Physiol Genomics
4: 109–126, 2000.

Analysis of molecular profile data using generative and discriminative methods

E. J. MOLER,* M. L. CHOW,* AND I. S. MIAN
Department of Cell and Molecular Biology, Radiation Biology and Environmental Toxicology Group, Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720

Received 20 December 1999; accepted in final form 4 September 2000

Moler, E. J., M. L. Chow, and I. S. Mian. Analysis of molecular profile data using generative and discriminative methods. *Physiol Genomics* 4: 109–126, 2000.—A modular framework is proposed for modeling and understanding the relationships between molecular profile data and other domain knowledge using a combination of generative (here, graphical models) and discriminative [Support Vector Machines (SVMs)] methods. As illustration, naive Bayes models, simple graphical models, and SVMs were applied to published transcription profile data for 1,988 genes in 62 colon adenocarcinoma tissue specimens labeled as tumor or nontumor. These unsupervised and supervised learning methods identified three classes or subtypes of specimens, assigned tumor or nontumor labels to new specimens and detected six potentially mislabeled specimens. The probability parameters of the three classes were utilized to develop a novel gene relevance, ranking, and selection method. SVMs trained to discriminate nontumor from tumor specimens using only the 50–200 top-ranked genes had the same or better generalization performance than the full repertoire of 1,988 genes. Approximately 90 marker genes were pinpointed for use in understanding the basic biology of colon adenocarcinoma, defining targets for therapeutic intervention and developing diagnostic tools. These potential markers highlight the importance of tissue biology in the etiology of cancer. Comparative analysis of molecular profile data is proposed as a mechanism for predicting the physiological function of genes in instances when comparative sequence analysis proves uninformative, such as with human and yeast translationally controlled tumour protein. Graphical models and SVMs hold promise as the foundations for developing decision support systems for diagnosis, prognosis, and monitoring as well as inferring biological networks.

microarrays; biological networks; graphical models; support vector machines; decision support systems; comparative molecular profile data analysis

PROFILING TECHNIQUES such as DNA microarrays, two-dimensional gel electrophoresis, capillary electrophoresis, and mass spectroscopy provide information on genes, proteins, metabolites, and other molecules (features) under defined conditions (for recent reviews see Refs. 11 and 43). Despite differences in how and which molecule is assayed, the problem can be generalized to one of analysis of a molecular profile matrix. Each row, a molecule profile vector, is the behavior of the same molecule under different conditions. Each column, an experiment profile vector, is the behavior of different molecules in the same experiment. Experiment profile vectors may be assigned class labels that reflect the source of the biological sample, for example, tumor or nontumor specimen. Molecule profile vectors may be labeled with information pertinent to the molecule of interest such as the presence or absence of a biochemical activity, oil-water partition coefficient, specific protein domain, and transcription factor binding site.

Currently, the most prevalent molecular profile matrices are those from transcription profiling studies in which the molecules are genes. For convenience, each functionally defined nucleic acid sequence whose expression level is monitored will be termed a "gene," irrespective of whether it is actually a gene, an expressed sequence tag, or DNA from another source. Although still in their infancy, computational methods have proved adept at extracting experimentally and clinically useful information from transcription profile data. These techniques include hierarchical clustering (10), gene shaving (19), self-organizing maps (16, 49, 51), *k*-means clustering (50), Boolean networks (26, 30, 45), linear modeling (9), principal component analysis (42), nonlinear modeling (53), Bayesian networks (BNs) (14), dynamic Bayesian networks (DBNs) (36), Support Vector Machines (SVMs) (5), and Petri nets (17, 32).

This work proposes a modular framework for the analysis of molecular profile data and domain knowledge that combines generative and discriminative methods. Initially, the framework is designed to address the distinct, yet complementary tasks of elucidating basic biological mechanisms and pathways and developing decision support systems for diagnosis, prognosis, and monitoring. The long-term goal is creation of an object-oriented system for prediction, inference, and experimental planning in which local relations (network fragments) can be integrated to build models that exhibit greater complexity. Here, specific generative and discriminative methods are employed. Graphical models were selected because of their structured stochastic nature and concomitant ability to

Article published online before print. See web site for date of publication (http://physiolgenomics.physiology.org).

*E. J. Moler and M. L. Chow contributed equally to this work.

Address for reprint requests and correspondence: I. S. Mian, Dept. of Cell and Mol. Biol., MS 74-197, Radiation Biology and Environ. Toxicol. Group, Life Sci. Div., Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720 (E-mail: SMian@lbl.gov).

model complex relations. SVMs were chosen because of their predictive performance capabilities when applied to classification, prediction, and regression problems. These general techniques permit creation of increasingly sophisticated models and analytical methods capable of yielding useful predictions and insights during each phase of framework development. Such models have good predictive accuracy (generalization) and lend themselves to human interpretation (explanation). They can handle missing data and/or "noisy" data arising from the stochastic nature of the underlying biological process (model noise) and errors occuring during sample preparation and/or measurement (observation noise). The models can incorporate prior knowledge, model hierarchical relationships, and utilize heterogenous data.

Here, working prototypes of tools for modules that address three statistical tasks associated with analysis of profile data are described. They are applied to published 1,988-feature experiment profile vectors from 62 human colon adenocarcinoma specimens labeled as tumor or nontumor (2). A naive Bayes model, a simple graphical model, is used to discover and characterize classes of experiment profile vectors (unsupervised learning). SVMs are employed to distinguish tumor from nontumor specimens and to assign the label of profile vectors not used for training (supervised learning). Two feature relevance experts are utilized to identify marker genes, genes that distinguish the two types of specimens (feature relevance, ranking, and selection). Insights into colon adenocarcinoma biology and future directions for the methodology are discussed.

MODELS

Learning Models from Profile Data

To make descriptions of the techniques more concrete, the profile data (2) to be analyzed will be described first. Human colon adenocarcinoma specimens were collected from 40 patients. For 22 of these patients, additional colon tissue specimens from "normal" regions were obtained. Each RNA sample obtained from these 62 specimens was hybridized to an Affymetrix oligonucleotide array complementary to more than 6,500 human genes and expressed sequence tags. Only genes with the highest minimal intensity across the samples were chosen for further study and made publicly available. Specimens labeled "normal" will be referred to as "nontumor" to differentiate them from specimens that would have come from individuals with no record of adenocarcinoma. The molecular profile matrix consists of sixty-two 1,988-feature experiment profile vectors labeled tumor or nontumor.

The aforementioned data can be represented as N {input, output} pairs or $\{(\mathbf{X}_L^n, d^n)\}_{n=1}^N$, where N is the number of profile vectors (here $N = 62$); $\mathbf{X}_L^n = [x_1^n, \ldots, x_L^n]$ is an L-feature profile vector ($L = 1988$); x_l^n is the "expression level" of gene l in profile vector n; and $d^n \in \{A, \ldots\}$ is a label that can take on a value such as A ($d^n \in$ {tumor, nontumor}). Here, x_l^n refers to the published value, but it could represent the result of any given transformation of "raw measurements." Although not the focus of this work, preprocessing image data and other data to arrive at meaningful x_l^n values is an essential component of reducing errors in downstream analyses of the type described here.

Learning predictive models from training data fall into two general headings. Unsupervised learning finds "natural groupings" using only the input variables. Here, this translates to identifying classes of experiment profile vectors by clustering the N L-feature input vectors $[\mathbf{X}_{\mathbf{1988}}^{\mathbf{1}}, \ldots, \mathbf{X}_{1988}^{62}]$. Supervised learning estimates a function from paired values of input and output variables with the aim of predicting the outputs for future, unseen input variables. This maps to utilising $(\mathbf{X}_{1988}^n, d^n)$ pairs to learn a model that can assign the output label tumor or nontumor to a new profile vector. Labeled input vectors are separated into positive and negative training examples. Here, tumor (nontumor) samples are considered to be positive (negative) training examples.

