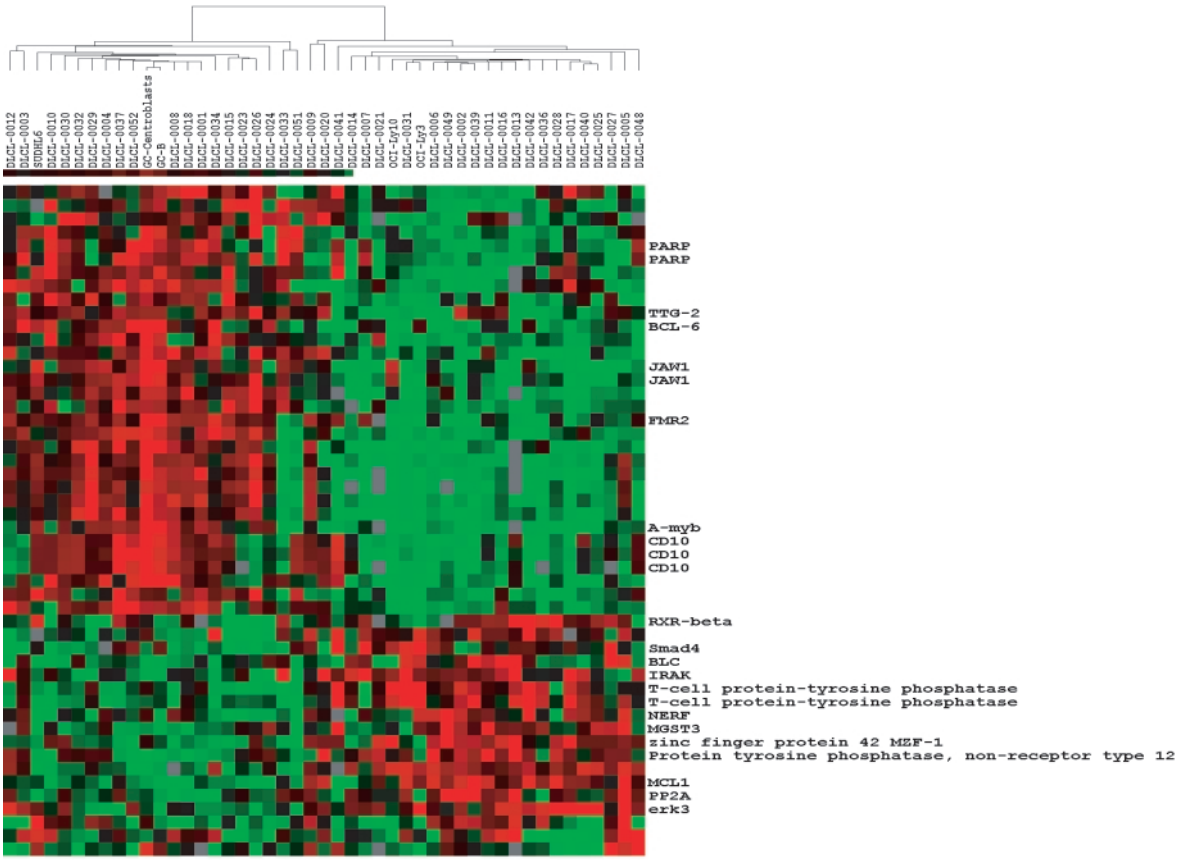

Fig. 5. Hierarchical clustering (Eisen *et al.*, 1998) of gene expression data for the lymphoma data using the top 50 genes that were selected using the GA/KNN method ($d = 40$).

ples were correctly classified while 2 training set samples (DLCL-0011 and DLCL-0024) were misclassified.

Reproducibility

To examine the reproducibility of gene selection, we repeated the same GA/KNN procedure on the same training set (from the original assignment) with different random seed numbers. Another 10 000 subsets of high-R^2 chromosomes were thus obtained through an independent run of the algorithm. The correlation between the ranks (*log*-transformed) from the two independent runs is shown in Figure 6 for each d. The reproducibility is high for all choices of d.

Stability

We chose $d = 40$ for the *stability* study, taking reassurance from the fact that the selection of optimal genes is evidently insensitive to this choice based on the above *sensitivity* and *reproducibility* studies.

