

Selecting maximally informative genes

Ioannis P. Androulakis*

Biomedical Engineering Department, Rutgers University, 617 Browser Road,
Piscataway, NJ 08854, USA

Received 22 September 2003; received in revised form 2 March 2004

Abstract

Microarray experiments are emerging as one of the main driving forces in modern biology. By allowing the simultaneous monitoring of the expression of the entire genome for a given organism, array experiments provide tremendous insight into the fundamental biological processes that translate genetic information. One of the major challenges is to identify computationally efficient and biologically meaningful analysis approaches to extract the most informative and unbiased components of the microarray data. This process is complicated by the fact that a number of uncertainties are associated with array experiments. Therefore, the assumption of the existence of a unique computational descriptive model needs to be challenged. In this paper, we introduce a framework that integrates machine learning and optimization techniques for the selection of maximally informative genes in microarray expression experiments. The fundamental premise of the approach is that maximally informative genes are the ones that lead to least complex descriptive and predictive models. We propose a methodology, based on decision trees, which identifies ensembles of groups of maximally informative genes. We raise a number of computational issues that need to be comprehensively addressed and illustrate the approach by analyzing recently published microarray experimental data.

© 2004 Elsevier Ltd. All rights reserved.

Keywords: Maximally informative genes; Microarray experiments; Genetic information; Machine learning; Optimization

1. Microarray experiments: brief introduction and major limitations

The goal of modern biology is to bridge the gap between the genetic information at its most elementary level (genotype: structure of the genetic material) and the collective expression of behavior (phenotype: characteristics of an organism that can be observed, assayed, measured). The phenotype is a function of genetics, environment and experimental conditions. The successive steps of translating sequence to structure to function are highly complex and convoluted, hence, very difficult to be described using first principles models. However, various experimental techniques have been developed that have revolutionized the way we look at complex biological systems since they allow to monitor changes during the process of transforming genetic information either at the gene level (genomics), the protein level (proteomics) or

the metabolite level (metabolomics). The goal of such analyses is to generate the data required to derive descriptive, as well as predictive, models that provide a fundamental understanding to the complexities of life with tremendous implications in improving life quality.

1.1. Monitoring changes at the genome level

The genetic information is stored in the DNA, the double-stranded polymer composed of four basic molecular units (nucleotides): adenine (A); guanine (G); cytosine (C) and thymine (T). In order for the genome to direct, or affect, changes in the cell a transcriptional program must be activated eventually dictating all biological transformations. This program is regulated temporarily according to an intrinsic program or in response to changes in the environment. The expression of the genetic information, which is stored in DNA, takes place in two stages: *transcription*, during which DNA is transcribed into mRNA, a single stranded complimentary

* Tel.: +1 732 445 0099; fax: +1 732 445 3753.

E-mail address: yannis@rci.rutgers.edu.

copy of the base sequence of the DNA, and *translation*, during which mRNA provides the blue-print for the production of specific proteins. Therefore, measuring the level of production of mRNA, thus measuring the expression levels of the associated genes, provides a quantitative assessment of the levels of production of the corresponding protein.

Innovative approaches such as cDNA and oligonucleotide microarrays were recently developed to extract genome-wide information related to gene expression (Bowtell, 1999; Brown & Botstein, 1999; Cheung et al., 1999; Lipshutz, Fodor, Gingeras, & Lockhart, 1999; Schena, Shalon, Davis, & Brown, 1995). During an expression experiment extracted mRNA is reverse-transcribed into more stable complementary DNA (cDNA) which is labeled using fluorescent dyes. Different colored dyes are used for different samples (probes). The probes are then tested by hybridizing to a DNA array. The array holds thousands of spots, each containing a different DNA sequence. Once the probes have hybridized, they are washed off and the array is scanned to determine the relative amount of each cDNA probe bound to any given spot. Quantitative imaging coupled with clone database information allows measurement of the labeled cDNA that hybridized to each target sequence. Image processing and data normalization are among the first, and very critical, computational filters required before the actual quantification of the expression experiment is defined (Dudoit, Wang, Callow, & Speed, 2000). Gene expression changes are usually measured relative to another sample. Comparative differences are used to assess the impact of gene expression to various regulatory pathways. A number of experiments can thus be designed to address a variety of issues. For instance:

- (a) Diversion from normal physiology is frequently accompanied by changes in gene expression patterns. Therefore, genes inappropriately transcribed cause diseases like cancer. Comparison of the expression profiles of such cells provides the basis for the understanding of the genetic causes of a disease.
- (b) By monitoring the changes in the expression levels of a genome in the presence of environmental changes provides the beginning for a fundamental understanding of the causes of the response in the presence of an environmental stimulus.

Gene expression microarray experiments have been celebrated as a revolution in biology, attracting significant interest, because they are slowly changing the working paradigm of biological research by allowing the analysis of the combined effects of numerous genetic and environmental components. The profound impact is that such global analysis methods will allow a fundamental shift from “. . . piece-by-piece to global analysis and from hypothesis driven research to discovery-based formulation and subsequent testing of hypotheses . . .” (Kafatos, 2002).

One of the major challenges is to extract in a systematic and rigorous way the biologically relevant components from the array experiments in order to establish meaning-

ful connections linking the genetic information to cellular function. Because of the significant amount of experimental information that is generated (expression levels of thousands of genes) computer-assisted knowledge extraction processes are the only realistic alternative for managing such an information deluge.

1.2. Understanding the limitations of array experiments

When designing algorithms that analyze array data it is imperative to understand the fundamental limitations of the experiment so as to account for those in the development of the algorithms. Array experiments are characterized by a number of inherent limitations:

1. The most fundamental implicit concept in an array experiment is the so-called “central dogma”. The main assumption is that the information flow is forward and sequential: transcription, translation, biotransformation. However, given the complexities of biological functions, it is not necessarily clear that this sequential description captures appropriately the convoluted relationship between gene expression and post-transcriptional/translational modifications or initial and secondary effects in gene expression. For instance, Hatzimanikatis and Lee (1999) tested the proposition whether simply monitoring mRNA expression data is sufficient to elucidate the relationship between genome sequence and gene regulation.
2. Array experiments are not necessarily “clean” experiments and the data they generate are interpretations of measurements rather than hard data. Furthermore, tremendous variability and uncertainty exist not only because of biological fluctuations, but also as a result of the processing of the experiment itself. Image processing plays a major role in array experiments and a result the information used as input to statistical analyses is an interpretation of the fluorescent image. Location of spot center segmentation, intensity, normalization, signal to noise ratio and interference are but a few of the issues that have to be addressed and appropriately interpreted.
3. Finally, we have to seriously consider the lack of available data in array experiments. Although a large number of genes are monitored during the experiment, we must also realize that, in general, we have a very limited number of cells that are analyzed and an even smaller, if any, number of repeats to statistically validate the robustness of the measurement. Simply put, we have a much larger number of independent (input) variables that we measure compared with the number of experiments (output variables) that we generate. In principle, when the ratio of experiments/variables is very small, it is highly unlikely that we can correctly capture the inherent non-linear structure of the experiments and the relationship between input and output variables.

