Gene mining: a novel and powerful ensemble decision approach to hunting for disease genes using microarray expression profiling

Xia Li^{1,2}, Shaoqi Rao^{3,*}, Yadong Wang² and Binsheng Gong¹

¹Department of Biomedical Engineering, Biomathematics and Bioinformatics, Harbin Medical University, Harbin 150086, China, ²Department of Computer Science, Harbin Institute of Technology, Harbin 150001, China and ³Departments of Molecular Cardiology and Cardiovascular Medicine, the Cleveland Clinic Foundation, 9500 Euclid Avenue, Cleveland, OH 44195, USA

Received December 12, 2003; Revised February 29, 2004; Accepted April 2, 2004

ABSTRACT

Current applications of microarrays focus on precise classification or discovery of biological types, for example tumor versus normal phenotypes in cancer research. Several challenging scientific tasks in the post-genomic epoch, like hunting for the genes underlying complex diseases from genome-wide gene expression profiles and thereby building the corresponding gene networks, are largely overlooked because of the lack of an efficient analysis approach. We have thus developed an innovative ensemble decision approach, which can efficiently perform multiple gene mining tasks. An application of this approach to analyze two publicly available data sets (colon data and leukemia data) identified 20 highly significant colon cancer genes and 23 highly significant molecular signatures for refining the acute leukemia phenotype, most of which have been verified either by biological experiments or by alternative analysis approaches. Furthermore, the globally optimal gene subsets identified by the novel approach have so far achieved the highest accuracy for classification of colon cancer tissue types. Establishment of this analysis strategy has offered the promise of advancing microarray technology as a means of deciphering the involved genetic complexities of complex diseases.

INTRODUCTION

The advent of DNA microarray technology has offered the promise of obtaining new insights into the secrets of life by monitoring the activities of thousands of genes simultaneously. These measurements or gene profiles provide a 'snapshot' of life that maps to a cross-section of genetic activities in a four-dimensional space of time and the biological entity. Although recent microarray experiments

(1,2) hold the promise of the innovative technology being used to classify biological types, development of powerful and efficient analysis strategies to perform more complex biological tasks, such as mining disease relevant genes and building genetic networks, remains a significant demand. According to the modes of learning algorithms, current analysis strategies can be classified into two groups. An unsupervised learning method such as a typical cluster analysis, by ignoring the biological attribute (label) of a DNA example (instance), directly works on genes themselves and is a useful tool to study functional genomics, but it is unable to efficiently relate differential gene expression profiles to phenotypes. On the other hand, supervised learning is a target-driven process in that a suitable induction algorithm is employed to identify the most contributed genes for a specific target, for example classification of biological types, gene mining or data-driven gene networking.

DNA microarrays have several unique characteristics, such as relatively few samples with a large dimension of feature gene space and a high signal-to-noise ratio. Therefore, attempts to analyze entire microarrays at once to immediately discover the pattern being sought would be a mistake. Microarrays provide massively parallel information, the analysis of which is a non-deterministic polynomial (NP)-hard problem (3). Current statistical methods are not sufficiently powerful to solve this NP-hard problem. Hence, several methods having flavors of modern data mining and machine learning have been proposed to reduce the feature gene space by extracting the contributed genes and excluding the noise features (4–9). Among these methods, decision-tree based approaches appear to be one of the best choices for both genetic data analysis and for more general settings (10,11), as it can partition the sample and feature gene space simultaneously and thus is especially efficient for heterogeneous microarray samples.

In this study, we have developed a novel tree-based ensemble method for analysis of microarray data or data with similar characteristics. First, we aimed at extracting an optimal gene subset(s) to achieve a high precision of classification (e.g. tumor versus normal tissue), the so-called prediction problem. One group of current methods marginally

*To whom correspondence should be addressed. Tel: +1 216 444 0056; Fax: +1 216 444 2682; Email: raos@ccf.org
Correspondence may also be addressed to Xia Li. Tel: +86 451 8661 5922; Fax: +86 451 8666 9617; Email: lixia@ems.hrbmu.edu.cn

filter noise genes based on rank, information gain and Markov blanket (12–14). The feature gene selection is independent of the resulting classifier(s) so that it is uncertain whether the selected subset of genes can lead to maximal prediction precision. Another group of methods for feature gene selection are wrappers and a hybrid of filtering and wrapping (15,16). In a wrapping approach, the algorithm for feature gene subset selection exists as a wrapper around the 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 training and external test sets. The feature subset with the highest evaluation is chosen as the final set on which to build a classifier. Because 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. We thus incorporated a decision tree-based wrapper in the novel ensemble approach. Several authors have 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) can be a third method to perform feature subset selection, known as ‘embedded algorithms’ (14). However, wrappers and embedded algorithms are often not clearly distinguished, with only slight differences in the feature searching strategies.

Second, we propose an innovative approach for mining complex disease genes. Instead of simply maximizing prediction accuracy, we identify genes that are mostly relevant to a disease itself. One might think that the two targets are equivalent, but we will show later by a numerical example that they are essentially different. In other words, the optimal gene subset for prediction is too simple to reflect the genetic complexity of complex diseases. For this purpose, we introduce a disease-relevance concept and define a relevance intensity (precise mathematical descriptions will be given later) to distinguish between disease-relevant genes and noise features. Relevance at large has been studied extensively over the last three decades and there is increasing interest in applications in a wide range of areas, in particular in the area of machine learning for feature subset selection, possibly owing to the advent of computational ability to handle massive high-dimensional data sets. A nice review and study of relevance concepts by Bell and Wang (17) reveals that either definitions or interpretations of relevance have evolved considerably, from a simple and intuitive relevance concept for marginally filtering a feature to the sophisticated mathematical formalism of the concept that is quantitative and normalized. In this study, we adopt and quantify with a relevance intensity the relevance formalism of Kohavi and John (15) (Definition 5, strong relevance; Definition 6, weak relevance), because this formalism is easily generalized to our specific application to microarrays and is appealing to capture the reality of biological complexities (epistasis or gene–gene interactions). The dual purposes of introducing the relevance concept were: (i) to characterize the target-dependent behavior and properties of a feature gene that distinguish it from the correlation metric for unsupervised learning on microarrays; (ii) to define a measure for mining disease-relevant genes.

The conventional subset selection approaches that aim at achieving a maximal accuracy of prediction would lead to

exclusion of a large number of partially relevant genes due to ‘redundancy’. To overcome this problem, we propose a novel ensemble approach to handling ‘redundant’ genes. We integrate an ensemble decision theory in supervised learning fields (18) into decision trees where multiple gene subsets are obtained from the training sets generated by a resampling technique and each putative feature gene is then evaluated in terms of a relevance intensity. An application of this approach to analyze two publicly available data sets [colon data (19) and leukemia data (2)] is given to demonstrate its performance and statistical properties for either mining relevant features based on their distribution over the gene forest using an ensemble voting approach or prediction evaluated using multiple external classifiers.