The generalization performance of a learning system is a measure of how well it performs on data not used for training. For a supervised learning system, labeled training examples are partitioned into two disjoint sets. The estimation set of positive and negative training examples is used to determine the parameters of the model and the test set to assess its performance. The label assigned by a trained model to a test example can be a true positive (known positive example, positive label), true negative (negative example, negative label), false positive (negative example, positive label), or false negative (positive example, negative label). Since the number of available training examples is limited (here $N = 62$), a "leave-one-out cross-validation" strategy is employed. A model estimated using $N - 1$ training examples is evaluated using the single test example. This procedure is repeated for each example in turn. The total number of models that make true positive, true negative, false positive, and false negative assignments is determined. Here, the generalization performance is defined the sum of the true positive and true negative assignments (the maximum possible generalization performance is N).

Graphical Models

Graphical models can be viewed as highly structured stochastic systems that provide a compact, intuitive, and probabilistic framework capable of learning complex relations between variables such as genes and other molecular or environmental factors. A BN is a graphical model annotated with conditional probabilities in which the graph is directed and contains no directed cycles (Fig. 1; for reviews see Refs. 23, 25, 40, and the introductory tutorial at **www.cs.berkeley.edu/~murphyk/Bayes/bayes.html**). Learning a model from data can be decomposed into the problem of learning the topology and/or the parameters. Many of the discrete time models proposed for reconstructing genetic networks from time series data are special cases of DBNs (36). The advantages of DBNs include

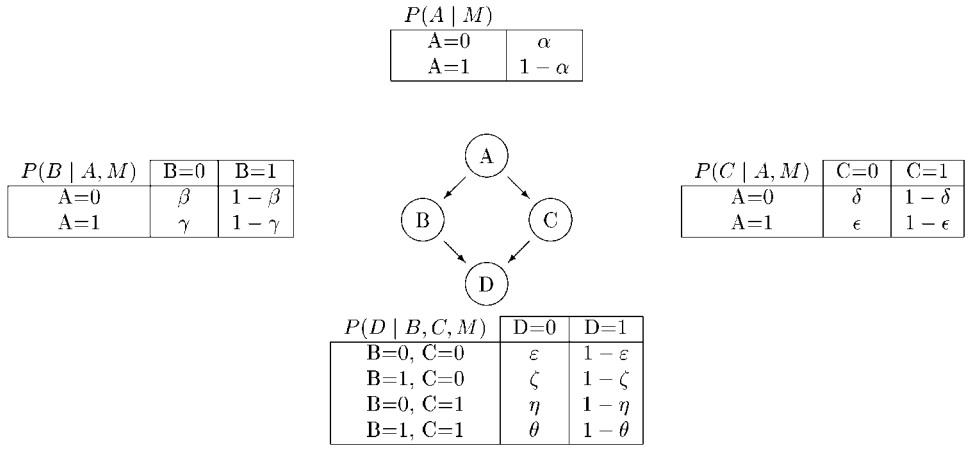

Fig. 1. The qualitative and quantitative aspects of graphical models illustrated using a simple model M consisting of four binary-valued variables A, B, C, and D. These variables could represent four genes in which the values indicate whether the gene is active (1) or inactive (0). The network topology is a directed acyclic graph in which nodes represent the variables of interest and influences between variables are encoded explicitly by the presence of edges between nodes; the edges have directionality and thus semantic meaning. The absence of an edge provides information about the independence between concepts: when two variables lack a connecting edge, nothing about the state of one variable can be inferred from the state of the other. The network parameters are the local probability models and conditional probability distributions ($0 \leq \alpha, \beta, \gamma, \delta, \epsilon, \varepsilon, \zeta, \eta, \theta \leq 1$). The joint distribution is the product of the individual node distributions, so, for the example shown, it is $P(A, B, C, D|M) = P(A|M)P(B|A, M)P(C|A, M)P(D|B, C, M)$. For example, the probability if all the genes are inactive is $P(A = 0, B = 0, C = 0, D = 0|M) = \alpha \times \beta \times \delta \times \varepsilon$.

the ability to model stochasticity, incorporate prior knowledge, and handle hidden variables and missing data in a principled way (36).

Friedman et al. (14) analyzed the expression levels of 800 yeast genes found previously to be regulated by the cell cycle (47). Their goal was to recover the structure of regulatory interactions between these 800 genes, a genetic network, by learning a BN from data tabula rasa, i.e., without any prior knowledge or assumptions. In contrast, the philosophy underlying the framework proposed here is that given the size and complexity of biological networks, reconstructing even fragments will necessitate incorporating prior knowledge into the model building process. Thus, clustering profile vectors using a simple graphical model known as a naive Bayes model is treated as an initial step toward learning "biologically plausible" (D)BNs. This procedure can identify classes of coexpressed genes and thus families of genes that are likely to be regulated by common factors. If known (unknown), these factors can be represented as observed (hidden) variables that influence coexpressed genes. Together with domain knowledge, this type of information provides important contraints on the space of (initial) network topologies and parameters that need to be explored during model learning. These high-level network fragment objects could then be assembled into larger biological networks within the same graphical model formalism as that used for clustering.

A naive Bayes model. In a naive Bayes model, a single unobserved variable is assumed to "generate" the observed data (here, sixty-two 1,988-feature experiment profile vectors). The hidden variable is discrete, and its possible states correspond to the underlying classes in the data. The data are produced by K models or data-generating mechanisms. These K models correspond to the K classes or clusters of biological interest (Fig. 2). A naive Bayes model can be viewed as a finite mixture model. If the functional form for the data-generating mechanism is a Gaussian, then the model is a Gaussian mixture model (Fig. 3).

The models have a number of attractive features that make them particularly well suited for unsupervised learning and exploratory analysis of profile vectors. These include a fixed topology so parameter estimation is the only learning problem to be solved, ease

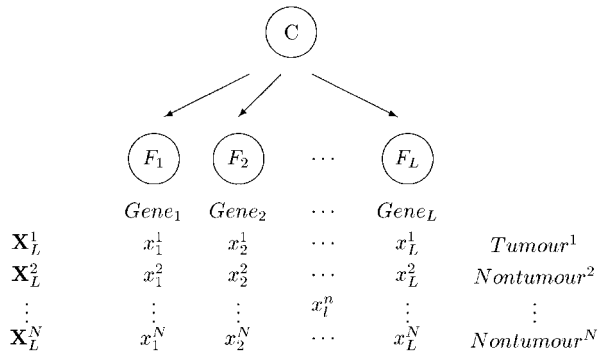

Fig. 2. A graphical model representation of a naive Bayes model M and its relationship to the labeled experiment profile vectors examined here. C is the hidden classification node that generates K alternative classes present in the observed $N = 62$, $L = 1{,}988$-feature experiment profile vectors (each row, $\mathbf{X}_L^n$, is a profile vector). Each F_l node represents the expression measurements for gene l. The network topology makes minimal assumptions about relationships in the data: the F_l nodes are conditionally independent given the class C. Thus, the joint distribution satisfies $P(F_1, \ldots, F_L, C|M) = P(C|M) \prod_{l=1}^{L} P(F_l|C, M)$. The "tumor" and "nontumor" labels are shown for illustration purposes only, since the model contains no explicit node for a variable "label."