Gene selection. In addition to gene selection using the training set from the original assignment, gene selection was carried out using the 'training' sets from the random and discrepant assignments, individually. The same number (10 000) of high-R^2 'chromosomes' were obtained. The correlation between the ranks of genes is shown in Figure 7. Approximately 25–37 of the top 50 genes obtained using the 'training' set from either the random or the discrepant assignment appeared in the list of top 50 genes obtained using the training set from the original assignment. Although this amount of overlap is highly significantly non-random, the selection of differentially expressed genes is more dependent on the sample used than one would like. We regard this as an inherent limitation produced by the limited number of available specimens. Other methods (e.g. those based on t-test-like measures) give similarly discouraging results. For the colon data, using maximally disjoint training sets, the two top-50 sets of genes based on the GA/KNN included only 25 that appeared in both. A comparable analysis using t-statistics to rank the genes yielded 27 that appeared in both lists (data

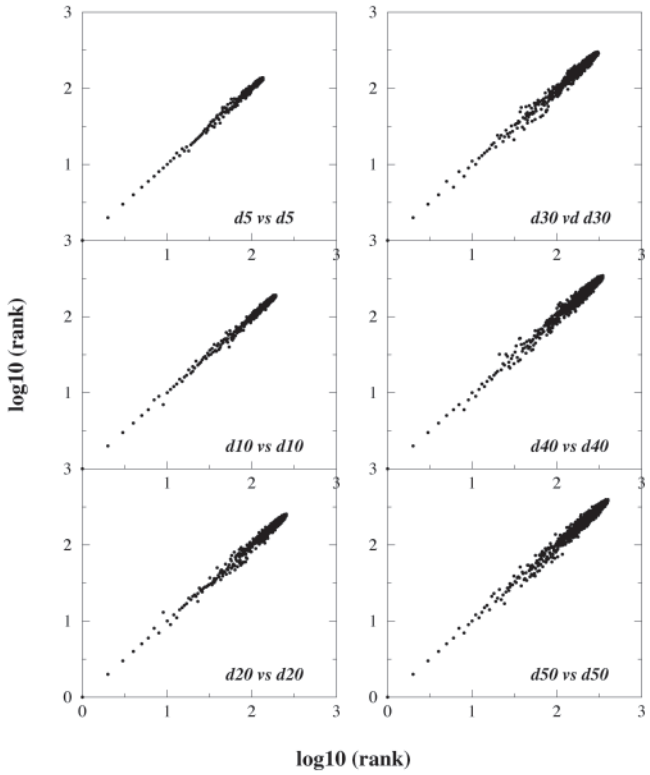

Fig. 6. The correlation between the ranks of genes ($\log_{10}$-transformed) from two independent runs for each d to assess reproducibility of the algorithm. For each run, 10 000 subsets of near-optimal (high-R^2) 'chromosomes' were obtained.

not shown). Similar degrees of disagreement were found for completely disjoint sets based on splitting of leukemia specimens (Golub *et al.*, 1999; data not shown). These results underscore the need for larger numbers of specimens to be studied in order to reliably identify differentially expressed genes.

Classification of test samples. The 50 most frequently selected genes obtained using each of the three 'training' sets of the same data were subsequently used to classify the corresponding 'test' samples. A sample was classified using 3 nearest *training* set neighbors (2 or 3 must agree). The results for the lymphoma data are shown in Table 2. For the 'test' set from the original assignment, all samples were correctly classified except DLCL-0005 and DLCL-0011. All samples in the 'test' sets from the random and discrepant assignments were correctly classified. The classification of colon data was more successful than that of the lymphoma data. All samples in all three 'test' sets were correctly classified except one (N8) ($k = 3$, majority rule; Table 3). When a consensus rule (and $k = 3$) was required, all samples in the test set from the original assignment were correctly classified (not shown). The same is true for the other two 'test' sets from the random and discrepant assignments except that one sample (N8) from the discrepant assignment became unclassifiable, perhaps due to an imbalance in normal and tumor samples in the corresponding 'training' sets (fewer normal samples). Nonetheless, in all cases, none of the 'test' samples were incorrectly classified. Similar results were found when a more stringent criterion ($k = 5$, consensus) was used (not shown).

DISCUSSION

The choice of *d*

Gene selection performed using a small d (e.g. 5) was computationally much faster than that using a large d (e.g. 50). However, a few genes dominated the selection when a small d was used. As d increased, the pattern of gene selection stablized. Although sample classification was fairly insensitive to the choices of d, a d in the range of 20–50 gave the best overall performance for all data analyzed. For other applications, similar systematic studies for the choices for d may be necessary.