ANALYTICAL METHODS

Definitions

Gene expression data with p probes and n DNA samples can be described by a $n \times p$ matrix, $X = (x_{ij})$, where x_{ij} represents the expression level for the j th gene (g_j) on the i th sample (X_i). When the phenotype of a DNA sample is known, the data for each sample consist of a vector of expression profile, $X_i = (x_{i1}, \dots, x_{ip})$ and a category label (y_i). Suppose that the DNA samples belong to K categories, $\omega_1, \omega_2, \dots, \omega_K$. Define a class label, y_i , to be an integer from 1 to K and let n_k be the number of samples in the k th category.

Definition 1. Given an inducer I and the microarray data D with p genes, $\{g_j\}$, $j = 1, 2, \dots, p$, from a multivariate distribution over the discrete phenotypic space $\{\omega_1, \omega_2, \dots, \omega_K\}$, an optimal subset of genes with the minimal misclassification rate, G' , is said to be a subset of genes such that based on it the partition of microarray samples by the induced classifier $C = I(D)$ leads to the best fit (with the highest goodness-of-fit statistic) to the observed phenotypic distribution.

Definition 2. A feature gene g_j is said to be completely relevant if it is included in all the classifiers and the removal of g_j alone will result in performance deterioration (in term of misclassification rate) of all the classifiers, obtained by learning on a number of training sets generated by a proper resampling technique. A feature gene g_j is partially relevant if it is not completely relevant and there exists a subset of feature genes, G , such that the performance of the induced classifier on G is worse than the performance on the union of G and g_j . A feature gene is irrelevant if it is neither completely nor partially relevant.

Ensemble feature selection

The proposed ensemble selection is a supervised learning approach based on a recursive partition tree. The basic procedures are as follows. First, a resampling technique is employed to build up pairs of training sets, $\{L_d\}$ ($d = 1, 2, \dots, m$), and test sets, $\{T_d\}$ ($d = 1, 2, \dots, m$), for learning and testing, respectively. Then, a binary tree is grown on L_d by a recursive partition algorithm. At each branching of the tree, the best feature gene is selected such that it leads to minimal impurity at the node. This binary recursive partition continues

until tree growth is stopped. For each tree grown, one subset of genes is extracted and tested on the holdout set T_d . This process for feature selection is repeated on each pair of $\{L_d\}$ ($d = 1, 2, \dots, m$) and $\{T_d\}$ ($d = 1, 2, \dots, m$), which consequently results in an ensemble of feature gene subsets (gene forests), $G_1, \dots, G_d, \dots, G_m$, $G_d = \{g_1^d, g_2^d, \dots, g_k^d\}$. The optimal gene subset for the purpose of classification that meets Definition 1, G' , is obtained by ranking on the subset ensemble $\{G_d\}$ based on their respective classification performance on the holdout test samples. By Definition 2, the disease-relevant genes, G^* , are mined by a vote for a weighted frequency that a particular gene is contained in the gene forests.

Construction of training and test data. Among numerous methods, we here introduce three common methods for constructing training sets and test sets. Bagging (10), abbreviated from Bootstrap Aggregation, is the most direct method, in which a training set of size N is randomly selected from microarrays and is used as a mother template and the rest of the microarrays are used as the test set. Each microarray may appear several times or not at all in any particular training set. In an n -fold cross-validation procedure (10), we randomly divide the data into n equal parts. One part of the data is used as the test set and the rest is used as the training set. We continue in this way through the entire data set to generate n combinations of non-overlapped training data sets and test data sets. The third method is to randomly select without replacement $1/n$ of data as the test set and the remaining as the training set.

Induction algorithm. For each resampling constructed pair, L_d and T_d , a recursive partition tree (20,21) is grown on the training set (L_d). The search for feature genes starts at the root of the tree and proceeds to its leaves. At each internal node, a decision is made with regard to the choice of a feature gene and a threshold value (cut-off) such that the class impurity is reduced to a minimum when a branch is made by an induction rule. After the optimal bifurcation is made, the microarray samples are divided into two non-overlapping subsets (two child nodes), dependent on whether for sample X_i the selected feature gene $g_j > \text{cut-off}$ or $g_j \leq \text{cut-off}$. The best cut-off for bifurcation is determined by iteratively searching the mid-points between two-ordered values of g_j considered to split the examples in the tree node. The same process is conducted successively until a leaf is reached or stopping criteria for tree growth are satisfied. The basic rule for selecting a feature gene and cut-off function is that such an induction rule leads to a maximal drop in category impurity at that particular layer of tree, i.e. a partition is sought to maximally reduce an impurity index (Gini inequality index) at node t :

$$E(t) = \phi[P(\omega_1 | t), \dots, P(\omega_K | t)]$$

$$= \sum_{k \neq m} P(\omega_k | t) P(\omega_m | t) = 1 - \sum_{k=1}^K P^2(\omega_k | t) \quad 1$$

Often, $P(\omega_k | t) = p_k = n_k/n$ ($k = 1, 2, \dots, K$). p_k is the probability that a sample belongs to the k th category at node t and

$$\sum_{k=1}^K p_k = 1.$$

Feature gene selection proceeds in a manner that the best gene is sought so that impurity is minimized at node t , from which a new bifurcation is attempted. Say, attempts are made to search for the best feature gene as well as its cut-off, which results in a new partition, s^* , of the data and also leads to a maximal decrease in impurity at this bifurcation. Note mathematically,

$$\Delta E(s^*, t) = \max_{s \in S} \Delta E(s, t) = \max_{s \in S} [E(t) - p_L E(t_L) - p_R E(t_R)] \quad 2$$

where $E(t_L)$ and $E(t_R)$ are impurity functions for the bifurcated nodes, respectively, and p_L and p_R are the frequencies of the two branching events at node t . S is a set consisting of all the possible bifurcation events at node t . The optimal branching, s^* , not only produces two child nodes (t' and t'') from the parent node t , but also extracts a feature gene at that node. The process for feature gene extraction is repeated for t' and t'' , respectively, until tree growth is stopped and a subset of feature genes, G_d , is obtained from the particular training set, L_d .

Evaluation of a tree (a subset of feature genes). To test whether the extracted subset of genes (G_d) is able to significantly distinguish the phenotypes of the holdout microarray samples (T_d), we propose a χ^2 statistic:

$$\chi^2 = [(n_{00}n_{11} - n_{01}n_{10})^2 / n] / [(n_{00} + n_{01})(n_{10} + n_{11})(n_{00} + n_{10})(n_{01} + n_{11})] \quad 3$$

where $n = n_{00} + n_{01} + n_{10} + n_{11}$ and n_{00} , n_{01} , n_{10} and n_{11} are the frequencies for true negative, false positive, false negative and true positive, respectively. This statistic follows an asymptotic χ^2 distribution with 1 degree of freedom.