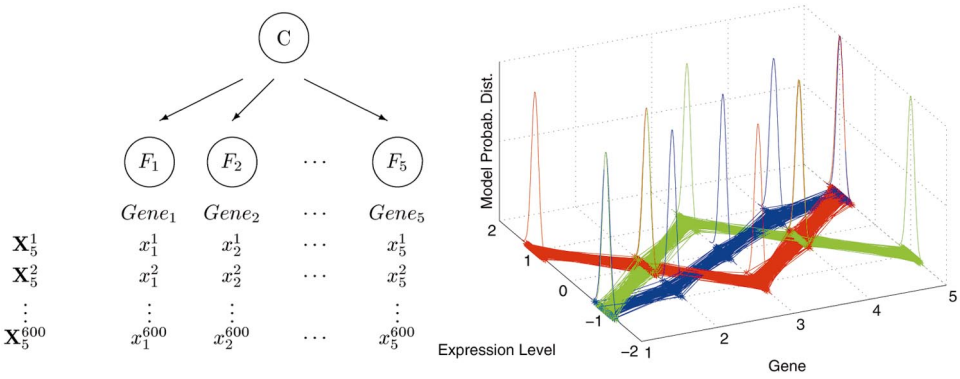

Fig. 3. A naive Bayes model for a toy data set in which 600 five-feature experiment profile vectors belonging to $K = 3$ classes (red, blue, green) are generated by Gaussian data-generating mechanisms. Note the network topology (*top left*), input data set (*bottom left*), and resultant model (*right,* the probability distributions are shown on the z-axis). Each of the 3×5 Gaussians (some overlap) has its own mean and standard deviation (probability parameters). For illustration purposes, the expression values in a given profile vector are connected by straight lines.

of implementation, speed, and ability to scale well as the size of the data increase. Modifying the variance characteristics of each component data-generating mechanism allows direct control over the variability permitted within each class. The question of how many classes K the data suggest can be treated in an objective manner. This model-based approach to clustering can handle missing data, noisy data, and uncertainty about class membership in a probabilistic manner. In the AutoClass implementation (7) of naive Bayes models (see below, *Naive Bayes Models: AutoClass*), a Bayesian approach is employed to derive the maximum posterior probability classification and the optimum number of classes K.

Given labeled input vectors, an "unsupervised naive Bayes model" refers to a model in which both the number of classes K and the $K \times L$ sets of probability parameters are estimated from unlabeled profile vectors. A "supervised naive Bayes model" refers to a model in which the number of classes is fixed a priori, and the probability parameters are calculated directly from the values of features assigned to classes. Here, the unsupervised naive Bayes model is one trained to discover and characterize the K classes present in the 62 unlabeled 1,988-feature experiment profile vectors. A supervised naive Bayes model is one estimated by first partitioning the profile vectors according to their tumor or nontumor label. The $K = 2 \times 1{,}988$ sets of probability parameters are computed directly from the 40 (or 22) expression levels of the 1,988 genes in the tumor (or nontumor) samples.

Support Vector Machines

In the context of pattern classification, an SVM constructs a hyperplane as the decision surface such that the margin of separation between positive and negative training examples is maximized (Fig. 4; for review, see Ref. 52 and the introductory tutorials at **www.kernel-machines.org**). This is achieved via an approximate implementation of the method of structural risk minimization, a principled approach rooted in statistical learning theory. This induction principle is based on the fact that the error rate of a learning machine on test examples (the generalization error rate) is bounded by the sum of the training-error rate and a term that depends on the Vapnik-Chervonenkis (VC) dimension. For separable data, an SVM produces a value of zero for the first term and minimizes the second VC term. Thus, SVMs generalize well when applied to pattern recognition problems. Compared to other machine learning algorithms, SVMs provide flexibility in choosing a similarity function, sparseness of solution when dealing with large data sets, the ability to handle large feature spaces, and the capacity to identify outliers.

A central notion in SVMs is the inner-product kernel $K(\mathbf{X}^i, \mathbf{X}^j)$, a measure of similarity between two input vectors $\mathbf{X}^i$ and $\mathbf{X}^j$. Depending on how the kernel is defined, different learning machines characterized by

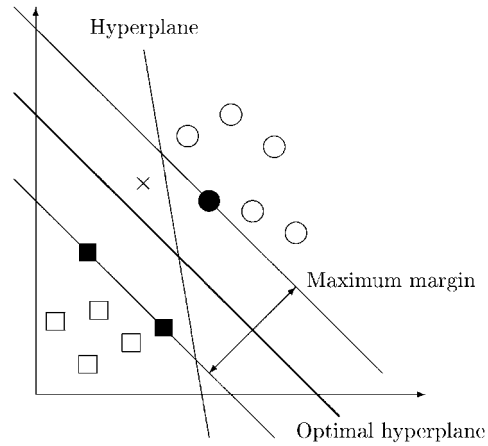

Fig. 4. The notion of separating hyperplanes illustrated using a classification problem in a two-dimensional input space. Given the 12 positive and negative training examples labeled as circles and squares, the task is to assign the output or class label for the input vector $\times$ of unknown label. Based on the known examples, the new point is most likely to be a circle. The size of the margin affects the generalization capacity of a hyperplane, its ability to assign correctly the label of a new data point. The "Hyperplane" would classify incorrectly the new point as a square. The "Optimal hyperplane" minimizes the generalization error because it separates the training data with a maximal margin. Only for the optimal hyperplane, therefore, does $\times$ lie inside the region of separation between the two classes *and* fall on the correct side of the decision surface. The filled data points at the margins are termed support vectors because they define the optimal hyperplane. The input data shown are linearly separable. If this were not the case, then mapping them into a higher-dimensional feature space would allow the hyperplane to be defined there as opposed to the input space occupied by the training data. This mapping can be achieved via use of a kernel function that encodes a similarity measure between two input vectors.

nonlinear decision surfaces can be constructed. The commonly used kernel employed here is a radial basis function, $K(\mathbf{X}^i, \mathbf{X}^j) = \exp(-\|\mathbf{X}^i - \mathbf{X}^j\|^2/2\sigma^2)$, where the width $\gamma = 1/2\sigma^2$ is a user-defined parameter. Another often-used kernel is a dot product, $K(\mathbf{X}_L^i, \mathbf{X}_L^j) = \sum_{l=1}^{L} x_l^i x_l^j$.

A MODULAR FRAMEWORK FOR ANALYSIS OF MOLECULAR PROFILE AND OTHER DATA USING GENERATIVE AND DISCRIMINATIVE METHODS

This section outlines tasks that are relevant to modules of the framework. The particular generative (graphical models, naive Bayes models) and discriminative (SVM) methods employed for unsupervised and supervised learning plus other specific algorithms are designed to be illustrative rather than comprehensive. Clearly, other learning systems, algorithms and approaches for defining and studying network fragments could be utilized. Given the limited expressive capability of any individual modeling method, it will be necessary to incorporate different models and model types to gain a comprehensive understanding of the myriad of processes that generate molecular profile data. The RESULTS present results from application of the methods explained in METHODS to the transcription profile data described in the subsection *Learning Models from Profile Data*.

Unsupervised Learning

Discovering and characterizing classes of profile vectors using an unsupervised learning method yields information on the fine structure of the data. Identifying classes of gene profile vectors can suggest genes whose products may have related functions as well as those that could be regulated by common transcription, environmental, or other factors. Clustering experiment profile vectors can indicate relationships between conditions and pathways. For example, if mutants with similar phenotypes fall into different classes, then the homeostatic mechanisms by which the biological endpoint is reached could differ. Alternatively, if mutants displaying seemingly unrelated physiological behaviors have similar profiles, then they may operate via common pathways.

For labeled profile vectors, an unsupervised learning method can be used to group the input vectors and the estimated classes compared to those that would be expected based on the output labels. The discrepancy between the number of classes estimated from the data and the number of known classes provides an indication of the homogeneity of the problem being addressed and/or quality of the data used. The estimated classes may correspond to subcategories of profile vectors with the same label, for example, tumor subtypes, or profile vectors with different labels. If an input vector with one label is assigned to a class that is dominated by examples from another class, then this could suggest mislabeled examples.