1.3. The need for an alternative approach

The above-mentioned limitations of the array experiments are but a few of the issues challenging the predictive capability of the interpretation of these measurements. Microarray analysis should not be considered as the conclusion to an experiment, but as a discovery mechanism to help determine which avenues to pursue further. However, it is also well accepted that array experiments provide tremendous insight to fundamental biological functions. Therefore, it is imperative to design the analysis techniques in a way that understands and captures these limitations. In the presence of such uncertainties, most biological questions will not be readily answered in a quantitative manner by array experiments. Instead, the most likely outcome from a functional genomics analysis is the next biological question to ask. Furthermore, given the nature of biological systems multiple hypotheses should be generated that could lead to various plausible alternatives that will subsequently be verified either theoretically or experimentally.

A number of excellent publications have focused on different aspects of gene expression experiments, primarily for clustering of cells and genes (Alizadeh et al., 2000; Allander et al., 2001; Alon et al., 1999; Bittner et al., 2001; Dudoit et al., 2000; Golub et al., 1999; Luo et al., 2001; Perou et al., 1999; Pollack et al., 1999; Ross et al., 2000). The development of novel computational approaches that exploit large warehouses of gene expression data have been identified as major enablers for realizing fully the potential of this technology (Basset, Eisen, & Boguski, 1999). The main focus of these analyses was to derive a single interpretation of the data. That is a single model that maximizes the predictive accuracy of a classifier for example. Further analyses of the computational results attempt to assign a certain level of significance to smaller subsets of genes whose expression patterns could potentially indicate a more direct involvement during the transcription process. Reducing the number of genes is significant not only from a biological point of view, since functionality can be inferred, but also from a computational point of view. The analysis of array experiments is much like an identification problem where we try to establish a link between an observation (i.e., whether a cell is cancerous or not) and a measurement (i.e., the expression of a gene). Reducing the number of measured variables reduces the degrees of freedom, hence avoids pointless over-fitting. Too few genes will not discriminate or predict, but too many genes might be introducing noise to the model rather than information. Therefore, the identification of informative genes is a significant component of an integrated computer assisted analysis of array experiments. However, in current practice the identification of such a critical sub-set of genes whose expression is informative is accomplished as a by-product of some other activity. For instance, by analyzing expression patterns in clustering, the loading of singular vector in SVD or by assessing the ability of certain genes to maximize the separability between classes.

The goal of our work is to develop a comprehensive modeling framework that, while optimizing the performance of a classifier

1. explicitly addresses the problem of identifying maximally informative genes, and
2. generates a number of possible alternatives and not a single interpretation of the data.

We further realize that in order to derive truly informative genes we must explicitly account for the complexity of the model. Our fundamental hypothesis being that among models that utilize the same number of variables the most “informative” ones are the those that introduce the least complexity (Ockam’s razor). We will, therefore, show that the complexity of the model has to be taken into account while searching for the most informative genes. Therefore, when looking for informative genes we would like to propose an approach that

1. explores the complexity of a number of models, and not a single model;
2. analyzes the robustness of the outcome of the model and not the model itself;
3. minimizes the complexity of the input/output relationship.

In this paper, we will first provide some of the fundamental concepts in machine learning required for our development. We, then, propose a multilevel optimization framework using classification trees as the discriminatory model and show how we model complexity. Finally, we use examples from publicly available datasets to illustrate the basic principles of the approach.

2. Machine learning preliminaries

2.1. Dimensionality reduction and feature subset selection

In a typical problem in machine learning, the input space is composed of n -dimensional vectors whose components are also denoted as “features”. Given this representation, a number of questions can be identified in order to extract regularities and patterns from the data. For example, in classification, we are looking for a “discriminant” function, which operates on the n -dimensional input space and whose output is a logical decision assigning the input feature vector to a specific class. The class defines a collection of items that all share some common relationship. A fundamental problem in machine learning is to find ways to succinctly represent this multi-dimensional input space. In other words, we are looking for ways to reduce the dimensionality of the data, hence overcoming the problem of over-fitting.

Techniques identifying optimal projections of the data on a few leading directions, such as principal component analysis, are commonly used in order to reduced the dimensionality of the representations. The main drawback of these methods is that the new “features” are linear combinations of the original

ones, and therefore, assigning a physical significance to them is still an open issue. Therefore, a number of approaches have been developed over the years that attempt to identify subsets of the original set of features that capture most of the intrinsic structure of the original set. The problem of feature selection can be stated simply as “. . . choose the best subset of N properties from a set of M ”. These authors formalized the process of selecting properties that contain the most discriminatory information in order to derive decision rules that lead to classification schemes. They present and analyze seven techniques for ranking the information content of each feature in an attempt to derive good sub-sets. Narendra and Fukunaga (1977) present a more formal approach based on a branch and bound scheme for addressing the very same problem. A recent review by Kohavi and John (1995) discusses a number of issues. More recently, Liu and Motoda (2000) also present ideas related to the coupling of information theory and feature selection. Recently, promising approaches that emerge are optimization-based approaches were also explored, such as support vector machines (Bradley and Masagarian, 1998).

Among the leading approaches for feature selection, especially when analyzing gene expression experiments, are methods that attempt to quantify the contribution of individual features to the separation of the classes in the data. Golub et al. (1999), for example, devised such a metric in order to derive ranking criteria for the genes. However, these methods fail to take into account mutual information between features.

2.2. Decision trees

Decision trees are one the most widely applicable non-parametric computational method used in classification. They are a well-established methodology and a number of excellent implementations are currently available (Breiman, Friedman, Olshen, & Stone, 1984; Quinlan, 1993). Decision trees, either classification or regression, are a particularly attractive type of models for three main reasons. First, they have an intuitive representation, the resulting model is easy to understand and assimilate by humans. Second, the decision trees are non-parametric models, no intervention being required from the user, and thus they are very suited for exploratory knowledge discovery. Third, scalable algorithms, in the sense that the performance degrades gracefully with the increase of the size of training data, exist for decision tree construction models.

Initially, one starts with a training set in which the classification label is known (pre-classified) for each record. All of the records in the training set are together in one big box. The algorithm then systematically tries to break up the records into two parts, examining one variable at a time and splitting the records on the basis of a dividing line in that variable. The object is to attain as homogeneous set of labels as possible in each partition. This splitting or partitioning is then applied to each of the new partitions. The process continues until no more useful splits can be found. The heart of the algorithm is the rule that determines the initial split rule. The splitting of records according to the rules defines a binary decision tree.