Selection of relevant genes based on ensemble decision. The above feature gene subset building algorithm is applied to the multiple training sets and test sets, generated by a particular resampling technique as described previously. Consequently, a set of feature subsets are obtained by inducing the decision trees, denoted $\{G_1, G_2, \dots, G_m\}$. Given the ensemble of feature gene subsets, strategies for extracting the optimal feature genes diverge for different targets. For the purpose of prediction or classification, a feature gene subset as a whole instead of individual genes is considered. The optimal gene subset is selected based on the performance of a particular subset of genes extracted from a training set on classification of its sister test set. Since the typical goal of disease gene mining is to find a relevant gene instead of simply maximizing classification accuracy on a test set, we have adopted this as our target in guiding our feature gene selection. In other words, the genes from the prediction-driven extraction of relevant genes are important for prediction but may not be so for deciphering the complex underlying genetic architecture of the disease itself. Furthermore, such a strategy for feature gene selection would fail to pick up partially relevant genes because many genes (e.g. a functional cluster) on microarrays are highly correlated and not necessarily included in an optimal subset to achieve maximal accuracy of prediction. Here, we provide an innovative approach to extract disease-relevant genes based on the

Table 1. The step-by-step recipe for the computational algorithm of the ensemble approach

Step 1. Construct structure of multiple training and test sets, $\{L_d\}$ and $\{T_d\}$ by resampling microarray instances.
Step 2. Set up the stop conditions for tree growth: having observations from only one class or having the maximum allowed instances in the node.
Step 3. Grow a tree on the particular training set, L_d . Select a feature gene at root node ($t = 0$) and determine a cut-off value to make an optimal partition (branching). At node t ($t = 1, \dots$), compute the prior probability, $P(\omega_j t) = n_j/n$ and then $E(t)$ from equation 1. Rank in descending order the expression level g_{ij} ($i = 1, 2, \dots, n$) for each feature gene, g_j ($j = 1, 2, \dots, p$). For the feature gene g_j , its corresponding partition is determined by the midpoints of two-ordered values of g_j : $s_{ij} = g_{ij} + (g_{ij} - g_{i+1,j})/2$. Compute all the $\Delta E(s, t)$ and record the corresponding s^* and g_j^d that satisfy equation 2. Use s^* to bifurcate the node t into two child nodes, t_1 and t_2 . Repeat the iterative operations (1–5) on nodes t_1 and t_2 until a stopping rule for tree growth is satisfied, i.e. a terminal node contains observations from only one class or has the maximum allowed instances in the node. The feature gene subset consists of the genes at the internal nodes and is denoted as: $G_d = \{g_1^d, g_2^d, \dots, g_d^d\}$.
Step 4. Evaluate the selected gene subset: compute the χ^2 value for G_d (equation 3) as well as other three indices, concordance rate, $Acc(G_d)$, prediction accuracy for true positives, $pP(G_d)$ and discriminant accuracy for true positives, $dP(G_d)$, respectively, where $Acc = (n_{00} + n_{11})/n$, $pP = n_{11}/(n_{01} + n_{11})$ and $dP = n_{11}/(n_{10} + n_{11})$. The notations for n , n_{00} , n_{01} , n_{10} and n_{11} are as given previously in the main text.
Step 5. Extract the feature gene subset, $G_d = \{g_1^d, g_2^d, \dots, g_d^d\}$.
Step 6. Repeat steps 3–5 and construct the set of feature gene subsets (gene forests), $\{G_d\}$.
Step 7. For each of feature genes g_k compute $FV(g_k)$ using equation 4.
Step 8. Randomly assign a label to each DNA sample. Then use the shuffled data to conduct steps 1–7. Compute the relevance intensity for each gene. Repeat the permutation N times to obtain an empirical threshold for FV_0^g at the specified significance level of β .
Step 9. Output the significantly relevant genes and their corresponding relevance intensities.

ensemble of feature gene subsets. Whether a feature gene is relevant to a disease depends on the magnitude of its relevance intensity (called an ensemble vote), FV .

For a particular feature gene g_k , define:

$$FV(g_k) = F(G_1, G_2, \dots, G_m) = [\sum_d w_d I(g_k, G_d)] / [\sum_d w_d] \quad 4$$

where $FV \in [0, 1]$ and $I(g_k, G_d)$ is an indicator function:

$$I(g_k, G_d) = \begin{cases} 1 & g_k \in G_d \\ 0 & \text{otherwise} \end{cases}$$

A weight, w_d , can be a measure for the classification performance of G_d , for example, $w_d = \chi_d^2$ or set $w_d = 1$ for an equal weight for all the feature gene subsets.

Because the distribution of FV is often unknown, we resort to a permutation approach in which we randomly assign a label (phenotype) to each microarray and $FV(g_k)$ is computed using the permuted data. Then, the empirical null distribution of $FV^0(g_k)$ is constructed. Given the empirical $FV^0(g_k)$ and a specified significance level, β (e.g. 0.05 or 0.01), a critical value FV_β^0 is obtained. A feature gene is selected if $FV \geq FV_\beta^0$ (one-tailed).

Computational algorithms

The numerical algorithm for the proposed ensemble approach, organized step by step, is given in Table 1. The computational algorithm has been realized on the MATLAB platform. The corresponding programming codes are available upon written request to the authors.

NUMERICAL APPLICATIONS

We report the results from analysis of two well-known data sets in the microarray literature: colon data, analyzed initially by Alon *et al.* (19), available at <http://microarray.princeton.edu/oncology/affydata/index.html>; leukemia data, analyzed initially by Golub *et al.* (2), available at <http://www.broad.mit.edu/cancer/>. The colon data set consists of absolute measurements from Affymetrix oligonucleotide arrays, with

62 tissue samples of 2000 human gene expressions (40 tumors and 22 normal tissues). The leukemia data set consists of 7070 preprocessed gene expressions in tissues derived from bone marrow aspirates (or peripheral blood) of 72 acute leukemia patients [25 acute myeloid leukemia (AML) samples and 47 acute lymphoblastic leukemia (ALL) samples]. First, we use the two data sets to demonstrate the utility of the proposed ensemble approach to extracting colon tumor-relevant genes or molecular signatures for leukemia subtypes. Then, using an independent external cross-validation procedure, we verify the classification performance of the three best subsets (trees) identified by the proposed ensemble approach, with comparisons with the 2000 genes or the 20 ensemble-identified relevant genes or 20 randomly selected genes as predictors, evaluated using multiple external classifiers including a support vector machine (SVM) with five different kernel functions, Fisher linear discriminator (FLD), logistic non-linear (logit) regression (LNR), K-nearest neighbors (KNN) and Mahalanobis distance (MD). In the analyses of the two data sets, the maximum allowed instances in a node is set to one to minimize the risk of losing any important feature gene because of the limited sample sizes. On the other hand, some trivial genes may be included during feature subset selection, but they can be efficiently removed from the following ensemble decision analysis. We do not perform either pre-pruning or post-pruning because for the above reason.