Here, naive Bayes models are utilized to cluster sixty-two 1,988-feature experiment profile vectors. Elsewhere, they have been used to cluster seventy-two 7,070-feature experiment (8) and 5,687 seventy-eight-feature gene (35) profile vectors.

Supervised Learning

Discriminating between profile vectors with different labels and assigning the label of a new profile vector using a supervised learning method is useful for a variety of tasks. For experiment profile vectors, this classification and prediction procedure can be a component of decision support systems for clinical and/or environmental diagnosis, prognosis, and monitoring. In cancer transcription profiling studies, for example, data from specimens that have different pathological characteristics or are subtypes of the same disorder can assist in developing systems for classifying and analyzing cancers from a molecular rather than morphological perspective. For gene profile vectors, this analytical approach can suggest potential biological roles for genes if the known labels correspond to biochemical, functional, or other physiological properties.

A trained model can assist in identifying input vectors that are most important in defining classes and pinpointing those that may have been mislabeled (outliers). In principle, an unsupervised method can be used to partition unlabeled input vectors into disjoint sets such that each class can be associated with an output label, which can then be employed subsequently by a supervised learning system.

Here, SVMs are used for supervised learning problems involving the sixty-two 1,988-feature labeled experiment profile vectors. A naives Bayes model trained to address the same discrimination problem performed considerably less well. Elsewhere, SVMs have been applied to seventy-two 7,070-feature experiment profile vectors with two to four different labels (8). SVMs have been utilized to classify 2,467 seventy-nine-feature yeast gene profile vectors and to assign functional roles for uncharacterised open reading frames (5).

Feature Relevance, Ranking, and Selection: Feature Relevance Experts

In a supervised learning problem, features in the input vector vary as to how relevant they are to the discrimination problem. For molecule profile vector classes, the extent to which experiments differentiate classes can vary. Given a compendium of gene profile vectors derived from experiments examining a range of genotypes and conditions, feature relevance can define experiments that best separate genes belonging to the same biochemical and/or functional classes. For experiment profile vector classes, molecules can vary as to how well they distinguish classes. In cancer profiling studies, molecules that discriminate tumor from nontumor samples are good candidates for subsequent in-depth experimental studies as well as developing decision support systems. Highly informative features, marker genes or marker experiments, can be identified by reducing the cardinality of input vectors such that the generalization performance of a supervised learn-

ing system is undiminished compared to one trained using all the features.

Assume that profile vectors are assigned to $T \geq 2$ classes, either those estimated from unlabeled profile vectors or those designated on the basis of output labels. The "relevance of a feature F_l" is defined as how well it distinguishes class c_i from c_j. If the relevance is zero, then the behavior of the feature in the two classes is the same; larger values signify increasingly greater differences and thus a greater ability to distinguish classes. The absolute magnitude is augmented with a sign such that a negative (positive) value signifies that the value of the feature is lower (higher) in c_j than in c_i. "Multiclass relevance" denotes how well the feature distinguishes class c_i from all other $T - 1$ classes. "Global relevance" signifies how well the feature distinguishes all T classes. Ordering features based on their relevance value ranks them in terms of how well they distinguish two specific classes, whereas the global relevance ranks them with regards to how well they distinguish all T classes. Different numbers of features can be selected either in terms of an absolute number such as the m top-, middle-, or bottom-ranked features or those with values above (below) a specified threshold.

Markers can be identified in a systematic manner with the aid of a feature relevance expert. Such an expert *1*) implements an algorithm for computing feature relevance, *2*) reorders features according to this value, *3*) selects subsets of ranked features for use in training a supervised learning system, and *4*) identifies markers based on feature subsets that generalize well, namely, assign the labels for input vectors not used for training. Preferably, a relevance measure should generate a monotonic ordering. Reducing the number of features by eliminating bottom-ranked ones should improve the generalization performance of supervised learning systems trained using the feature subsets. There should be optimal subsets that maximize the performance. Finally, retaining fewer features should degrade the performance. If optimal subsets have the same or better generalization performance than the full repertoire, then these features are likely to be particularly useful markers. Different feature relevance experts can be evaluated by determining the generalization performance of supervised learning systems trained using the m-ranked features of each expert.

For profile vectors with multiple labels, feature relevance depends on the biological question being posed. For example, the relevance of a gene in differentiating pathological states may or may not be related to its ability to discriminate between tissue types. Consider experiment profile vectors with two sets of binary labels, "tumor/nontumor" and "liver/colon." The relevance of a gene for the tumor/nontumor problem requires comparing its expression values in profile vectors labeled (tumor,liver)/(tumor,colon) and (nontumor,liver)/(nontumor,colon). In contrast, the relevance of the same gene for the liver/colon problem requires comparing (liver,tumor)/(liver,nontumor) with (colon,tumor)/(colon,nontumor).

Here, specific algorithms for calculating the relevance and global relevance of a feature are described that are based on the probability parameters of naive Bayes model classes. A feature relevance expert based on an unsupervised naive Bayes model generalizes better than one employing a supervised naive Bayes model (see above, *A naive Bayes model*).

External Knowledge as an Aid to Interpretation

The time taken to explore complex relationships revealed by analysis of profile data can be reduced by a systematic environment that extracts, organizes, and integrates external knowledge into the interpretation procedure. Gene ontologies and controlled vocabularies (3, 6) are key components in the creation of such environments. Two-way unsupervised learning, discovering classes of experiment and gene profile vectors, integrated with a comprehensive knowledge base could highlight markers and correlations for further study. For example, if genes and experiments are cross-indexed to external information in a qualitative and quantitative manner, then it should be easier to uncover statistically and/or biologically significant associations between profile vector classes and, for example cell type, developmental stage, small molecule concentration, environmental condition, signaling pathway, and so on. Specific gene vector classes may be correlated with protein products having similar functions, noncoding regions, protein-protein interactions, and so on.

Elsewhere (35), the associations between 45 classes estimated from 5,687 seventy-eight-feature *Saccharomyces cerevisiae* gene profile vectors and four types of external knowledge were determined. The results were used to suggest potential functions and physiological roles for specific genes.

Decision Support Systems for Diagnosis, Prognosis, and Monitoring

A decision support system is a knowledge-based systems aimed at organizing relevant experimental and other data for the purpose of assisting users make decisions about real world problems. Experiment profile vectors from cancer transcription profiling studies can be used to distinguish between specimens of known (sub)type and to assign the label for new specimens. Since the consequences of misdiagnosis are potentially deleterious, the supervised learning method and training data underlying such a decision support system should maximize sensitivity and specificity. Not all the genes monitored in a profile study are required to assign the label for a specimen of unknown origin with a high degree of accuracy. Some genes may even decrease prediction accuracy. Hence, feature relevance, ranking, and selection is an important component of creating prototypes of clinically useful systems. For a given data set and fixed number of features, there are likely to be a number feature subsets of this size that

have similar generalization performances when used to address the same supervised learning problem. Thus, features ranked highly by a majority of, or all the experts in a mixture of feature relevance experts should be robust and reliable markers.

Here, a feature relevance expert is used to identify markers for colon adenocarcinoma. Elsewhere (8), each of the top 50 genes from three different feature relevance experts were shown to generalize as well as each other and the full repertoire of 7,070 genes. However, the specific genes in these subsets were not identical. Thus, genes at the union of these subsets (125 genes in total) were proposed as candidate for developing a prototype decision support system for distinguishing two subtypes of leukemia.

Networks for Experimental Design, Planning, and Inference

Inferring or verifying networks for use in diagnostic reasoning, causal reasoning, and assessing the effects of intervention will require fusing data and results from the other modules of the framework. For example, groupings of gene profile vectors and the identification of common noncoding regions will provide important constraints on learning the topology and/or parameters of a network. Identifying markers using feature relevance experts can pinpoint molecules that should be represented explicitly in efforts to infer networks from profile data using techniques such as graphical models.