Fayyad and Irani (1990) showed that given two decision trees consistent with the data, the simplest one, i.e., the one requiring the fewest number of partitions (classification rules) should be preferred. In the machine learning community, the notion that the accuracy of the model is associated with its simplicity dates back at least as far as the 14th century. It was then that William of Ockam first proposed his famous razor: “one should not increase, beyond what is necessary, the number of entities required to explain anything”. More recently, machine learning researchers have followed this principle in biasing their algorithms toward finding hypotheses with simple representations. Holte (1993) reports the results of experiments measuring the performance of very simple rules on the data-sets commonly used in machine learning research. The idea of simplicity in classification rules is the one that we further explored in this study and its implications for feature selection further explored.

2.3. Feature selection and measures of complexity

For the purpose of the analysis, we present here we assume that our search is supervised. In other words, we will assume that enough corroborating information regarding class assignment of the records (cells) exists so that class assignment is known a priori. Hence, one should, in principle, be able to derive a set of rules that, given the complete set of features, will determine correctly the class assignment with the least complexity. Our main assumption is, therefore, the following: a minimal set of maximally informative features is the smallest set of features, for a given classification algorithm, which achieves the best possible classification at minimum complexity. Furthermore, the complexity of the classification scheme is defined as the number of rules, alternatively the number of leaf nodes, in the binary decision tree. This is our first approach and currently additional complexity measures are under investigation.

Recently information theoretic criteria were used to formulate a similar problem in the context of logistic linear regression (Li & Wang, 2002). These approaches make use of a balanced representation of complexity, as measured by the number of parameters to be determined in a linear model, and accuracy. Two such measures typically used are:

1. Akaike information criterion (AIC) = $-\log(L) + 2K$;
2. Bayesian information criterion (BUIC) = $-\log(L) + \log(N)K$;

where K is the number of parameters in the model and L , the maximum likelihood.

3. A flexible framework for selecting maximally informative genes

Based on the previous discussion, we wish to develop a framework for selecting maximally informative genes such that

- (a) the accuracy of the classifier, based on decision trees, is maximized;
- (b) the classifier has minimum complexity, as measured by the number of rules;
- (c) the least number of features is used.

The two measures of complexity (b and c) are both required in order to best define the overall complexity of the classifier. The search for the informative features can be, therefore, formally defined in terms of the following multi-level optimization problem:

$$\begin{aligned}
 & \min \|C^{\text{calc}} - C^{\text{expt}}\| \\
 & \text{subject to :} \\
 & C^{\text{calc}} = T(\lambda_i, \quad i = 1, \dots, N) \\
 & \text{subject to :} \\
 & \min \text{ complexity} \\
 & \text{subject to :} \\
 & \text{complexity} = T(\lambda_i, \quad i = 1, \dots, N) \\
 & \min \sum_{i=1}^N \lambda_i \\
 & \lambda_i = \begin{cases} 1, & g_i \in I_G \\ 0, & g_i \notin I_G \end{cases}
 \end{aligned} \tag{1}$$

$$I_G = \{g_i \ni \lambda_i = 1\} \subseteq G = \{g_i, i = 1, \dots, N\}$$

We will now present the details of this formulation.

3.1. Modeling and optimization

The objective in (1) measures the accuracy of the classifier. C^{calc} and C^{expt} are vectors containing the class assignment of the samples. C^{expt} denotes the actual assignment whereas C^{calc} is the assignment derived based on the classifier. The latter depends of the number of features and the particular decision tree that is derived and is implicitly defined via the use of the classifier, denoted in our formulation as $C^{\text{calc}} = T(\lambda_i, i = 1, \dots, N)$. The “norm” $\|C^{\text{calc}} - C^{\text{expt}}\|$ can be defined in a number of different ways: count of the number of erroneous predictions, percent of erroneous prediction, etc. The functional form does not impact the development. Once the classifier has been applied to a given set of features its complexity, in terms of the number of classification rules required, is identified and is denoted as “Complexity” in (1). Clearly, the complexity of the classification tree, i.e., the number of rules, is a function of the algorithm used for the construction of the tree as well as the specific sub-set of features that is used. This defines the second level of the optimization as we search for the minimum possible complexity in the classifier. As noted earlier, the classification tree and the corresponding rules for a given set of features are derived based on the C4.5 classification algorithm (Quinlan, 1993). The interested reader is advised to consult the original reference for more details regarding the construction of trees based on the C4.5 implementation. These are not presented here. The

algorithm is very robust one and it can be very easily incorporated within the overall scheme. However, it should be pointed out that the fundamental hypothesis defining the relationship between informative features and the complexity of the classification rules does not dependent on the classification algorithm. Finally, feature selections are modeled through the use of appropriate binary variables, one for each gene (g_i), such that

$$\lambda_i = \begin{cases} 1, & g_i \in I_G \\ 0, & g_i \notin I_G \end{cases} \tag{2}$$

The value of the binary variable is 1 if the particular gene is to be incorporated in the classifier, 0 otherwise. The minimal set of informative genes is defined as $I_G = \{g_i \in \lambda_i = 1\} \subseteq G = \{g_i, i = 1, \dots, N\}$. The set of informative genes is a subset of the original set of genes. The third level or required decision is thus defined by minimizing the number of active features in the model.

3.2. Model simplification and solution methodology

The optimization problem defined in Eq. (1) is a multi-level non-linear integer optimization problem, an alternative way to introduce model complexity in terms of the number of features, that does not require the solution of the third level feature minimization problem, is to derive a formulation parametric in the number of features. That is replacing the third level by a constraint of the form: $N_G = \sum_i \lambda_i = M < N$ and the problem is solved parametrically for increasing values of M in order to identify $G(M)$ and the corresponding error $E(M, G(M))$. Furthermore, a penalty formalism is used to combined two of the objectives as follows: $\langle \text{ClassErr} \rangle + \text{ClassTree}$. This approach replaces the norm $\|C^{\text{expt}} - C^{\text{calc}}\|$ and the second level optimization defining complexity with a metric which sums the contributions in terms of the accuracy of the classifier as well as the size of the classification tree. The classification error, $\langle \text{ClassErr} \rangle$, is set to 0 if no wrong assignments are made, otherwise is set equal to the number of miss-classified samples. The complexity in terms of the size of the classification tree, ClassTree , is accounted for by augmenting the objective with the term representing the actual number of classification rules.

These “simplifications” although they do not alter the character of the problem, they significantly reduce its computational complexity as they avoid the explicit treatment of the multi-objective optimization problem defined in (1). We have shown (Androulakis & Hatzimanikatis, 2004) that the cardinality of the minimal set of informative features is quite small for a number of array expression experiments. However, the actual size of the optimization problem is very substantial. Mathematical programming optimization techniques, Floudas (2000) are prohibitive given the size and non-linear nature of the problem. Therefore, a variant of

simulated annealing (Aarts, Korst, & van Laarhoven, 1997) has been implemented. In our implementation, since the number of active genes is maintained constant (M), the simulated annealing moves are defined in a way that the number of active genes (features) is maintained constant. More specifically:

1. choose randomly i and j such that $\lambda_i + \lambda_j = 1$;
2. swap the value of λ_i and λ_j .