Gene mining: disease-relevant genes

Example 1. Mining colon tumor-relevant genes. A 5-fold cross-validation (CV) resampling approach is used to construct the training and test sets. First, colon tumor and normal samples are randomly divided into five non-overlapping subsets of roughly equal size, i.e. tumor subsets D_i ($i = 1, 2, \dots, 5$) and normal subsets N_i ($i = 1, 2, \dots, 5$). A random combination of D_i and N_i constitutes a test set and the rest of the subsets are used as the training set. The 5-fold CV resampling produces 25 pairs of training and test sets. It is noteworthy that this cross-validation process is 'stratified' with respect to the class variable of the task, with an implicit purpose to identify multiple underlying genetic paths for the

Table 2. Ensemble decision analysis of 2000 gene expressions identifies 20 highly significant (empirical $P < 0.01$) colon tumor relevant genes

Gene no.	Relevance	Sequence	Name
M26383	0.462	Gene	Human monocyte-derived neutrophil-activating protein (MONAP) mRNA, complete cds
M63391	0.272	Gene	Human desmin gene, complete cds
T58861	0.193	3' UTR	60S ribosomal protein L30E (<i>K.lactis</i>)
R39465	0.190	3' UTR	Eukaryotic initiation factor 4A (<i>O.cuniculus</i>)
R87126	0.190	3' UTR	Myosin heavy chain, nonmuscle (<i>Gallus gallus</i>)
H55933	0.146	3' UTR	<i>Homo sapiens</i> mRNA for homolog to yeast ribosomal protein L41
D14812	0.132	Gene	<i>Homo sapiens</i> KIAA0026 mRNA, complete cds
H55758	0.118	3' UTR	α -Enolase (Human)
U14973	0.108	Gene	Human ribosomal protein S29 mRNA, complete cds
R39465 ^a	0.098	3' UTR	Eukaryotic initiation factor 4A (<i>O.cuniculus</i>)
T62947	0.098	3' UTR	60S ribosomal protein L24 (<i>Arabidopsis thaliana</i>)
T51849	0.091	3' UTR	Tyrosine-protein kinase receptor ELK precursor (<i>R.norvegicus</i>)
Z24727	0.060	Gene	<i>Homo sapiens</i> tropomyosin isoform mRNA, complete cds
T65938	0.052	3' UTR	Translationally controlled tumor protein (human)
H87465	0.051	3' UTR	Pre-mRNA splicing factor SRP75 (<i>H.sapiens</i>)
M69135	0.050	Gene	Human monoamine oxidase B (MAOB) gene, exon 15
H08393	0.049	3' UTR	Collagen $\alpha 2(XI)$ chain (<i>H.sapiens</i>)
R54097	0.044	3' UTR	Translational initiation factor 2 β subunit (human)
T72863	0.038	3' UTR	Ferritin light chain (human)
J02854	0.036	Gene	Myosin regulatory light chain 2, smooth muscle isoform (human);contains element TAR1 repetitive element

^aTwo probes that correspond to the same gene (R39465) were used in the microarray experiment.

target phenotype. We repeat the resampling 20 times and obtain 500 pairs of training and test sets. The proposed gene extraction approach is then applied to each pair. In order to obtain a statistical measure of significance for each gene, a null distribution FV^0 is constructed, as described previously. An empirical threshold of 0.035 at the significance level of 0.01 is chosen, denoted as $FV^0_{\beta} = 0.035$ ($\beta = 0.01$). The extracted colon tumor-relevant genes of high significance ($P < 0.01$), obtained by analyzing 500 pairs, are given in Table 2. Pairwise relationships between the relevant genes in terms of Pearson's correlations of absolute gene expressions (above the diagonal) or pairwise joint relevance (below the diagonal), evaluated by the concordance rate of two genes appearing in the same trees among the gene forests of 1000 artificially grown trees, are shown in Table 3.

In total we identified 20 highly significant (empirical $P < 0.01$) relevant genes. The top one was M26383, human monocyte-derived neutrophil-activating protein (MONAP) mRNA, complete cds, which appears in 230 or ~46% of trees. Given so strong a relevance intensity, it can be postulated that this gene plays a pivotal role as a central hub for the gene network that maps to the underlying pathological complexity of colon cancer. Molecular experiments evidence indicated that MONAP was constitutively overexpressed by human tumor lines (22). Strikingly, MONAP was ranked as the top colon tumor gene using the versatile measures information gain, sum of variances, twoing rule and Gini index, and as the second most important gene ranked by the summary minority measure (<http://genomics10.bu.edu/yangsu/rankgene/compare-alon-colon-cancer-top100.html>), but surprisingly not among the top 100 genes identified by Student's t -test, implemented using the RankGene package (23). The second most significant gene identified by ensemble decision analysis in this study is human desmin, complete cds, which appears in ~27% of trees in the gene forests. It is not at all surprising that this gene was also identified by RankGene to be either the top one using the measures sum minority, max minority and one-dimensional SVM or as the second using the

measures information gain, sum of variances, twoing rule and Gini index. T51849, tyrosine-protein kinase receptor ELK precursor (*Rattus norvegicus*), ranked 12th based on its relevance intensity to colon tumor phenotype, is also worth further investigation. A large microarray experiment with leukemic blasts from 360 acute pediatric lymphoblastic leukemia patients identified expression of this gene as highly associated with a leukemia subtype, *E2A-PBX1* (24), suggesting that this gene might have a pleiotropic action in multiple cancers. Application to this colon tumor data supports our speculation that the ensemble decision approach is efficient for extracting 'redundant' genes. An extreme example is R39465, which was repeated twice in the microarray experiment. We have successfully identified both replicates, although there are some variations in relevance ranks (ranked 4th and 10th, respectively).

To investigate whether joint relevance shares some characteristics with Pearson's correlation of gene expressions, which serves as the basis for unsupervised functional clustering, we grew 1000 trees and then computed the two indices using the top 20 tumor-relevant genes identified from ensemble decision analysis of the colon microarray data. Pearson's correlation of gene expressions for the 20 genes is a target-independent index, while the joint relevance is a target-driven relationship index for genes, which might be a useful candidate for reverse engineering of the underlying disease gene networks. The joint relevance is computed as the concordance rate of two genes appearing in a same tree. Of 190 Pearson's correlations, 45 are negative and the rest are positive. According to their absolute magnitude, 46 are strongly correlated (correlation $\rho \geq 0.5$), with the two replicates of R39465 being the strongest (0.974), 76 are correlated at the medium level ($0.2 \leq \rho < 0.5$) and the rest (68 estimates) are weakly correlated ($\rho < 0.2$). The averaged Pearson correlation over 190 estimates was 0.326 (± 0.232). In contrast, the joint relevance matrix is highly sparse, with 49 non-zeros (mean $\pm$ SD, 0.028 ± 0.030). Strongly relevant genes tend to have strong joint relevance with other genes, in

Table 3. Pairwise joint relevance (the below diagonal) and Pearson's correlation (the above diagonal) matrix for the top 20 colon tumor relevant genes, identified by ensemble decision analysis of 2000 gene expression profiles