Choice for the termination R^2

Preliminary studies using a variety of termination R^2 suggest that gene selection is more sensitive than test-set classification to the choice (data not shown). A slightly less stringent termination criterion (e.g. $R^2 = (M-1)/M$ or $(M-2)/M$, M = number of samples in training) had little effect on the selection of the most 'important' genes, although the relative rank order of genes will vary. A search with a less stringent termination criterion is computationally more rapid. In addition, genes of predictive importance that fail for a few samples may still be selected. Thus, a slightly less stringent termination criterion may be desirable.

The number of near-optimal 'chromosomes' needed

Since typical array data consist of a large number of genes and a small number of samples, for a given data set, many subsets of genes that by chance discriminate between different classes of sample may exist. It is important to obtain many such subsets. To determine if enough high-R^2 'chromosomes' have been obtained, one may divide the solutions into two equal-size groups and compare the ranks of gene selection. A high correlation would indicate adequate sampling.

The choice for the number of top genes for classification

With a choice of only a few top genes, the classification may not be reliable, whereas too many top genes will add noise to the classification. For the lymphoma data, the best

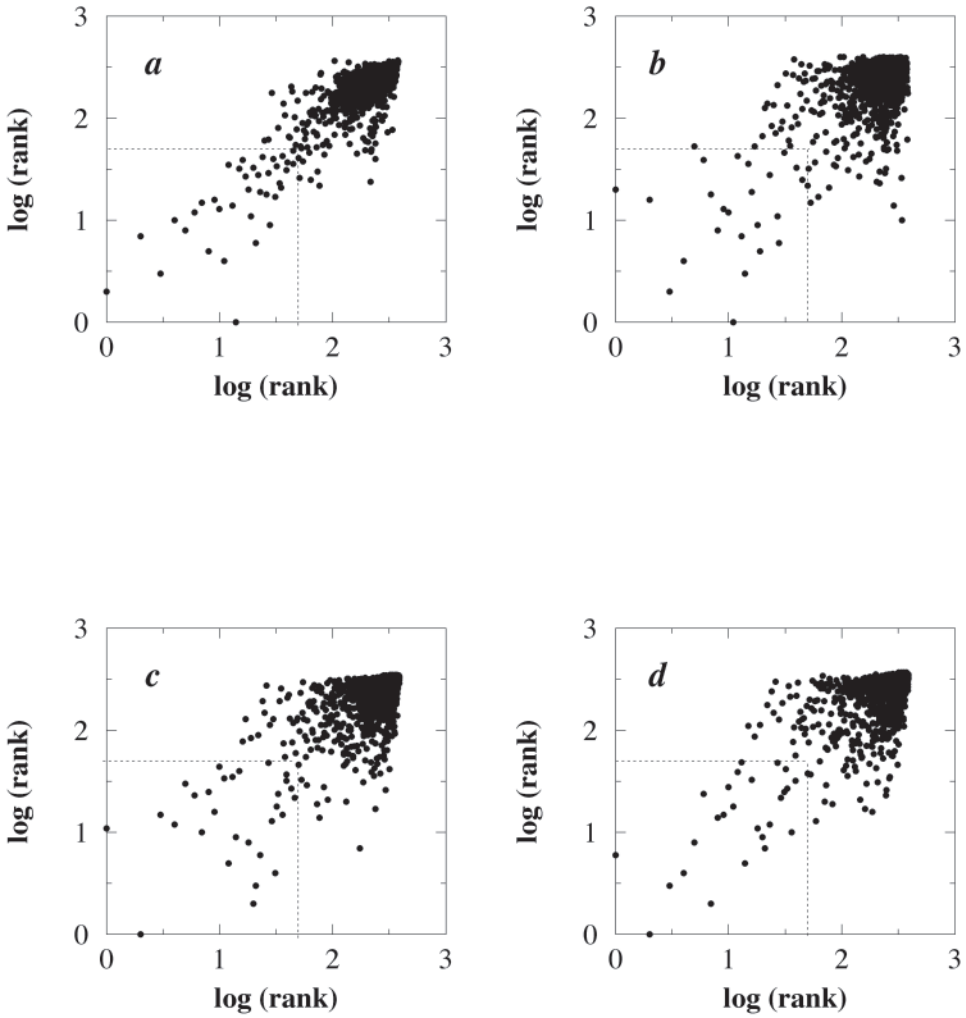

Fig. 7. The correlation between the ranks of genes ($\log_{10}$-transformed) from two 'training' sets of the same data. Each data set was divided into a training and test set in three different ways (referred to as original, random, and discrepant assignment, see Section Sensitivity, reproducibility, and stability studies in Section Methods). Each of the three 'training' sets was used to select 10 000 near-optimal 'chromosomes' that can discriminate between the different classes of samples in the 'training' set. The genes were ranked according to the frequency of selection in the 10 000 near-optimal 'chromosomes' with the top-most gene assigned a rank of 1 (0 after $\log_{10}$-transformation). (a) colon data, rank from *orginal* assignmen (x-axis) versus rank from random assignment; (b) colon data, rank from original assignment (x-axis) versus rank from discrepant assignment; (c) lymphoma data, rank from original assignment (x-axis) versus rank from random assignment; (d) lymphoma data, rank from original assignment (x-axis) versus rank from discrepant assignment. The dotted lines indicate $\log_{10}50$.