Nevertheless, we are currently exploring more rigorous approaches for the robust and deterministic solution of the aforementioned optimization problem.

3.3. Generating ensembles of solutions

The inherent uncertainties characterizing the array experiments, as discussed earlier, require the derivation of a set of robust solutions as opposed to a unique combination of genes that optimally perform a particular task. It is, therefore, imperative to generate stable ensembles of solutions with similar, if not identical, predictive and complexity characteristics. The analysis of these ensembles should lead to the identification of emerging patterns, that is combinations of genes whose coordinated expression profiles define stable and robust sub-structures, members of a collection of accurate models. We should, therefore, require the algorithms to possess the ability of generating ensembles of solutions whose concurrently expressed genes will define targets for further analysis based on the argument that robust emerging patterns have captured key inherent and fundamental biological characteristics.

The stochastic nature of simulated annealing makes it ideal for similar types of analyses. Multiple runs are expected to generate, i.e., converge to, diverse solutions. However, further analysis of the generated optimal models will reveal whether common motifs exist across the various solutions that have been identified. In the following section, we will demonstrate how multiple solutions lead to the identification of robust patterns of expression. Furthermore, we will demonstrate that the robust components of the solution vectors can be rationalized biologically providing experimental evidence to the fact that computational robustness can lead to the generation of interpretable hypotheses and could be linked to biological robustness as well.

4. Analysis of computational results

In an upcoming publication (Androulakis & Hatzimanikatis, 2004), we present a large computational study in which various publicly available data-sets are analyzed and maximally informative solutions are identified. In this paper, we will be focusing on a single case so as to illustrate some of the main computational points that serve as the guidelines for our current development.

4.1. Generating multiple sets of maximally informative genes

“Small round blue cell tumors” (SRBCT) is a descriptive category encompassing a large number of malignant tumors that tend to occur in childhood. They are united by their similar histo-pathological appearance. However, subtle clues may be present to distinguish between the tumors. For proper characterization, pathologists often employ immunohistochemistry, electron microscopy, and molecular analysis for chromosomal abnormalities. The SRBCTs of childhood include neuroblastoma (NB), rhabdomyosarcoma (RMS), non-Hodgkin lymphoma (NHL), and the Ewing family of tumors (EWS). Currently, no single biological or chemical test exists that can detect SRBCTs. Khan et al. (2001) presented a comprehensive study in which a large number of genes were monitored. The data were reduced by SVD decomposition and the leading factors were used to train an artificial neural network (ANN) to build a predictive diagnostic device. This study constitutes a milestone since it was the first attempt to use microarray experiments in a predictive way to explore the potential application of using such methods for tumor diagnosis. Computationally, it is a very interesting problem. It is of relatively modest size, containing the expression levels of 2303 genes, with 63 cells, belonging to four cancer types, used for training purposes. An additional set of 25 cells will be used for testing the accuracy of the classifier. This is discussed later in the text. Khan et al. (2001) construct their ANN by using as inputs the projection of the expression measurements onto the first 10 principal directions as determined by a principal component analysis of the raw expression data. An exhaustive sensitivity analysis identified a sub-set containing the 96 most informative genes. The genes are ranked based on their ability to discriminate the four classes.

Implementation of our proposed approach determines that the minimal number of informative genes is 3. Multiple runs of the selection algorithm consistently identify 9 combinations of three genes that perfectly classify the data with a corresponding tree of nine nodes. The total number of genes participating in the 9 combinations is 19 and the list of gene

21652	784224	1435862
377048	784224	1435862
745019	784224	1435862
308231	784224	1435862
745019	784224	1435862
308231	784224	841620
308231	784224	813742
308231	745019	784224
769716	859359	1435862
295985	745019	784224

Fig. 1. Ensemble of 9 three-gene combinations that perfectly classify the SRBCT data.

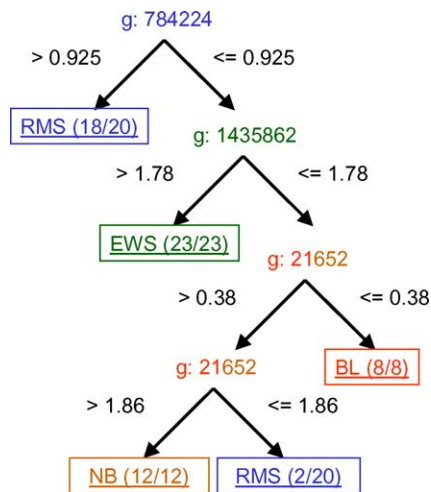


Fig. 2. Optimal classification tree using three genes and nine nodes for perfect sample assignment.

identifiers is depicted in Fig. 1 (we have used the gene identifiers as provided in the original publication of Khan et al., 2001). The classification tree for the first solution is depicted in Fig. 2. Fig. 3 monitors the progression of a typical run as the combined objective encompassing error and complexity is minimized. Each point in this plot corresponds to a distinct, i.e., different combination of genes, solution identified during the search. Qualitatively, the sum of the two curves expresses a weighted average between model accuracy (classification error) and model complexity (nodes of the classification tree). This is of course very reminiscent to the various information criteria often used in model selection such as the Akaike and Bayesian information criteria (AIC and BIC, respectively), Burnham and Anderson (1998). The plot exemplifies the interplay between model accuracy and model complexity. The basic conjecture made in this paper is that model complexity as measured by the number of nodes in the tree, and not by the number of features alone, should be the indicator differentiating the various models.

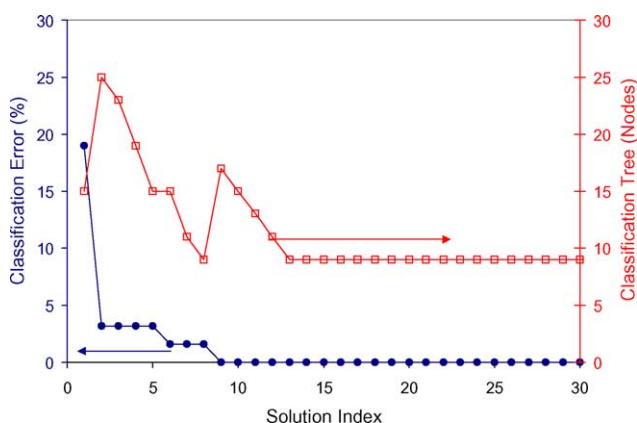


Fig. 3. Progress of the search for optimal gene combinations of cardinality.

According to our results, although 9 independent solutions were obtained half of them share two common genes, 784224 (FGFR4) and 1435862 (MIC2). Fig. 4 depicts the associated rules for the solutions that share the two genes that form this “emerging pattern”. By analyzing these solutions we see that the FGFR4 and MIC2 genes are primarily responsible for differentiating the RMS and EWS cancer types. The third gene is used to further differentiate between the two additional cancer types. In fact, what we have uncovered is a robust combination of genes with significant biological implications.