Gene	M26383	M63391	T58861	R39465	R87126	H55933	D14812	H55758	U14973	R39465 ^a	T62947	T51849	Z24727	T65938	H87465	M69135	H08393	R54097	T72863	J02854
M26383																				
M63391	0.034																			
T58861	0.049	0.095																		
R39465	0.139	0.042	0.031																	
R87126	0.000	0.000	0.000	0.000																
H55933	0.077	0.038	0.020	0.018	0.039															
D14812	0.000	0.000	0.000	0.000	0.000	0.000														
H55758	0.075	0.008	0.000	0.000	0.021	0.000	0.000													
U14973	0.026	0.005	0.013	0.000	0.032	0.000	0.000	0.004												
R39465 ^a	0.000	0.078	0.040	0.000	0.002	0.000	0.058	0.002	0.000											
T62947	0.017	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000										
T51849	0.097	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000									
Z24727	0.052	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.009	0.014								
T65938	0.001	0.046	0.023	0.000	0.004	0.000	0.025	0.000	0.000	0.000	0.001	0.000	0.000							
H87465	0.025	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000						
M69135	0.043	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.010					
H08393	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000				
R54097	0.040	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000			
T72863	0.020	0.003	0.004	0.000	0.009	0.000	0.000	0.001	0.000	0.000	0.000	0.004	0.003	0.000	0.000	0.000	0.000	0.000		
J02854	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.001	0.000	

^aTwo probes that correspond to the same gene (R39465) were used in the microarray experiment.

agreement with the fact that the recursive partition kernel can efficiently treat gene–gene interactions because of its sequential nature. Correlation analysis of the two relationship indices suggests a lack of consensus between the two measures, with an estimate of -0.184 ($P = 0.011$), implying inefficiency of the target-independent relationship measure, Pearson's correlation of gene expressions, as a tool for seeking the gene cluster(s) to differentiate phenotypes.

Example 2. Mining molecular signatures for leukemia subtypes. Here, the target phenotypes are two distinct leukemia subtypes, AML and ALL. Thus, an ensemble decision analysis is conducted to identify the significant molecular signatures (subtype-relevant genes) that underpin the complex molecular mechanisms for distinction between the two subtypes. These data contain measurements corresponding to ALL and AML samples from bone marrow and peripheral blood. The raw gene expressions were re-scaled with a linear regression model by the data provider such that overall intensities for each chip are equivalent (2). Again, the same 5-fold CV resampling approach as in Example 1 is used to generate 500 replicates of pairs. An empirical threshold of 0.012 at a significance level of 0.01 is used to evaluate a putative molecular signature. Ensemble decision analysis of 7070 gene expressions identifies 23 highly significant ($P < 0.01$) feature genes (Table 4).

The top molecular signature, U16954 (AF1q mRNA), is a strongly relevant gene for distinction of the two acute leukemia subtypes and appears in $>55\%$ trees. Multiple lines of evidence from molecular biological studies imply that this gene is involved in leukemia development and progression. There are 117 NCBI nucleotide entries pertinent to this gene. The accumulated evidence over recent years has rendered it one of the most important molecular targets for leukemia. Cytogenetic studies (25–29) suggest that aberrations of chromosome band 1q11–q23, where AF1q resides, are among the most common chromosomal alterations in non-Hodgkin lymphoma. Tse *et al.* (1995) found that AF1q mRNA is highly expressed in the thymus but not in peripheral lymphoid tissues (26). Given so strong a relevance intensity for AF1q and encouraged by the wealth of biological evidence, we conceive that this gene is a key molecular determinant for phenotypic refinement of acute leukemia. It would also be interesting to explore its pivotal role as a central hub for the gene network that can differentiate the subtypes. Surprisingly, the importance of this gene was not revealed with either the eight marginal RankGene measures or by the data providers (2). On the other hand, there is good consensus on the importance of several other genes, including M27891 (CST3 Cystatin C), M23197 (CD33 antigen) and X95735 (Zyxin), between several independent analysis approaches [RankGene, Golub's method (2) and this study]. Again, the genes identified by Student's *t*-test are markedly different from the results obtained using other measures as implemented in the RankGene package (23) or using the proposed ensemble method.

An interesting concordance between the analyses of the two data sets is observed. Analyzing the colon data, we identified an important colon cancer gene, tyrosine-protein kinase receptor ELK precursor (*R.norvegicus*). In this data analysis, we identified its counterpart, protein tyrosine kinase, as a

Table 4. Ensemble decision analysis of 7070 gene expressions identifies 23 highly significant (empirical $P < 0.01$) molecular signatures for subtyping acute leukemias

Gene no.	Relevance	Name
U16954	0.554	(AF1q) mRNA
M27891	0.288	CST3 Cystatin C (amyloid angiopathy and cerebral hemorrhage)
M23197	0.258	CD33 CD33 antigen (differentiation antigen)
X95735	0.194	Zyxin
L42176	0.114	(clone 35.3) DRAL mRNA
L40391	0.086	(clone s153) mRNA fragment
M31523	0.082	TCF3 Transcription factor 3 (E2A immunoglobulin enhancer binding factors E12/E47)
U19517	0.070	(apoargC) long mRNA
U46499	0.062	Glutathione S-transferase, microsomal
L09209	0.052	APLP2 Amyloid β (A4) precursor-like protein 2
M20902	0.030	APOC1 Apolipoprotein CI
U73704	0.030	48 kDa FKBP-associated protein FAP48 mRNA
M84526	0.028	DF D component of complement (adipsin)
HG1496-HT1496	0.024	Adrenal-specific Protein Pg2
L19871	0.024	ATF3 Activating transcription factor 3
L40636	0.024	(clone FBK III 16) Protein tyrosine kinase (NET PTK) mRNA
U70063	0.020	Acid ceramidase mRNA
M11722	0.018	Terminal transferase mRNA
D88422	0.016	Cystatin A
M83652	0.016	PFC Properdin P factor, complement
L33477	0.014	(clone 8B1) Br-cadherin mRNA
L33842	0.012	(clone FFE-7) Type II inosine monophosphate dehydrogenase (IMPDH2) gene, exons 1–13
M95787	0.012	22 kDa smooth muscle protein (SM22) mRNA

significant molecular signature for differentiation of the leukemia subtypes. Whether this concordance implies a biological coherence between different cancers awaits verification. In short, the majority of the results from the ensemble analysis can be explained and are supported by molecular biological evidence or alternative approaches. However, caution should be taken in interpretation of the feature gene selection for refinement of the leukemia phenotype (two disease subtypes, without a normal control). The genes relevant to leukemia subtypes should be considered as important molecular signatures which may also be the disease genes. The target-driven interactions between these genes are presumably mapped to the underlying complex network bridging the two biological attributes.