METHODS

MATLAB (**www.mathworks.com**) was used for data analysis, visualization, and algorithm and application development. All computations were performed on a Sun Ultra 60 workstation.

Naive Bayes Models: AutoClass

In AutoClass C version 3.3 (7), the continuous F_l nodes are modeled using Gaussian probability density functions and the discrete classification node C using a Bernoulli distribution (Fig. 2). Training a model involves using profile vectors to estimate the number of classes K for node C and the probability parameters for each F_l node. Starting from random initial descriptions for a specified number of classes, a gradient descent search through the space of descriptors is performed. At each step of the model search procedure, the current descriptions are used to assign probabilistically each profile vector to each class. The observed values for each profile vector are used to update class descriptions, and the procedure is repeated until a specified convergence criterion is reached. The program iterates through different numbers of classes to determine the best taxonomy.

Overfitting, finding a model in which the number of classes K is equal to the number of profile vectors N, is ameliorated as follows. A variant of the expectation-maximization (EM) algorithm is used to search through model-space with the condition that each profile vector belong to some class (the sum of all class probabilities is one). A penalty is incurred for adding more classes. Increasing the number of classes decreases the prior probability of each class unless the additional class improves the likelihood of the data. The model-space that needs to be searched can be constrained by setting a lower bound on the variance of the data-generating mechanism. For each gene l, the level of observation noise (measurement error) and/or natural variation in expression between samples (patients) can be used to set this value in a data-dependent manner. Thousands of models are estimated, each starting from different random number seeds. Each resultant model, a locally optimum solution in the parameter space, is scored. These model marginals are compared to find the model that best describes the data.

The input data are the sixty-two 1,988-feature experiment profile vectors, $[\mathbf{X}^1_{1988}, \ldots, \mathbf{X}^{72}_{1988}]$ where $\mathbf{X}^n_{1988} = [x^n_1, \ldots, x^n_{1988}]$. The expression level of gene l in profile vector n, x^n_l, is used as is, i.e., the published data (2) are not rescaled, shifted, normalized, or modified. Since the measurement error and intrinsic variability are unknown, the minimum value of the standard deviation of the Gaussian for each class, $\sigma_{k,l}$, is set to 0.1 of the standard deviation of the Gaussian for the expression values across all N samples, $x^1_l, \ldots, x^N_l$. The output consists of *1*) K, the number of classes, *2*) an $N \times K$ likelihood matrix where each element is the likelihood of experiment profile vector n given class c_k, $P(\mathbf{X}^n_L|c_k, M)$, and *3*) a $K \times L$ parameter matrix where each element is the mean and standard deviation of the Gaussian modelling class c_k and gene l, $(\mu_{k,l}, \sigma_{k,l})$. For the data set here, the marginal for the best model is significantly higher than the other models. The final results do not depend on the order in which input vectors are entered into the model.

Support Vector Machines: SVM^{light}

SVM$^{\text{light}}$ version 3.02 (24) has a fast optimization algorithm, can handle many thousands of support vectors, can be trained using tens of thousands of training examples, and supports a variety of kernel functions. The input data are labeled profile vectors and a kernel function plus any of its associated parameters. Although there is no formal mechanism for selecting the most appropriate class of kernel function for a particular problem, empirical evidence suggests that a radial basis function is a reasonable choice. This kernel performed well when applied to biological classification problems arising from transcription profiling (5, 8) and protein fold recognition (M. L. Chow and I. S. Mian, unpublished information) studies. Based on the latter work and tests using the data examined here, the width of the radial basis function $\gamma = 1/2\sigma^2$ is set to 0.01. Elsewhere (5, 8), the value of γ is set in a data-dependent manner by choosing σ to be equal to the median of the Euclidean distances from each positive example to the nearest negative example (5). The output from the learning module is a binary classification model which can be used to assign the label for a test example.

SVMs are trained and evaluated using the leave-one-out cross-validation procedure described above (*Learning Models from Profile Data*). To account for unequal numbers of positive and negative training examples, each estimation set is balanced by duplicating as many randomly chosen examples as necessary from the smaller set to yield the same number of examples as the larger set. The generalization performance achieved is the total number of SVMs that make true positive and true negative assignments for their test example. A false positive or false negative assignment occurs when the test example falls on the wrong side of the decision boundary.

Feature Relevance: Naive Bayes (Global) Relevance

The relevance of a feature (see above, *Feature Relevance, Ranking, and Selection: Feature Relevance Experts*) measure proposed here is termed the naive Bayes relevance (NBR). It

is based on the probability of a profile vector class k given the observed value of the feature l, $P(c_{k,l}|x_l^n)$. Using Bayes rule and assuming that classes c_i and c_j are independent and equally likely a priori, the NBR is defined as

$$NBR_{ij}(F_l) \equiv \log\left[\frac{4}{N}\sum_{n=1}^{N} P(c_{i,l}|x_l^n)P(c_{j,l}|x_l^n)\right]$$

$$\approx \log\left[\frac{4}{N}\sum_{n=1}^{N} \frac{P(x_l^n|c_{i,l})P(x_l^n|c_{j,l})}{[P(x_l^n|c_{i,l}) + P(x_l^n|c_{j,l})]^2}\right]$$

The factor of 4 ensures the minimum and maximum values are 0.0 and 1.0 (rather than 0.0 and 0.25). As calculated above, the NBR is a fast and crude approximation of the joint density $P(c_{i,l}, c_{j,l}|x_l^1, \ldots, x_l^N, M)$. Since the data-generating mechanism is taken to be a Gaussian, terms on the righthand side can be evaluated from the mean and variance of the Gaussian modeling class k and gene l

$$P(x_l^n|c_{k,l}) = \frac{1}{\sqrt{2\pi}\sigma_{k,l}} \exp -\frac{1}{2}\left[\frac{x_l^n - \mu_{k,l}}{\sigma_{k,l}}\right]^2$$

The sign for the NBR value is obtained from $sign(\mu_{j,l} - \mu_{i,l})$. A negative (positive) sign indicates that the expression level in c_i is higher (lower) than that in class c_j. If $NBR_{ij}(F_l) = 0.0$, then the expression in class c_i is identical to that in c_j. The larger the absolute value, the more distinct the expression levels and the more likely gene F_l is to be a marker. Given $K \geq 3$ classes, the naive Bayes global relevance (NBGR) is the sum of the NBR over pairwise combinations of the classes

$$NBGR(F_l) = \frac{1}{K}\sum_{i=1}^{K}\sum_{j=i+1}^{K} NBR_{ij}(F_l)$$

Naive Bayes Model-Based Feature Relevance Expert

The probability parameters for the K classes of an unsupervised naive Bayes model (see above, *Graphical Models*) are used to calculate $NBR_{ij}(F_1), \ldots, NBR_{ij}(F_{1988})$ and $NBGR(F_1), \ldots, NBGR(F_{1988})$. The 1,988 genes are reordered according to their NBR and NBGR values. The ranking based on the NBGR values is termed the "K-class unsupervised NBGR ranking." The probability parameters for the $K = 2$ classes of a supervised naive Bayes model are used to calculate NBGR values. The ranking based on these NBGR values is termed the "$K = 2$ supervised NBGR ranking."

For each ranking, representative gene subsets are created by selecting different numbers of top-ranked genes. Each subset is employed to create training examples for leave-one-out cross-validation studies in which the input vectors contain only the selected genes. Rather than working directly with the original expression levels, x_l^n, each value is normalized using $x_l^n/[\sum_{l \in S} (x_l^n)^2]^{1/2}$ where S is the gene subset of interest. For simplicity and to illustrate the basic approach, genes are ranked once using all N training examples and not for each $N - 1$ estimation set.