FGFR4 seems to be a strong indicator of RMS since 90% of the training data that belong to this category are classified according to the first rule, which only involves FGFR4 (Fig. 1). This is very consistent with available biological information indicating the FGFR4, a tyrosine kinase receptor expressed during myogenesis, is known to have inappropriate expression patterns that link them to breast cancer (Dickson, Spencer-Dene, Dillon, & Fantl, 2000). Although FGFR4 warrants further study, it cannot be singled out since it is not only present in other cancers but also, based on the data just presented, that gene alone does not fully predict all RMS occurrences in the data. MIC2 is a gene known to be associated with EWS; and in fact, immunostaining of it is currently used to diagnose EWS (Kovar et al., 1990). However, once again, it alone cannot be used to differentiate EWS as 20% of the RMSs exhibit MIC2 overexpression, based on the rules that were identified (Fig. 1). Therefore, one should begin to rationalize:

- the synergistic effects of genes that co-occur, and
- the relationship of the single genes that are interchangeable.

The advantage of such a methodology is that it allows the theoretical and experimental biologist to rationalize the actual implications of the genetic behavior of the individual genes which appear interchangeable, as well as the synergistic effects of the genes that appear to be active simultaneously. The approach allows also for specifying the required complexity. As a result, imposing an additional constraint that sets the number of active genes can also identify a number of additional solutions with four or more genes.

Recently, Deutsch (2003) generated a number of, unspecified, predictors containing 12 genes. However, we determined that the minimum required number of genes is 3, which is a significant reduction in complexity. Furthermore, we determined that the stable 2-gene is composed of genes whose biological importance can be justified.

4.2. Predictions on unseen data

The main focus of this work is in identifying improved approaches for analyzing available microarray data in an effort to identify potential leads that drive the observed responses that warrant further analyses. However, because arrays can provide significant biological information at a very refined

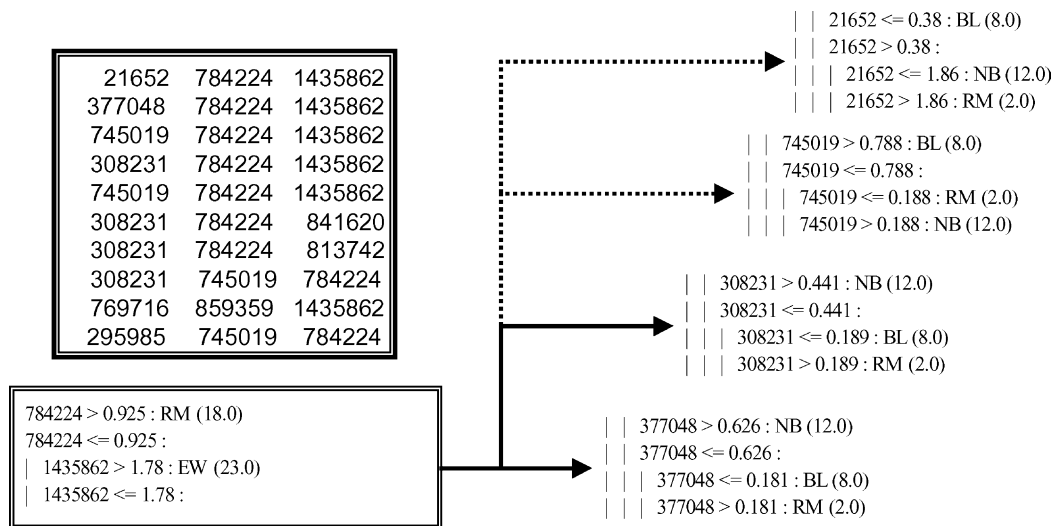


Fig. 4. Structure of solutions sharing the two-gene (784224, 1435862) emerging pattern.

resolution, it is also hoped that they could also be used to develop diagnostic devices so as to eliminate current practices that in many cases involve subjective decision. The novel work of Khan et al. (2001) was a first step in this direction. Therefore, building predictive models is one of the areas that will have to be addressed properly.

One of the advantages of the approach that was described is the fact that an ensemble of solutions (optimal classifiers) can be generated. This leads naturally to exploring the possibility of using those ensembles in order to perform predictions on unseen data. The idea of ensemble methods in machine learning has been explored in a number of occasions (Dietterch, 2000). The general concept being that in order to construct classifiers operating on new data points, a weighted prediction vote combining predictions of independent classifiers should be implemented. The idea of sets of predictive models is also used in re-sampling approaches such as “bagging” (Breiman, 1996) and “boosting” (Freund and Schapire, 1996). Bagging generates classifiers produced by different bootstrap (replicates) samples. A number of samples, with replacement, are generated and a classifier is built for each sample. A final classifier is built by producing an average prediction from all the classifiers. In boosting, the fundamental difference is that samples are not treated equally but appropriate weights are assigned to the points. This is critical because, unlike bagging, the classifiers are not built in parallel but rather sequentially. The results from one classifier affect the weights of the samples that are to be used in subsequent classifiers. The final set of classifiers will once again vote of the new data points. Recently, Ho (1998) analyzed the concept of “decision forests” based on the C4.5 classification algorithm, which is the classifier used in our studies as well. The idea of the approach is to generate sequences of classifiers by using random combinations of smaller dimension feature spaces. The actual number of tree that are built is chosen in an ad hoc manner.

All approaches that combine classifiers essentially produce a very similar overall discriminant function of the form:

$$g_c(x) = \frac{1}{n_t} \sum_{j=1}^{n_t} P(c|v_j(x)) \quad (4)$$

In other words, we combine the assignment probabilities to any class c for a given feature vector x for of all classifiers v_j , $P(c|v_j(x))$, in order to obtain the estimate for the probability of sample x belonging to class c .

Khan et al. (2001) provided a training set with 63 samples to be used for the training of the classifier, and a set of 20 samples to be used for testing. The 10 sets derived were used to develop a predictive decision “forest”. The ensemble of classification trees predicts the correct assignment for all the samples. The class assignment probabilities are depicted in Fig. 5. Our predictions are a significant improvement of the ones presented by Li and Wang (2002) where a model based on linear discriminant analysis was presented. Only three samples (15, 18, and 20) are not overwhelmingly predicted, however, majority vote identifies the correct class.

4.3. Characterizing robustness in the presence of noise

Noise in gene expression experiments in an unavoidable reality. It is significant to estimate the impact of noise in order to evaluate the biological implication of the results as well as the stability characteristics of the computational approach. It should be pointed out that the impact of noise on the analysis of arrays has not received, yet, significant attention despite its tremendous importance.