Gene mining: classification of biological types

Again, we use the colon tissue data to explore the utility of the proposed ensemble approach for classification. Working on the same data allows us to show the differences between the two targets. As a result of the ensemble gene subset selection, three best subsets determined by their classification performance on the holdout samples using equation 3 are obtained, all with a χ^2 value of 9.118 ($P = 0.003$). Best subset 1 (Best tree 1) contains four genes: M26383 (human monocyte-derived neutrophil-activating protein mRNA, MONAP), T51849 (tyrosine-protein kinase receptor ELK precursor, *R.norvegicus*), Z24727 (*Homo sapiens* tropomyosin isoform mRNA) and H55758 (*H.sapiens* α -enolase). Best subset 2 also contains four genes: M26383, T94993 (*H.sapiens* fibroblast growth factor receptor 2 precursor), T58861 (60S ribosomal protein L30E, *Kluyveromyces lactis*) and R39465 (eukaryotic initiation factor 4A, *Oryctolagus cuniculus*). Best subset 3 contains five genes: M63391 (*H.sapiens* desmin gene), D14812 (*H.sapiens* KIAA0026 mRNA, complete cds),

H44011 (myosin heavy chain, non-muscle type A, *H.sapiens*), T58861 and H55933 (*H.sapiens* mRNA homolog of yeast ribosomal protein L41). To unravel the relationships between the two targets, classify the biological types and mine disease-relevant genes, we construct a classification rule using all 20 colon tumor-relevant genes. To allow for (lessen) selection bias due to either the same approach being used for feature gene selection and prediction or the induced rule being tested on tissue samples that had been used in the first instance to select the feature genes, we perform a *de novo* validation procedure called external cross-validation, with newly permuted data sets and with separate classifiers from that used for feature gene subset selection. The classifiers considered are a SVM with five different kernel functions, FLD, LNR, KNN and MD, reflecting the diversity of discriminant methods putatively useful for microarray data analysis. The averaged results from analysis of 500 pairs of data sets are shown in Tables 5 and 6, evaluated in terms of concordance rate (*Acc*), prediction accuracy for true positive (*pP*) and discriminant accuracy for true positive (*dP*), respectively.

Although there are some variations between the different classifiers, on average all four subsets (three best trees and the feature set with the 20 relevant genes) identified by our ensemble approach perform comparably with or better than the feature set with all 2000 genes. Best tree 3, although it does not include the top gene (M26383), achieves the highest performance across the multiple external classifiers and even performs better than the feature set of the top 20 colon tumor-relevant genes, with the highest performance (92.1%) attained using a SVM with a polynomial 2-D kernel, which is the highest attainable so far (4). The second best feature set is the top 20 relevant genes, reflecting the fact that the relevant genes are extracted from trees, which are in turn built with a target of high classification performance given a data structure.

Table 5. Classification performance of the support vector machine with different kernels, based on all the 2000 genes (2000 genes), 20 ensemble selected relevant genes (Ensemble 20 genes) and best trees

Kernel	Feature set	Acc	pP	dP
Linear	Ensemble 20 genes	0.888	0.907	0.931
	Best tree 1	0.844	0.852	0.931
	Best tree 2	0.802	0.844	0.861
	Best tree 3	0.919	0.951	0.927
	2000 genes	0.824	0.874	0.866
Polynomial 1-D	Ensemble 20 genes	0.888	0.906	0.930
	Best tree 1	0.832	0.817	0.967
	Best tree 2	0.814	0.815	0.935
	Best tree 3	0.887	0.896	0.943
	2000 genes	0.646	0.646	1.000
Polynomial 2-D	Ensemble 20 genes	0.888	0.907	0.931
	Best tree 1	0.822	0.848	0.897
	Best tree 2	0.783	0.840	0.831
	Best tree 3	0.921	0.942	0.941
	2000 genes	0.888	0.928	0.902
Polynomial 3-D	Ensemble 20 genes	0.847	0.879	0.900
	Best tree 1	0.836	0.858	0.907
	Best tree 2	0.778	0.831	0.838
	Best tree 3	0.911	0.935	0.932
	2000 genes	0.823	0.874	0.863
Polynomial 4-D	Ensemble 20 genes	0.835	0.875	0.864
	Best tree 1	0.842	0.888	0.879
	Best tree 2	0.774	0.836	0.823
	Best tree 3	0.887	0.932	0.898
	2000 genes	0.821	0.872	0.863

The reported values are averaged over analyses of 500 pairs of training and learning data sets generated by a 5-fold cross-validation resampling. Acc, pP and dP stand for concordance rate, prediction accuracy for true positives and discriminant accuracy for true positives, respectively.

Nevertheless, this feature set is neither necessarily the most economical (minimal) nor the most efficient set for classification or prediction because there are 'redundant' features among the top 20 genes (e.g. the two replicates of R39465). Indeed, mining these 'redundant' genes is one of major goals for ensemble decision analysis of microarrays.

DISCUSSION

Current methods for feature gene (or subset) extraction are guided by the target of prediction or classification of biological types, the basic strategy of which is to search for a single subset that leads to the best prediction of biological types, for example tumor versus normal tissue. Because of the very nature of these approaches and the target sought for, a large number of highly correlated genes or 'redundant' features need to be excluded from the 'best' subset. However, these so-called 'redundant' genes are in fact genes of critical importance in elucidating the complex genetic architecture of a complex disease. These genes might be neighboring genes (co-regulating genes) in a biological/biochemical path or genes in a parallel path or genes having epistatic actions. Strictly speaking, current methods aimed at prediction or tumor classification are unable to uncover this hidden pattern of genes on microarrays. As a result, one of the purposes of this study was to provide molecular biologists with a powerful and useable tool for extracting disease-relevant genes, a major theme in the post-genomic era. To our knowledge, we are the first to explicitly formalize the

task of mining disease-relevant genes using microarrays. Nevertheless, several scientists have recognized this issue and have been implicitly working in this direction (30,31). Likewise, the concept of ensemble decisions has been implicitly utilized both for gene selection for biological classification (32) and for mining disease-relevant genes (31).

The underlying biological complexity with which our proposed methodology deals is genetic heterogeneity, which is a thorny issue for both genetic linkage analysis (33,34) and microarray data analysis. The outward 'same' phenotype (e.g. affected or normal) can result from very different genetic and non-genetic pathways. A typical example is the discovery of novel tumor subtypes using genome-wide DNA profiles (2). The underlying rationale for our approach is: by sampling DNA samples (different combinations of phenotypes) and by tree-based recursive partitioning we can separate a mixture of heterogeneous samples into relatively homogeneous subgroups such that DNA samples belonging to one subgroup will presumably have the same genetic mechanism responsible for the differential phenotypic variations and by repeating this sampling strategy a large number of times we are able to capture the hidden multiple genetic pathways that lead to numerous genetic subtypes for a complex disease. Our analysis strategy is in fact an extension of the method of Shannon *et al.* (34), who proposed recursive partitioning of sibpair data into relatively more homogeneous subgroups such that within-subgroup analysis resulted in an increased power to detect linkage using Haseman-Elston regression (35).