Supervised Learning System: SVM vs. Naive Bayes Model

In addition to being a generative model for unsupervised learning, a naive Bayes model can be used for supervised learning and prediction. Given a model that has grouped training data into K classes, the posterior probability of each class given a test example $P(c_k|\mathbf{X}_L^n)$ is computed. The test example is assigned to the class which maximizes this value. To compare SVMs and naive Bayes models as supervised learning systems, N supervised naive Bayes models are trained and tested using the same leave-one-out cross-validation strategy employed to evaluate SVMs (see above, *Support Vector Machines: SVM^{light}*). The generalization performance of these two systems is compared using feature subsets derived from the K-class unsupervised NBGR ranking and the $K = 2$ supervised NBGR ranking.

Outliers and Potentially Mislabeled Specimens

Support vectors define the location of the decision surface (solid symbols in Fig. 4) whereas nonsupport vectors (open symbols) do not participate in its specification. One method for identifying outliers and potentially mislabeled specimens is pinpointing positive and negative training examples that are the support and nonsupport vectors. For each leave-one-out SVM, the training examples that constitute the support vectors and nonsupport vectors are ascertained. An "invariant support vector training example" is one that is a support vector in all the $N - 1$ SVMs which placed it in the estimation set. Similarly, an "invariant nonsupport vector training example" is one that is never a support vector. This approach presumes no mislabeled examples and uses a hard margin for SVM training. A soft margin would permit training examples to violate the decision boundary subject to some penalty.

RESULTS

Unsupervised Learning Using Naive Bayes Models

An unsupervised naive Bayes model trained using sixty-two 1,988-feature experiment profile vectors identified four classes that will be referred to as *classes 1–4.* The two classes that might be expected a priori given the tumor and nontumor labels differ from the number estimated from the data. Each profile vector is assigned to the class that maximizes the posterior probability $P(c_k|\mathbf{X}_L^n)$. The results are *class 1,* 2 nontumor, 19 tumor; *class 2,* 9 nontumor, 8 tumor; *class 3,* 11 nontumor, 4 tumor; *class 4,* 0 nontumor, 9 tumor. While *classes 1* and *4* contain primarily tumor specimens, *classes 2* and *4* are mixtures. The four classes may reflect the composition of the tissue specimens. While tumor specimens were biased toward epithelial tissue, nontumor specimens probably included a mixture of tissue types (2). Tumor specimens in *classes 2* and *3* may contain a high degree of nonepithelial tissue. The results indicate that transcription profile data can distinguish tumor from nontumor specimens and suggest the homogeneity of the original specimens.

The published study clustered the profile vectors by means of a binary tree computed using an algorithm based on deterministic annealing (2). *Clusters 1* contained *3* nontumor and 35 tumor specimens, whereas *cluster 2* contained 19 nontumor and 5 tumor specimens; 36 of the 38 specimens assigned to *cluster 1* belong to *classes 1, 2,* or *4.* Although *class 4* is a subset of *cluster 1,* the specimens are scattered throughout the clustering tree. Hence, the unsupervised naive Bayes model defines an important subgroup of tumor specimens not detected by the binary clustering method.

Supervised Learning Using SVMs

Table 1 shows that when all 1,988 genes are used, 55 (89%) of the SVMs make consistent assignments. *Classes 1* and *4* are associated with consistent assignments, whereas *classes 2* and *3* contain the three nontumor and four tumor inconsistent assignments (boxed in Table 1). The discrepancy between the generalization performance achieved, 55, and the maximum possible, 62, indicates the divergence between the known labels in Alon et al. (2) and the assigned labels here. The seven false negative and false positive assignments (boxed) are valid only within the context of the original labels. Possible explanations for these seven "differences," especially the two for *patient 36,* include *1*) deficiencies in the SVM learning method used, *2*) specimens may have been mislabeled as a result of human error, and *3*) pathologically "normal" regions of the colon could have substantial tumor-like properties from a molecular standpoint.

Table 1. *Results from leave-one-out cross-validation studies that used SVMs and all 1,988 genes to distinguish tumor from nontumor patient specimens*

Patient	*Class 1*		*Class 2*		*Class 3*		*Class 4*	
1		T†			N†			
2	N†	T†						
3			N*	T				
4		T	N†					
5			N	T*				
6	N†	T*						
7			N	T†				
8			[N†]	T†				
9			N	T				
10			N*	T*				
11		T†			N			
12			N†	T				
13								T
14		T*						
15		T†						
16		T*						
17		T*						
18		T						
19		T						
20		T*						
21								T
22		T*						
23		T						
24								T*
25		T*						
26								T*
27		T*	N†					
28		T			N†			
29					N			T*
30						[T†]		
31								T*
32					N†			T*
33					N	[T†]		
34					[N†]			T
35				[T†]	N			
36					[N†]	[T†]		
37		T†						
38		T*						
39					N			T
40					N†	T†		
Total	2	19	9	8	11	4	0	9

Each of the 62 specimens is listed under the naive Bayes model class to which it is assigned. True positive or true negative assignments are consistent with the known label (2), whereas false positive or false negative assignments are inconsistent. If the assignment made by an SVM for its test example is inconsistent, then the specimen is boxed. For example, when the nontumor specimen of *patient 8* comprised the test set, an SVM trained using the remaining 61 specimens assigned it a "tumor" label. †Invariant support vector training example. *Invariant nonsupport vector training example. T, tumor (positive training example); N, nontumor.

Analysis of the invariant support and nonsupport training examples can suggest outliers and mislabeled samples. There are 24 invariant support vector training example: 12/22 nontumor cases and 12/40 tumor cases († in Table 1). There are 18 invariant nonsupport vector training examples: 2/22 nontumor cases and 16/40 Tumor cases (asterisks in Table 1). These latter 18 examples should form the core of any revised training set, because they are the most unambiguous and none belong to *class 3,* the class that appears to be the most problematic in terms of tissue composition and inconsistent assignments. The six inconsistently assigned, invariant support vector training examples (boxed and † in Table 1) can be flagged as requiring further investigation to clarify their labels. These are the nontumor specimens of *patients 8, 34,* and *36* and the tumor specimens of *patients 30, 33,* and *36.*

Overall, the results suggest the presence of three subtypes of specimens: those that are clearly tumor (*classes 1* and *4*), those that are mainly nontumor (*class 3*), and those that are heterogeneous or have a mixed tissue composition (*class 2*). The two tumor classes could, for example, indicate different pathways for reaching the same biological endpoint and/or variation in the treatment schedules or clinical histories of the patients.

NBR: Genes That Distinguish Class 4 From 1, 2, or 3

The NBR measure quantitates the degree to which gene F_l distinguishes *class i* from *j* [see above, *Feature Relevance: Naive Bayes (Global) Relevance*]. Genes with the highest and lowest values are the top- and bottom-ranked genes, respectively. Since Table 1 suggests that *class 4* is perhaps the most interesting, this Class will be employed as an exemplar to illustrate the utility of the overall approach (Fig. 5). The aim of the subsequent analysis is not to provide a detailed discussion of all the genes and their potential roles, but to demonstrate that NBR values provide a useful mechanism for pinpointing biologically plausible candidates for subsequent in-depth studies. For example, genes that distinguish *class 4* from the other three classes include immunoglobulin superfamily receptors (Fc receptor hFvRn) and laminin receptors. Immunoglobulin receptors are known to be associated with malignant transformation and dissemination of colon tumors (48).

Relative to *class 4,* the following genes are upregulated in *classes 1, 2,* and *3* (red, Fig. 5).