Estimating and quantifying the inherent noise in expression experiments is still an open issue. A number of contributors impact the measured values:

	EWS	BL	NB	RMS
1	0.20	0.00	<u>0.70</u>	0.10
2	<u>0.90</u>	0.10	0.00	0.00
3	0.10	0.00	0.00	<u>0.90</u>
4	<u>1.00</u>	0.00	0.00	0.00
5	0.00	<u>0.85</u>	0.15	0.00
6	0.00	0.00	<u>1.00</u>	0.00
7	0.15	0.00	0.00	<u>0.85</u>
8	<u>0.90</u>	0.00	0.00	0.10
9	0.00	0.10	<u>0.90</u>	0.00
10	<u>0.80</u>	0.00	0.15	0.05
11	0.00	0.00	<u>0.90</u>	0.10
12	0.10	0.00	0.00	<u>0.90</u>
13	0.05	<u>0.90</u>	0.05	0.00
14	<u>1.00</u>	0.00	0.00	0.00
15	<u>0.45</u>	0.15	0.05	0.35
16	<u>1.00</u>	0.00	0.00	0.00
17	0.00	0.00	0.00	<u>1.00</u>
18	0.30	0.15	<u>0.35</u>	0.20
19	0.00	0.00	0.00	<u>1.00</u>
20	0.25	0.20	<u>0.45</u>	0.10

Fig. 5. Prediction on 20 unseen SRBCT samples.

1. Cells from the same organ but different donors are used in order to study a particular cancer type, for example. However, donors do not share similar genetic background or environmental history, and therefore, variability within the samples is implicitly induced.
2. DNA extraction and hybridization cannot be identically replicated hence contributing to variability in the measured expression values.
3. Expression levels are comparative measurements between a target and a sample. Alterations in fluorescent intensity are used to differentiate the expression levels. These images are then digitized and the values normalized to allow comparisons (Schadt, Cheng, Byron, & Wong, 2001; Yang, Haddar, Tomas, Aksaker, & Papoutsakis, 2003). Therefore, the actual measurement is an interpretation rather than hard data.

Predicting, and modeling, noise in these experiments is a very difficult task. Tu, Stolovitzky, and Klein (2002) proposed a model in which hybridization noise is proportional to the expression level. Lee, Kuo, Whitmore, and Sklar (2000) present a mixture model for representing the observed expression level.

Unlike approaches that tend to eliminate the noise in the data, we would like to explore the hypothesis that informative gene expression patterns survive in the presence of noise. The purpose of our analysis is to produce estimates of the stability and robustness of the maximally informative genes. The assumption that we make is that the error is expressed in aggregate manner. That is, all the three components mentioned earlier can affect the reported value, however, we operate of the aggregate measurement. Furthermore, we assume that

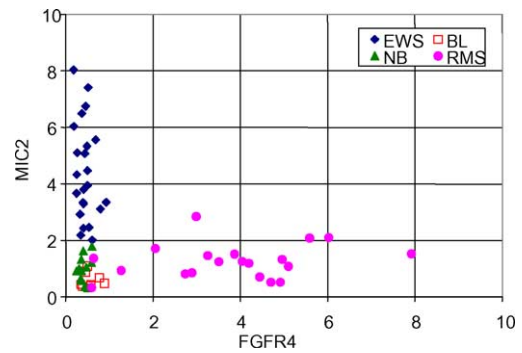


Fig. 6. Distribution of MIC2 and FGFR4 expression values in SRBCT samples.

there is no uncertainty in the class assignment. That is we have enough corroborating evidence to assign samples to classes (benign, malignant, etc.). However, the various errors affect the values, are recorded in terms of expression levels. Given the above assumptions, we performed a number of simulations and selections following a number of strategies. In order to simulate “noise” in the data, we assume that the expression values follow distributions on a per gene, per class basis. We analyze the available data, and we assume that

$$f(j|s \in C) \approx N(\mu^{\text{exp}}(j|s \in C), \sigma^{\text{exp}}(j|s \in C))$$

or

$$f(j|s \in C) \approx U(\min(j|s \in C), \max(j|s \in C))$$

It has to be emphasized once again that our goal is not to develop noise models but rather to determine the robustness of the emerging patterns of informative genes in the presence of noise. In our studies we performed the experiments in two different modes:

- We assume that the data are corrupted and we re-optimize the gene selection process;
- We assume that the classification rules are given and we test them on noise-corrupted data.

In either case, we conclude that frequently emerging solutions are by far more robust in the presence of noise. When training for a new classifier using noise-corrupted data, the emerging pattern (FGFR4, MIC2) survives as a stable component of the solution vector. When deriving the rules, using the emerging pattern (FGFR4, MIC2) as a component of the solution, while training with the original data and subsequently testing the classifier with noise-corrupted data, the performance of the classifier remains good. Fig. 6 demonstrates geometrically why the combination of FGFR4 and MIC2 is robust in terms of its ability to classify the SRBCT data. These two genes are significantly discriminating between two of the classes, as seen from the geometry of the classification rules, which makes them survive even increased levels of noise.

An open issue in discrimination analysis is the lack of a systematic way for exploring the geometry of the classification rules in order to make estimates regarding the

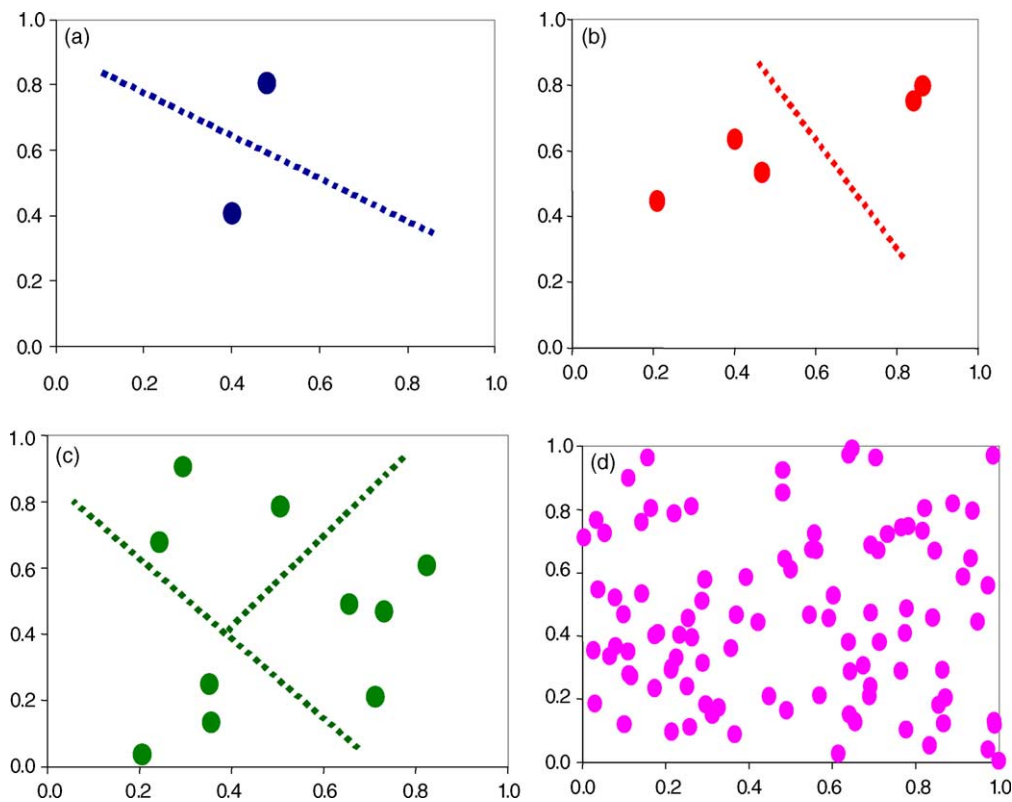


Fig. 7. Effect of small sample size.