classification for the test set samples using a consensus rule corresponds to a window of approximately 50 to 200 top genes (Figure 4). Other postprocessing techniques such as linear discriminant analysis or clustering analysis could be applied using top genes selected by the GA/KNN.

Information/noise content

It is clear that not all expression data are relevant to the distinction between different classes of samples. For the lymphoma data, when the test set samples were classified using 3 nearest *training* set neighbors, all 4026 genes were only able to classify 31% of the test set samples using the consensus rule (Figure 4). When a majority rule was used, approximately 61% of the test set samples were correctly classified (not shown). These results indicate that much of the expression data does not contribute information to distinguishing between the two subtypes of lymphomas. Similar results (Li *et al.*, 2001) were obtained for colon (Alon *et al.*, 1999) and leukemia data (Golub *et al.*, 1999; data not shown).

Table 2. Classification of 'test' set samples of lymphoma data[a]

Original[b]	Class[c]	Predicted	Random[b]	Class[c]	Predicted	Discrepant[b]	Class[c]	Predicted
DLCL-0008	0	0	DLCL-0009	0	0	DLCL-0015	0	0
DLCL-0005	1	**0**	OCI-Ly10	1	1	DLCL-0009	0	0
DLCL-0006	1	1	DLCL-0041	1	1	GC-centroblasts	0	0
DLCL-0037	0	0	DLCL-0008	0	0	DLCL-0049	1	1
DLCL-0021	1	1	DLCL-0023	0	0	DLCL-0027	1	1
DLCL-0011	1	**0**	DLCL-0021	1	1	DLCL-0025	1	1
DLCL-0002	1	1	DLCL-0052	0	0	DLCL-0024	0	0
DLCL-0042	1	1	DLCL-0011	1	1	DLCL-0030	0	0
DLCL-0036	1	1	DLCL-0017	1	1	GC-B	0	0
DLCL-0013	1	1	SUDHL6	0	0	OCI-Ly10	1	1
DLCL-0010	0	0	DLCL-0016	1	1	DLCL-0029	0	0
DLCL-0051	0	0	DLCL-0028	1	1	DLCL-0048	1	1
DLCL-0012	0	0	DLCL-0039	1	1	OCI-Ly3	1	1

[a]Original, random, and discrepant are the three 'training' sets resulting from multiple splitting of the same lymphoma data (see text for details).

[b]Classification from Alizadeh *et al.* (2000).

[c]A sample is classified as 0-germinal center B-like DLBCL or 1-activated B-like DLBCL using the top 50 genes obtained for each assignment by 3-nearest *training* set neighbors using KNN with a majority rule (2 or 3 must agree). Bold type indicates incorrect classifications. See text for details.