Precursors for both complement C1s and C1r. These proteases are responsible for the lectin pathway activation and proteolytic activity of the C1 complex of complement, an activation system designed for the

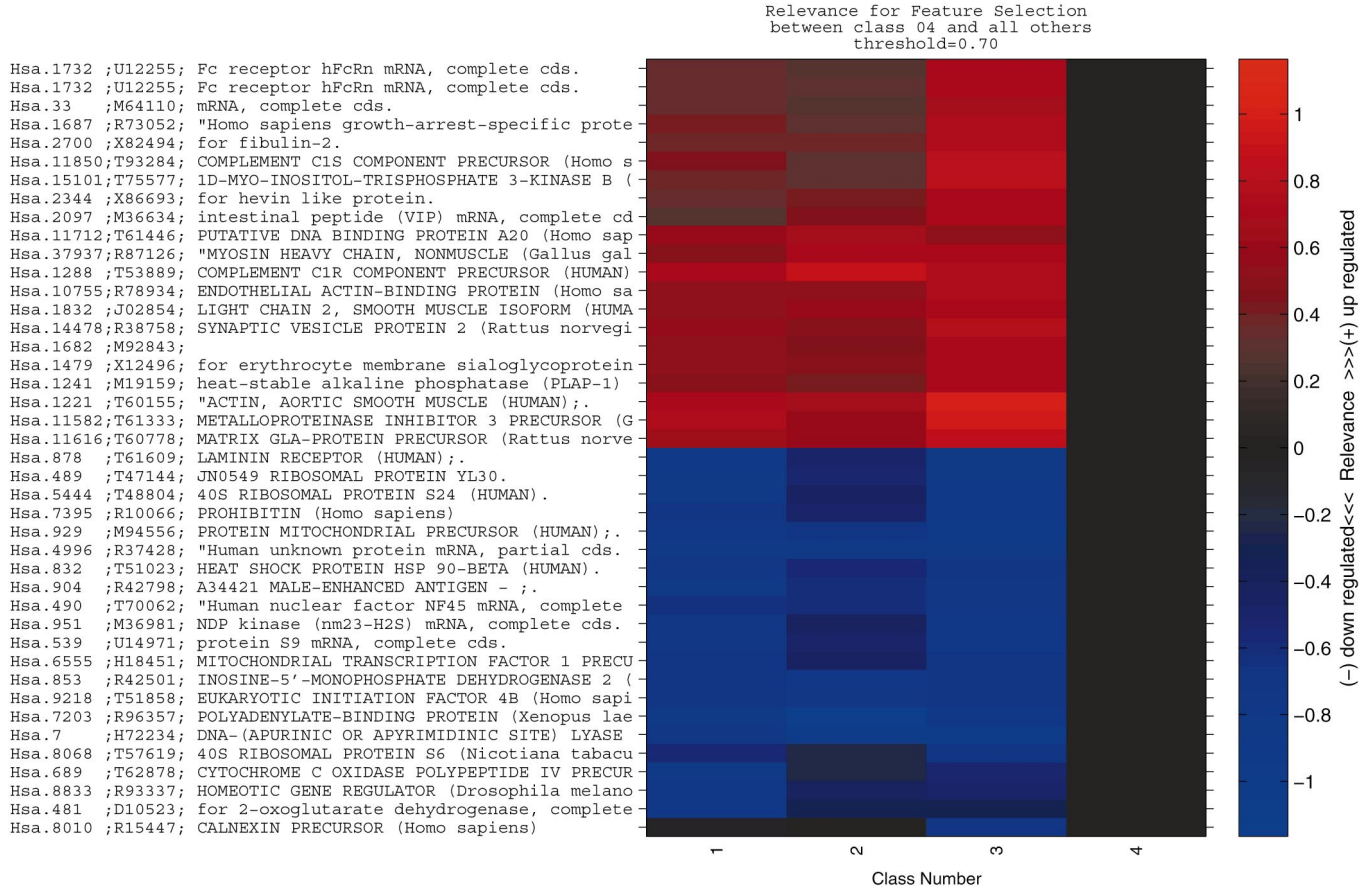

Fig. 5. Genes that best distinguish *class 4* from the other three classes according to the naive Bayes relevance (NBR) measure. Each row represents the same gene F_l and the columns, from *left* to *right,* its $NBR_{41}(F_l)$, $NBR_{42}(F_l)$, $NBR_{43}(F_l)$, and $NBR_{44}(F_l)$ values. NBR values are color-coded such that more intense colors signify higher values and thus noteworthy marker genes. The precise color indicates the sign of the NBR values and thus the direction of change of the expression level. Positive (red) denotes that, relative to *class 4,* the gene is upregulated in the other class. Negative (blue) indicates downregulation. The 1,988 genes are filtered to show only those for which $|NBR_{4j}(F_l)| \geq 0.7$, i.e., those best able to discriminate *class 4* from *1, 2,* or *3.* Relative to *class 4,* "Hsa.1221: ACTIN, AORTIC SMOOTH MUSCLE (HUMAN)" is the gene that is most upregulated in any of the other classes. Relative to *class 4,* "Hsa.7203: POLYADENYLATE-BINDING PROTEIN (*Xenopus laevis*)" is the gene that is most downregulated in any of the other classes. The NBR is a log scale, but the brightness of a color is a linear scale.

elimination of pathogens. The lectin pathway plays an important role in innate immunity.

Fibulin-2. This extracellular matrix (ECM) protein is present in the basement membrane and stroma of many tissues, and its expression pattern suggests an essential role in organogenesis, particularly in embryonic heart development.

Hevin. This ECM protein is important for the adhesion and trafficking of cells through the endothelium. Hevin has been shown to be downregulated in non-small cell lung cancer (4) and metastatic prostate adenocarcinoma (38).

Vasoactive intestinal peptide. Vasoactive intestinal peptide (VIP) has been implicated as an important factor in several inflammatory conditions of the human gut.

Tumour necrosis factor-α inducible protein A20. This putative DNA binding protein is a Cys_2/Cys_2 zinc finger protein induced by a variety of inflammatory stimuli and characterised as an inhibitor of cell death.

The cytoskeletal proteins actin and myosin and endothelial actin-binding protein. Relative to *class 4,* these proteins are downregulated in *classes 1, 2,* and *3* (blue, Fig. 5).

Polyadenylate-binding protein. This protein recognises the 3′ mRNA poly(A) tail and plays critical roles in eucaryotic translation initiation and mRNA stabilization/degradation.

DNA-apurinic or apyrimidinic site lyase APE1/HAP1. This protein plays an important role in DNA repair and in the resistance of cancer cells to radiotherapy.

KAP-1. This protein ($TIF1^{\beta}$/KRIP-1; human unknown protein mRNA; R37428) may be a corepressor for the large class of KRAB-containing zinc finger proteins (1).

Calnexin precursor. This protein is a chaperone that promotes the correct folding and oligomerisation of many glycoproteins. A study of protein changes associated with ionizing radiation-induced apoptosis in hu-

man prostate epithelial tumor cells indicated that the proteins levels of this molecular chaperone are higher in such dying cells (41).

Inosine 5′-monophosphate dehydrogenase 2. Inosine 5′-monophosphate dehydrogenase 2 (IMPDH isoform 2) enzyme is the rate-limiting enzyme in the de novo synthesis of guanine nucleotide. Of the two isoforms, IMPDH isoform 2 is selectively upregulated in neoplastic and replicating cells and is thus considered to be a sensitive target for cancer chemotherapy (reviewed in Ref. 12).

Overall, the results suggest that tumor specimens belonging to *classes 1* and *4* have very distinctive properties. For example, NDP kinase (nm23-H2S) is known to be associated with tumor metastasis (13), but the levels in these classes are very different. There are marked differences in genes related to cell growth, protein synthesis, energy metabolism, oxidative stress, and apoptosis. Greater knowledge of the clinical histories of the patients from which these tumor specimens were taken may reveal the origins of these differences. One possibility based on the expression patterns of calnexin and IMP is that patients whose tumor samples are assigned to *class 4* may have received radiation or other therapy.