Our ensemble decision approach has some analogies to random forests (11). For example, both approaches use a tree-based model as the framework for constructing a gene forest. However, there are several key differences. First, the research targets are different. Our ensemble method extracts relevant genes for classification of biological types or for disease gene mining, whereas most random forests focus on improvements in classification accuracy. The two approaches diverge after building the forests. After a large number of trees have been generated in random forests, they are used directly to vote for the most popular class, while in our ensemble method, the best feature subsets for prediction are selected based on the performance of an individual tree in classifying the external test samples. Second, a tree for random forests is constructed based on random selection of a feature subspace, while the tree for our ensemble decision approach is constructed based on random selection of a sample subspace. In other words, random forests are built on a one-directional (features) search and are thus inefficient in dealing with genetic heterogeneity (because it does not separate the hidden genetic subtypes of samples). On the other hand, our ensemble decision approach is built on a two-directional search (samples and features) under the assumption that genetic heterogeneity might be exhibited among the study samples. Third, although the random forests approach has been widely used for classification issues, its application to microarray data analysis and the extraction of disease genes has not been well investigated. Actually, this seminal method was proposed to achieve high accuracy of classification and robustness to noise. Although the author proposed a statistical measure to evaluate the importance of a feature variable, it is guided by classification (11) and thus it is uncertain whether this measure can be a good index for selecting a disease-relevant gene. It is a well-known

Table 6. Performance of four external classifiers, with 20 ensemble selected genes (Ensemble 20 genes), best trees or randomly selected 20 genes (Random 20 genes) as predictors

Method	Feature set	Acc	<i>pP</i>	<i>DP</i>
Fisher linear discriminant	Ensemble 20 genes	0.871 (0.088)	0.906 (0.089)	0.903 (0.095)
	Best tree 1	0.792 (0.107)	0.855 (0.105)	0.830 (0.124)
	Best tree 2	0.742 (0.104)	0.846 (0.094)	0.741 (0.142)
	Best tree 3	0.882 (0.087)	0.949 (0.070)	0.869 (0.122)
	Random 20 genes	0.592 (0.120)	0.751 (0.153)	0.600 (0.194)
Logistic regression	Ensemble 20 genes	0.765 (0.039)	0.853 (0.112)	0.787 (0.159)
	Best tree 1	0.826 (0.085)	0.858 (0.093)	0.890 (0.101)
	Best tree 2	0.794 (0.114)	0.850 (0.098)	0.841 (0.154)
	Best tree 3	0.866 (0.097)	0.922 (0.065)	0.877 (0.135)
	Random 20 genes	0.660 (0.125)	0.729 (0.105)	0.775 (0.108)
<i>K</i> -nearest neighbors	Ensemble 20 genes	0.887 (0.089)	0.909 (0.084)	0.925 (0.102)
	Best tree 1	0.635 (0.116)	0.713 (0.069)	0.742 (0.208)
	Best tree 2	0.824 (0.094)	0.850 (0.049)	0.895 (0.141)
	Best tree 3	0.835 (0.027)	0.873 (0.006)	0.858 (0.044)
	Random 20 genes	0.759 (0.086)	0.773 (0.093)	0.910 (0.077)
Mahalanobis distance	Ensemble 20 genes	0.724 (0.094)	0.874 (0.091)	0.673 (0.139)
	Best tree 1	0.776 (0.139)	0.913 (0.114)	0.732 (0.178)
	Best tree 2	0.823 (0.087)	0.849 (0.086)	0.893 (0.096)
	Best tree 3	0.904 (0.061)	0.933 (0.061)	0.925 (0.063)
	Random 20 genes	0.602 (0.144)	0.770 (0.156)	0.565 (0.188)

The reported values are averaged over analyses of 500 pairs of training and learning data sets generated by a 5-fold cross-validation resampling. Standard deviations are in parenthesis. *Acc*, *pP* and *DP* stand for concordance rate, prediction accuracy for true positives and discriminant accuracy for true positives, respectively.

fact, supported by this study, that the relevance of a feature does not imply that it is in the optimal feature subset and irrelevance does not imply that it should not be in the optimal feature subset for prediction (15).

An appealing and competitive alternative to recursive a binary partition model is a genetic algorithm coupled with a suitable evaluator [for example, SVM (X.Li and S.Rao, unpublished results) or KNN (8)] as the genetic algorithm itself is merely a search algorithm. A genetic algorithm is an adaptive search engine that mimics the natural selection process in genetics. 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 are computed, to avoid becoming stuck in a local optimum, which can occur with other search techniques. In addition, its merit over most other search algorithms, in which once a feature is taken in (or removed) it is never considered again, has rendered it a promising method for feature gene selection over a high-dimensional space. To apply this algorithm to hunt for target-relevant genes, several issues need to be resolved. First, a suitable fitness function (i.e. how well a feature subset survives over a specific criterion) has to be defined to map the biological reality (36). Second, the number of genes contained in the optimal feature gene subsets can be very large in the early generations. The coupled classifier(s) must have the capacity to handle very high-dimensional data, for which SVM is one of the best choices.

The globally optimal gene subsets identified by the novel ensemble approach have so far achieved the highest accuracy for classification of colon cancer tissue types. In the study, we have in fact performed two cross-validations, internal and external. The purpose of the internal cross-validation was to identify the 'optimal' gene subset(s). Then, the 'optimal'

subset(s) was subject to a further test by external cross-validation. The external cross-validation procedure with *de novo* permutation of the data 500 times and with evaluations using external classifiers that are separated from the internal classifier are performed to obtain an unbiased estimate of classification performance for the ensemble selected subset(s), although we cannot vouch for it that this external validation procedure is completely free from biases because the limited sample size does not allow us to reserve a part of the data as an independent test data set. Nevertheless, the estimate of final classification accuracy averaged over the *de novo* permuted 500 data sets and evaluated using external classifiers should have removed the biases due to the same algorithm being used in both learning and testing and the estimate is conceived to be the upper limit of the predicted accuracy if a novel and unseen, unlabeled DNA sample is to be classified. Another motivation for the sophisticated verification is to evaluate the generalization performance of a feature gene subset, for which Tasmardinos and Aliferis have given a nice discussion (Definition 2: feature selection problem 2) (37).

This study suggests that the proposed ensemble method holds the promise of deciphering some of the secrets of life from enormous amounts of molecular data generated by modern molecular technology. Application to two well-known microarray data sets proves it to be a powerful tool not only for prediction or classification but also for disease gene discovery. We propose the extraction of critical disease-relevant genes through multiple feature subsets, each being selected based on its classification performance. By resampling a large number of learning samples, the majority of strongly relevant and partially relevant genes can be mined. After this work, the next step is to address a more involved biological question: how do these genes act or interact to lead to the manifestation of a

disease phenotype, or so-called target-driven gene networking, which is currently under investigation.

ACKNOWLEDGEMENTS

We thank two anonymous reviewers for their comments on an early version of the manuscript. This work was supported in part by the National High Tech Development Project of China (grants nos 2003AA2Z2051 and 2002AA2Z2052), the National Natural Science Foundation of China (grants nos. 30170515 and 30370798) and the Cardiovascular Genetics Funds from Cleveland Clinic Foundation of USA.