robustness of the rules. Ho (2002) discusses the significance of understanding the geometry of separating boundaries among classes. We are currently in the process of addressing issues related to the geometric structures and relating those to an information content for various rules. We are rigorously exploring ideas such as length of class boundaries and space covering by ε -neighborhoods (Ho, 2000) to provide rigorous justification of our results.

4.4. Optimal versus random selection

The analysis of microarray data is complicated by the fact that, in general, the experiment is such that we have a very large number of observables, i.e., gene expression values, but a relatively small number of experiments. In fact, the problem of “small sample size” is a fundamental problem in pattern recognition and classification literature (Trunk, 1979). The problem was revisited a number of times and more recently Jain and Zongker (1997) restated the fact that the quality of the set of selected features is strongly affected by the size of the samples that is available. In the context of analyzing microarray samples the problem was also recently addressed by Pan, Lin, and Le (2002) and Hwang et al. (2002), among others.

Lets us illustrate the problem of small sample size with a simple example in two variables. Different realizations of the “system” were generated and are depicted in Fig. 7. Realizations differ in the number of observations that are recorded

each time. Let us consider the various cases in greater detail. By observing cases (a)–(c), one could argue that a structure begins to emerge. However, in (d), we realize that the true underlying generating process is one that generates realizations following a uniform distribution, hence, there is not emerging coherent structure. Looking at small, hence biased, sub-samples could not have revealed this fact. The implication is that for a sufficiently small sample size even irrelevant features can, by chance, lead to apparently accurate interpretations of the experimental data. Approaches that determine the minimum required size have been recently proposed (Hwang, Schmitt, Stephanopoulos, & Stephanopoulos, 2002). However, some knowledge of the distribution of data points is required in order to guarantee that homogeneous sampling of the entire population is achieved so that samples are not biased. Defining arbitrary diverse, or focused, sets of samples is not a trivial matter (Agrafiotis, 1997).

These issues cast a shadow of doubt on methods used for the analysis of microarray data. The main criticism is usually that even a trivial random selection of features (genes) should be able to provide very reasonable results given the extremely low values of the ratio samples over genes to choose from. In principle, the relative size of the training sample to feature space dimensionality is one of the key problems associated with the complexity of a discrimination problem (Ho, 2002). At this point, we will address this issue based on our current computational experience (we are currently exploring fundamentally analyses of the approach). The simplest

cases to be analyzed are the ones where a single gene was able to perfectly classify the arrays samples. In Androulakis and Hatzimanikatis (2004), we analyze a number of such cases and demonstrate computationally that random selections are not very efficient even for simple 1-rule problems. However, the probability of randomly selecting a set that perfectly classifies a more complicated problem, such as the BRCT one, is practically zero. Therefore, our experimental evidence suggests that systematic methodologies are required in order to appropriately address the question of identifying truly informative gene.

5. Conclusions and future work

Analysis of genome-wide expression profiles is currently revolutionizing modern biology. A number of computational approaches have recently been proposed for the analysis of volumes of expression data in an attempt to develop biologically interpretable models. We identified a number of limitations of methods that attempt to derive a single comprehensive model for analyzing genomic data. Therefore, we advocate the generation of multiple hypotheses for interpreting array data. We proposed a framework that combines optimization and classification algorithms for extracting ensembles of sets of maximally informative features. The fundamental assumption of our approach is that the minimal set of informative features achieves the best classification in the simplest way. We proposed the use of decision trees for deriving robust classifiers and offer ways for modeling and controlling their complexity. We demonstrated that robust emerging patterns of gene expression have a higher probability of generating biologically relevant interpretations. Our approach provides the domain expert with very critical information relating to multiple sets of optimal features. The availability of sets of solutions to choose from enhances the analysis capabilities and expands the scope of array experiments. Preliminary discussion was devoted to the issue of robustness of the classification rules and the impact of experimental noise. Under investigation is a more complete framework that will make the class assignment process part of the overall framework. In our upcoming work, we will demonstrate how the feature selection problem should be recast to properly account for the robustness of the solution and not just its prediction accuracy, how comprehensive complexity measures can be derived and integrated within our framework. Finally under investigation are approaches that will improve the computational efficiency of the algorithm by developing a generalized and integrated optimization framework.

Acknowledgments

The author would like to acknowledge critical input and suggestions provided by V. Hatzimanikatis (Northwestern University).

References

- Aarts, E. H., Korst, J., & van Laarhoven, P. J. (1997). Simulated annealing. In E. H. Aarts & J. K. Lenstra (Eds.), *Local search in combinatorial optimization*. John Wiley and Sons.
- Agrafiotis, D. (1997). Stochastic algorithms for maximizing molecular diversity. *Journal of Chemical Information and Computer Sciences*, 37, 841–851.
- Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., et al. (2000). Different types of diffuse b-cell lymphoma identified by gene expression profiling. *Nature*, 403, 503–511.
- Allander, S. V., Nupponen, N. N., Ringner, M., Hostetter, G., Maher, G. W., Goldberger, N., et al. (2001). Gastrointestinal stromal tumors with KIT mutations exhibit a remarkably homogeneous gene expression profile. *Cancer Research*, 61, 8624.
- Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., et al. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide analysis. *Proceedings of the National Academy of Sciences*, 96, 6745–6752.
- Androulakis, I. P., & Hatzimanikatis, V. (2004). Informative gene selection from gene expression experiments [in preparation].
- Basset, D. E., Eisen, M. B., & Boguski, M. S. (1999). Gene expression informatics—It's all in your mine. *Nature Genetics*, 21, 51–55.
- Bowtell, D. D. L. (1999). Options available—From start to finish—For obtaining expression data by microarray. *Nature Genetics*, 21, 25–32.
- Brown, P. P., & Botstein, D. (1999). Exploring the new world of the genome with DNA microarrays. *Nature Genetics*, 21, 33–37.
- Bittner, M., Meltzer, P., Chen, Y., Jiang, Y., Seftor, E., Hendrix, M., et al. (2001). Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, 406, 536.
- Bradley, P., & Masagarian, O. (1998). Feature selection via concave minimization and support vector machines. In *Proceedings of the 13th international conference on machine learning* (pp. 820–890).
- Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. New York: Chapman & Hall (Wadsworth Inc.).
- Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24, 123–140.
- Burnham, K. P., & Anderson, D. R. (1998). *Model selection and inference*. Springer.
- Cheung, V. G., Morley, M., Aguilar, F., Massimi, A., Kucherlapati, R., & Childs, C. (1999). Making and reading microarrays. *Nature Genetics*, 21, 15–19.
- Deutsch, J. M. (2003). Evolutionary algorithms for finding optimal gene set microarray prediction. *Bionformatics*, 19(1), 45–52.
- Dickson, C., Spencer-Dene, B., Dillon, C., & Fantl, V. (2000). Tyrosine kinase signalling in breast cancer: Fibroblast growth factors and their receptors. *Breast Cancer Research*, 2, 191–196.
- Dietterch, T. G. (2000). Ensemble methods in machine learning. In F. Roli (Ed.), *Proceedings of the first international workshop on multiple classifier systems, Lecture notes in computer science*. New York: Springer Verlag.
- Dudoit, A., Wang, Y. H., Callow, M. J., & Speed, T. P. (2000). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Technical report 578. Stanford University.
- Fayyad, U. M., & Irani, K. B. (1990). What should be minimized in a decision tree? In *Proceedings of the eight national conference on artificial intelligence* (pp. 749–754).
- Floudas, D. C. (2000). *Deterministic global optimization: Theory algorithms and applications*. Kluwer Academic Publishers.
- Freund, Y., & Schapire, R. (1996). Experiments with a new boosting algorithm. In *Proceedings of the 13th international conference on machine learning* (pp. 148–156).
- Golub, T. R., Slomin, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., et al. (1999). Molecular classification of cancer: Class