Table 3. Classification of 'test' set samples of colon data[a]

Original[a]	Experiment[b]	Predicted[c]	Random[a]	Experiment[b]	Predicted[c]	Discrepant[a]	Experiment[b]	Predicted[c]
T28	1	1	T31	1	1	T1	1	1
N28	0	0	T28	1	1	N1	0	0
N29	0	0	T18	1	1	T2	1	1
T29	1	1	N11	0	0	N2	0	1
T31	1	1	T26	1	1	T3	1	1
T32	1	1	T4	1	1	N3	0	0
N32	0	0	T14	1	1	T4	1	1
N33	0	0	N40	0	0	N4	0	0
T34	1	1	T22	1	1	T5	1	1
T35	1	1	N1	0	0	N5	0	0
N35	0	0	N8	0	0	T6	1	1
T37	1	1	N5	0	0	N6	0	0
T38	1	1	T38	1	1	T7	1	1
T39	1	1	N29	0	0	N7	0	0
N39	0	0	N12	0	0	T8	1	1
T40	1	1	T8	1	1	N8	0	**1**
N40	0	0	T1	1	1	T9	1	1

[a]Original, random, and discrepant are the three 'training' sets resulting from multiple splitting of the same colon data (see text for details).

[b]Biological classification based on Alon *et al.* (1999).

[c]A sample is classified as 0-normal or 1-tumor using the top 50 genes obtained for each assignment by 3-nearest *training* set neighbors using KNN with a majority rule (2 or 3 must agree). Bold type indicates incorrect classifications. See text for details.

Relationship between gene selection and *t*-statistics

Again (Li *et al.*, 2001), there was no strong correlation between the frequency of gene selection and the magnitude of the gene's Student t-statistic. For instance, an unknown gene (clone 1339086) had an insignificant t-statistic (0.88) but was ranked 31 among 4026 genes. The reason for the selection of genes with small magnitude t-statistics may be the following. A gene that is not differentially expressed, and thus would appear to be non-discriminative, may be important for sample distinction when considered with other genes. For such a case, one would expect that the frequency of selection of the gene with certain others should be higher than that predicted based on their individual frequencies. In fact, the frequency of co-selection (P_{ABC}) for the unknown gene (clone 1339086, t-statistic $= 0.88$) and

two other top 50 genes (an unknown clone 1353 036, rank 18 and *BCL-6*, rank 49) was 10 times higher than would be predicted based on their individual frequencies ($P_A \times P_B \times P_C$). Other examples of co-selection exist. Thus, the GA/KNN method, by selecting sets of genes based on their joint ability to discriminate, can identify genes that are important jointly, but do not discriminate individually. Another scenario that could lead to selection of a gene with a small t-value would be when two classes of samples are compared, but disparate subclasses exist. Certain genes could be highly differentially expressed in one subclass versus another. Such genes could well be informative (as evaluated by the KNN) although the t-statistics may be small.

Comparison with other approaches

Although hierarchical clustering (Alon *et al.*, 1999; Ben-Dor *et al.*, 1999; Bittner *et al.*, 1999; Getz *et al.*, 2000a; Hartuv *et al.*, 2000) has been widely used in grouping genes/samples that have similar patterns of expression, this method alone does not fully address the problem of gene selection for sample classification. Differentially expressed genes are often selected as the informative genes, but other important genes may not be identified (see above). Even so, selecting genes that are differentially expressed from a cluster map may be difficult and time consuming, when several different classes of samples are involved and the expression is not homogeneous within a class. The number of genes selected could be large. Moreover, the relative importance of these genes in sample classification can not be fully assessed using clustering analysis. On the other hand, the GA/KNN method is able not only to select a subset of informative genes but also to assess the relative predictive importance of these genes. Furthermore, gene selection by the GA/KNN is carried out by less subjective methods using all genes.

Other computational methods that identify a subset of genes for sample classification have been reported. For instance, Golub *et al.* (1999) applied neighborhood analysis to identify a subset of genes that discriminate between the two types of leukemia, AML and ALL, using a separation measure similar to the t-statistic. Ben-Dor *et al.* (1999) applied a boosting technique (Freund and Schapire, 1997) to search for a threshold (expression level) for each gene that will maximally discriminate between two types of samples. Although differing in technical details, both approaches are univariate, that is, they examine one gene at a time, and multi-gene correlated expression patterns are not fully used. Furthermore, both approaches implicitly assume that samples show similar expression within the same class (type), i.e. they do not allow for clumping within subcategories. Unlike these two approaches, the GA/KNN is a multivariate approach for which sample heterogeneity is accommodated, so that subtypes within a class are allowed. When applied to the leukemia test set, the method of Golub *et al.* (1999) correctly classified 29/34, while the GA/KNN correctly classified 33/34. Moreover, the subset of genes found by the GA/KNN not only discriminated between AML and ALL, but also revealed the existence of two subtypes within ALL without applying prior knowledge (Li *et al.*, 2001). Detailed discussion comparing the GA/KNN with other approaches is given elsewhere (Li *et al.*, 2001).