REFERENCES

- DeRisi, J.L., Iyer, V.R. and Brown, P.O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680–686.
- 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.
- Burke, H.B. (2000) Discovering patterns in microarray data. *Mol. Diagn.*, **5**, 349–357.
- Ambroise, C. and McLachlan, G.J. (2002) Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl Acad. Sci. USA*, **99**, 6562–6566.
- Bo, T. and Jonassen, I. (2002) New feature subset selection procedures for classification of expression profiles. *Genome Biol.*, **3**, RESEARCH0017.
- Chow, M.L., Moler, E.J. and Mian, I.S. (2001) Identifying marker genes in transcription profiling data using a mixture of feature relevance experts. *Physiol. Genomics*, **5**, 99–111.
- Hastie, T., Tibshirani, R., Eisen, M.B., Alizadeh, A., Levy, R., Staudt, L., Chan, W.C., Botstein, D. and Brown, P. (2000) 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biol.*, **1**, RESEARCH0003.
- 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.
- Sun, M. and Xiong, M. (2003) A mathematical programming approach for gene selection and tissue classification. *Bioinformatics*, **19**, 1243–1251.
- Breiman, L. (1996) Bagging predictors. *Machine Learn.*, **24**, 123–140.
- Breiman, L. (2001) Random forests. *Machine Learn.*, **45**, 5–32.
- Mills, J.C. and Gordon, J.I. (2001) A new approach for filtering noise from high-density oligonucleotide microarray datasets. *Nucleic Acids Res.*, **29**, e72.
- Hall, M. (1998) Correlation-based feature selection for machine learning. PhD thesis, University of Waikato, Hamilton.
- Blum, A.L. and Langley, P. (1997) Selection of relevant features and examples in machine learning. *Artif. Intell.*, **97**, 245–271.
- Kohavi, R. and John, G.H. (1997) Wrappers for feature subset selection. *Artif. Intell.*, **97**, 273–324.
- Xing, E.P., Jordan, M.I. and Karp, R.M. (2001) Feature selection for high-dimensional genomic microarray data. In *Machine Learning: Proceedings of the Eighteenth International Conference*, San Francisco. Morgan Kaufmann, San Mateo, CA.
- Bell, D.A. and Wang, H. (2000) A formalism for relevance and its application in feature subset selection. *Machine Learn.*, **41**, 175–195.
- Dietterich, T.G. (2000) Ensemble methods in machine learning. In Kitter, J. and Roli, F. (eds), *First International Workshop on Multiple Classifier Systems, Lecture Notes in Computer Science*. Springer Verlag, New York, NY, pp. 1–15.
- 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.
- Guo, Z., Li, X. and Rao, S. (2001) *Analysis of Medical Data: An Introduction to Bioinformatics*. Harbin Publishers, Harbin, China.
- Zhang, H., Yu, C.Y., Singer, B. and Xiong, M. (2001) Recursive partitioning for tumor classification with gene expression microarray data. *Proc. Natl Acad. Sci. USA*, **98**, 6730–6735.
- Kowalski, J. and Denhardt, D.T. (1989) Regulation of the mRNA for monocytic-derived neutrophil-activating peptide in differentiating HL60 promyelocytes. *Mol. Cell. Biol.*, **9**, 1946–1957.
- Su, Y., Murali, T.M., Pavlovic, V., Schaffer, M. and Kasif, S. (2003) RankGene: identification of diagnostic genes based on expression data. *Bioinformatics*, **19**, 1578–1579.
- Yeoh, E.J., Ross, M.E., Shurtleff, S.A., Williams, W.K., Patel, D., Mahfouz, R., Behm, F.G., Raimondi, S.C., Relling, M.V., Patel, A. *et al.* (2002) Classification, subtype discovery and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, **1**, 133–143.
- Lestou, V.S., Ludkovski, O., Connors, J.M., Gascoyne, R.D., Lam, W.L. and Horsman, D.E. (2003) Characterization of the recurrent translocation t(1;1)(p36.3;q21.1-2) in non-Hodgkin lymphoma by multicolor banding and fluorescence *in situ* hybridization analysis. *Genes Chromosomes Cancer*, **36**, 375–381.
- Tse, W., Zhu, W., Chen, H.S. and Cohen, A. (1995) A novel gene, AF1q, fused to MLL in t(1;11)(q21;q23), is specifically expressed in leukemic and immature hematopoietic cells. *Blood*, **85**, 650–656.
- Busson-Le Coniat, M., Salomon-Nguyen, F., Hillion, J., Bernard, O.A. and Berger, R. (1999) MLL-AF1q fusion resulting from t(1;11) in acute leukemia. *Leukemia*, **13**, 302–306.
- Watanabe, N., Kobayashi, H., Ichiji, O., Yoshida, M.A., Kikuta, A., Komada, Y., Sekine, I., Ishida, Y., Horiukoshi, Y., Tsunematsu, Y. *et al.* (2003) Cryptic insertion and translocation or nondividing leukemic cells disclosed by FISH analysis in infant acute leukemia with discrepant molecular and cytogenetic findings. *Leukemia*, **17**, 876–882.
- Li, Z.G., Wu, M.Y., Zhao, W., Li, B., Yang, J., Zhu, P. and Hu, Y.M. (2003) Detection of 29 types of fusion gene in leukemia by multiplex RT-PCR [in Chinese]. *Zhonghua Xue Ye Xue Za Zhi*, **24**, 256–258.
- Szabo, A., Boucher, K., Jones, D., Tsodikov, A.D., Klebanov, L.B. and Yakovlev, A.Y. (2003) Multivariate exploratory tools for microarray data analysis. *Biostatistics*, **4**, 555–567.
- Chilingaryan, A., Gevorgyan, N., Vardanyan, A., Jones, D. and Szabo, A. (2002) Multivariate approach for selecting sets of differentially expressed genes. *Math. Biosci.*, **176**, 59–69.
- Li, L., Darden, T.A., Weinberg, C.R., Levine, A.J. and Pedersen, L.G. (2001) Gene assessment and sample classification for gene expression data using a genetic algorithm/k-nearest neighbor method. *Comb. Chem. High Throughput Screen.*, **4**, 727–739.
- Province, M.A., Shannon, W.D. and Rao, D.C. (2001) Classification methods for confronting heterogeneity. *Adv. Genet.*, **42**, 273–286.
- Shannon, W.D., Province, M.A. and Rao, D.C. (2001) Tree-based recursive partitioning methods for subdividing sibpairs into relatively more homogeneous subgroups. *Genet. Epidemiol.*, **20**, 293–306.
- Haseman, J.K. and Elston, R.C. (1972) The investigation of linkage between a quantitative trait and a marker locus. *Behav. Genet.*, **2**, 3–19.
- Cho, S.J. and Hermesmeier, M.A. (2002) Genetic algorithm guided selection: variable selection and subset selection. *J. Chem. Inf. Comput. Sci.*, **42**, 927–936.
- Tsamardinos, I. and Aliferis, C.F. (2003) Towards principled feature selection: relevance, filters and wrappers. In *Ninth International Workshop on Artificial Intelligence and Statistics*, Key West, FL.