- discovery and class prediction by gene expression monitoring. *Science*, 286, 531–537.
- Hatzimanikatis, V., & Lee, K. H. (1999). Dynamical analysis of gene networks requires both mRNA and protein expression information. *Metabolic Engineering*, 1, E1–E7.
- Ho, T. K. (1998). C4.5 decision forests. In *Proceedings of the 14th international conference on pattern recognition* (pp. 545–549).
- Ho, T. K. (2000). Complexity of classification problems and comparative advantages of combined classifiers. In *Proceedings of the 1st international workshop on multiple classifier systems* (pp. 21–23).
- Ho, T. K. (2002). A data complexity analysis of comparative advantages of decision forests constructors. *Pattern Analysis and Applications*, 5, 102–112.
- Holte, R. C. (1993). Very simple classification rules perform well on most commonly used databases. *Machine Learning*, 11, 63–91.
- Hwang, A., Schmitt, W. A., Stephanopoulos, G., & Stephanopoulos, G. (2002). Determination of minimum sample size and discriminatory expression patterns in microarray data. *Bioinformatics*, 18(9), 1184–1193.
- Jain, A., & Zongker, D. (1997). Feature selection: Evaluation, application and small sample size performance. *IEEE Transactions on Pattern Analysis and Machine Learning*, 19, 153–158.
- Kafatos, F. C. (2002). A revolutionary landscape: The restructuring of biology and its convergence with medicine. *Journal of Molecular Biology*, 319, 861–867.
- Khan, J., Wei, J. S., Ringer, M., Saal, L. H., Landanyi, M., Westerman, F., et al. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*, 7, 673.
- Kohavi, R., & John, J. G. (1995). Wrappers for feature subset selection. Technical report. Computer Science Department, Stanford University.
- Kovar, H., Dworzak, M., Strehl, S., Schnell, E., Ambros, I. M., Ambros, P. F., et al. (1990). Overexpression of the pseudoautosomal gene MIC2 in Ewing's sarcoma and peripheral primitive neuroectodermal tumor. *Oncogene*, 5(7), 1067–1070.
- Lee, M.-L. T., Kuo, F. C., Whitmore, G. A., & Sklar, J. (2000). Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations. *Proceedings of the National Academy of Sciences*, 97, 9834–9839.
- Li, W., & Wang, Y. (2002). How many genes are needed for a discriminant microarray data analysis. In S. M. Lin & K. F. Johnson (Eds.), *Methods of microarray data analysis* (pp. 137–150). Kluwer Academic Publishers.
- Lipshutz, R. L., Fodor, S. P. A., Gingeras, T. R., & Lockhart, D. L. (1999). High density oligonucleotide arrays. *Nature Genetics*, 21, 20–24.
- Liu, H., & Motoda, H. (2000). *Feature selection for knowledge discovery and data mining*. Kluwer Academic Publishers.
- Luo, J., Duggan, D. J., Chen, Y., Sauvageot, J., Ewing, C. M., Bittner, M. L., et al. (2001). Human prostate cancer and benign prostatic hyperplasia: Molecular dissection by gene expression profiling. *Cancer Research*, 61, 4683.
- Narendra, P. N., & Fukunaga, K. (1977). A branch and bound algorithm for feature subset selection. *IEEE Transactions of Computers*, C-26(9), 917.
- Pan, W., Lin, J., & Le, C. T. (2002). How many replicates of arrays are required to detect gene expression changes n microarray experiments? A mixture model approach. *Genome Biology*, 3(5), 1–10.
- Perou, C. M. S. S., Jeffrey, M., van de Rijn, C. A., Rees, M. B., Eisen, D. T., Ross, A., et al. (1999). Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proceedings of the National Academy of Sciences*, 96, 9212–9217.
- Pollack, J. R., Perou, C. M., Alizadeh, A. A., Eisen, M. B., Pergamenschikov, A., Willimans, C. F., et al. (1999). Genome-wide analysis of DNA copy-number changes using cDNA microarray. *Nature Genetics*, 23, 41–46.
- Quinlan, R. J. (1993). *C4.5: Programs for machine learning*. Morgan Kaufmann Publishers.
- Ross, D. T., Scherf, U., Eisen, M. B., Perou, C. M., Spellman, P., Iyer, V., et al. (2000). Systematic variation in gene expression pattern in human cancer cells. *Nature Genetics*, 24, 227–234.
- Schadt, E. E., Cheng, L., Byron, E., & Wong, W. H. (2001). Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. *Journal of Cellular Biochemistry*, 37, 120–125.
- Schena, M., Shalon, D., Davis, R., & Brown, P. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270, 467.
- Trunk, G. V. (1979). A problem of dimensionality: A simple example. *IEEE Transactions on Pattern Analysis and Machine Learning*, 1(3), 306–307.
- Tu, Y., Stolovitzky, G., & Klein, U. (2002). Quantitative noise analysis for gene expression microarray experiments. *Proceedings of the National Academy of Sciences*, 99(22), 14031–14036.
- Yang, H., Haddar, H. H., Tomas, C., Aksaker, K., & Papoutsakis, E. T. (2003). A segmental nearest neighbor normalization and gene identification method gives superior results for DNA-array analysis. *Proceedings of the National Academy of Sciences*, 100(3), 1122–1127.