In conclusion, the GA/KNN method is a multivariate approach in which the joint discriminative ability of several genes is analyzed. Many subsets of discriminative genes are obtained, after which a single predictor set is formed by examining the frequency of gene selection. The predictor can be used to classify unknown samples. The GA/KNN method accommodates heterogeneity within the classes, which facilitates subclass discovery. The method could potentially be useful in uncovering a group of genes that serve to fingerprint subtypes of the disease and the selected genes could thus aid in refining cancer diagnosis, improving assessment of prognosis, and suggesting carcinogenic mechanisms.

ACKNOWLEDGEMENTS

The work was partially completed when L.L. was employed at NIOSH, West Virginia. We thank an anonymous reviewer for his/her careful review of the manuscript. We also thank Drs David Umbach and Shyamal Peddada (NIEHS) and Sidney Soderholm, Doug Landsittel, and Gene Demchuk (NIOSH/CDC) for careful reading of the manuscript and insightful discussion. L.G.P. acknowledges a grant of computer time from the North Carolina Supercomputing Center and support from NIH (HL-06350). Most of the computations were carried out on SGI workstations.

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., Powell,J.I., Yang,L., Marti,G.E., Moore,T., Hudson,J. Jr, Lu,L., Lewis,D.B., Tibshirani,R., Sherlock,G., Chan,W.C., Greiner,T.C., Weisenburger,D.D., Armitage,J.O., Warnke,R., Levy,R., Wilson,E., Grever,M.R., Byrd,J.C., Botstein,D., Brown,P.O. and Staudt,L.M. (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. *Proc. Natl Acad. Sci. USA*, **96**, 6745–6750.

Ben-Dor,A., Shamir,R. and Yakhini,Z. (1999) Clustering gene expression patterns. *J. Comput. Biol.*, **6**, 281–297.

Ben-Dor,A., Bruhn,L., Friedman,N., Nachman,I., Schummer,M., Yakhini,Z. and Ben-Dor,A. (2000) Tissue classification with gene expression profiles. In *Proceedings of the Fourth Inter-*

national Conference on Computational Molecular Biology (RECOMB2000). ACM Press, New York.

Bittner,M., Meltzer,P. and Trent,J. (1999) Data analysis and integration: of steps and arrows. *Nature Genet.*, **22**, 213–215.

Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14 863–14 868.

Freund,Y. and Schapire,R.E. (1997) A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, **55**, 119–139.

Getz,G., Levine,E., Domany,E. and Zhang,M.Q. (2000a) Superparamagnetic clustering of yeast gene expression profiles. *Physica* A, **279**, 457–464.

Getz,G., Levine,E. and Domany,E. (2000b) Coupled two-way clustering analysis of gene microarray data. *Proc. Natl Acad. Sci. USA*, **97**, 12 079–12 084.

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

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., Bloomfield,C.D. and Lander,E.S. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Hartuv,E., Schmitt,A.O., Lange,J., Meier-Ewert,S., Lehrach,H. and Shamir,R. (2000) An algorithm for clustering cDNA fingerprints. *Genomics*, **66**, 249–256.

Holland,J.H. (1975) *Adaptation in Natural and Artificial Systems*. The University of Michigan Press, Ann Arbor, IL.

Judson,R. (1997) Genetic algorithms and their use in chemistry. In Lipkowitz,K.B. and Boyd,D.B. (eds), *Reviews in Computational Chemistry*, Vol. 10, VCH, New York, pp. 1–66.

Li,L., Darden,T.A., Weinberg,C.R. and Pedersen,L.G. (2001) Gene assessment and sample classification for gene expression data using a genetic algorithm/k-nearest neighbor method. *Combinatorial Chemistry & High Throughput Screening*, **4**, in press.

Massart,D.L., Vandeginste,B.G.M., Deming,S.N., Michotte,Y. and Kaufman,L. (1988) The k-nearest neighbor method. In *Chemometrics: A Textbook (Data Handling in Science and Technology, Vol. 2)*. Elsevier Science, New York, pp. 395–397.