In some instances, differential expression at the gene level is mirrored at the protein level. Prohibitin and IMPDH-2 are proteins that have been shown to exhibit differential protein expression in normal and neoplastic human breast epithelial cell lines (54). The levels of the latter enzyme in tumor cell lines was elevated 2- to 20-fold relative to the levels in normal cells. Relative to tumor *Class 4,* the expression levels of the genes for these enzymes exhibit a similar pattern in that they are downregulated in the other classes.

NBGR: Genes That Distinguish All Classes

The NBGR measure quantitates the degree to which gene F_l distinguishes all four classes. Genes with the highest and lowest values are the top- and bottom-ranked genes, respectively. The top 50 NBGR ranked genes are listed in Table 2. Of the feature subsets examined (discussed below, *Naive Bayes Model-Based Feature Relevance Experts*), the top 50 represents the smallest number of features that generalize as well as all 1988 genes. Selected genes of potential interest are as follows.

Serum response factor. Serum response factor (SRF) regulates transcription of many serum-inducible and muscle-specific genes. It binds to the serum response element, a DNA sequence required for the transcription of a number of genes in response to growth factor or mitogen stimulation. A number of these types of genes are present in the top 50: β′-actin, myosin light chain, and profilin I. This is consistent with the observation that signal-regulated activation of SRF is mediated by changes in actin dynamics (46). These genes might provide an indication of the migratory capacity of the cells in the specimens and hence their propensity for metastasis.

Ferritin. Low serum ferritin levels are associated with patients having serious gastrointestinal pathologies such as neoplasia and acid peptic disease (29). Previous work has shown that the majority of colorectal adenocarcinomas exhibit ferritin expression (20), but the clinical significance remains unknown.

Tra1/GRP94/GP96. This molecular chaperone been suggested to be useful in cancer immunotherapy (39). The level of the protein is higher in human breast cancer cell lines compared to normal basal epithelial cell lines (15). Figure 5 indicates that HSP 90-β, another of member of the heat shock protein 90 family to which Tra1 belongs, is downregulated in *classes 1, 2,* and *3* relative to *class 4.*

In a manner analogous to comparative sequence analysis, comparative analysis of molecular profile data may be useful for inferring the potential physiological roles of genes. Such comparison of the expression patterns of orthologous and paralogous proteins can be illustrated using "translationally controlled tumor protein" (TCTP, HRF P23), the ninth ranked gene. TCTP is a eucaryotic cytoplasmic protein found in several normal and tumor cells that is suggested to have a general, yet unknown, housekeeping function (44). Comparative sequence analysis (data not shown) provides few insights into the biological role of this evolutionarily conserved protein and a protein that may have a role in colon cancer. A naive Bayes model trained using 5,687 seventy-eight-feature yeast gene profile vectors found 45 classes (35). The yeast TCTP homologue (TCTP_YEAST; YKL056C) is found in a class populated with genes from the MIPS (33) protein functional category "PROTEIN SYNTHESIS: ribosomal proteins." Physiologically, therefore, and consistent with other genes in the top 50, human TCTP may be a ribosome-associated protein.

Marker Genes for Understanding Colon Adencarcinoma Biology

One mechanism for generating a set of candidates for subsequent study is by taking the union of the NBGR top 50 listed in Table 2 and the genes shown in Fig. 5. Experimental data support the notion that these 89 genes may be biologically relevant. For example, the set includes genes shown to be differentially expressed in mucus-secreting cells and undifferentiated HT-29 colon cancer cells: transcripts encoded by the mitochondrial genome, components of the protein synthesis machinery, ferritin, and TCTP (37). Alterations in the distribution and/or adhesiveness of laminin receptors in colon cancer cell lines may be associated with increased tumorigenicity (27). A study of cultured colon cancer cells suggests that laminin may play an important role in hematogeneous metastasis by mediating tethering and spreading of colon cancer cells under blood flow (28). In general, the markers are involved in cell signaling, adhesion and communication, immune response, heat shock, and DNA repair. Adhesion receptors and cell surface-associated molecules mediating cell-matrix and

Table 2. *The top 50 genes that distinguish all four classes according to the naive Bayes global relevance measure*

Gene ID		Gene Annotation
Hsa.689	H55933	*Homosapiens* mRNA for homolog to yeast ribosomal protein L41
Hsa.5444	R39465	Eukaryotic initiation factor 4A (*Oryctolagus cuniculus*)
Hsa.2191	R39465	Eukaryotic initiation factor 4A (*O. cuniculus*)
Hsa.2097	R85482	Serum response factor (*H. sapiens*)
Hsa.1682	U14973	Protein S29 mRNA, complete cds
Hsa.7395	R02593	60S Acidic ribosomal protein P1 (*Polyorchis penicillatus*)
Hsa.1479	T51496	60S Ribosomal protein L37A (human)
Hsa.11850	H80240	Inter-α-trypsin inhibitor complex component II precursor (*H. sapiens*)
Hsa.2344	T65938	Translationally controlled tumor protein (human)
Hsa.1221	T55131	Glyceraldehyde-3-phosphate dehydrogenase, liver (human)
Hsa.490	T72863	Ferritin light chain (human)
Hsa.878	H86060	Negative factor (Simian immunodeficiency virus)
Hsa.549	X63432	mRNA for mutant β-actin (β′-actin)
Hsa.11616	H20709	Myosin light chain alkali, smooth-muscle isoform (human)
Hsa.904	U14971	Ribosomal protein S9 mRNA
Hsa.8068	T52342	Human tra1 mRNA for human homolog of murine tumor rejection antigen gp96
Hsa.539	L28809	dbpB-like protein mRNA
Hsa.7203	T63508	Ferritin heavy chain (human)
Hsa.6080	H09263	Elongation factor 1-α 1 (*H. sapiens*)
Hsa.15101	T49423	Breast basic conserved protein 1 (human)
Hsa.572	H79852	60S Acidic ribosomal protein P2 (*Babesia bovis*)
Hsa.11582	J02763	Gene
Hsa.951	R22197	60S ribosomal protein L32 (human)
Hsa.33	T59954	Thymosin β-4 (human)
Hsa.1288	H80240	Inter-α-trypsin inhibitor complex component II precursor (*H. sapiens*)
Hsa.18897	T95018	40S Ribosomal protein S18 (*H. sapiens*)
Hsa.1687	H86060	Negative factor (Simian immunodeficiency virus)
Hsa.8583	T63484	Human ornithine decarboxylase antizyme (Oaz) mRNA
Hsa.7	R02593	60S Acidic ribosomal protein P1 (*P. penicillatus*)
Hsa.489	M11799	Class I HLA-Bw58 gene
Hsa.3349	**T61609**	Laminin receptor (human)
Hsa.11712	T62220	Calpactin I light chain (human)
Hsa.539	**T51574**	40S Ribosomal protein S24 (human)
Hsa.2710	T48041	Human mRNA fragment for the β-2 microglobulin
Hsa.2700	T96832	Interferon-α receptor precursor (*H. sapiens*)
Hsa.929	H54676	60S Ribosomal protein L18A (human)
Hsa.491	R86975	40S Ribosomal protein S28 (human)
Hsa.8147	T63258	Elongation factor 1-α 1 (human)
Hsa.695	**T57619**	40S Ribosomal protein S6 (*Nicotiana tabacum*)
Hsa.27537	T88723	Ubiquitin (human)
Hsa.14478	R36455	Nucleolar transcription factor 1 (*H. sapiens*)
Hsa.45293	T61602	40S Ribosomal protein S11 (human)
Hsa.6555	T58861	60S Ribosomal protein L30E (*Kluyveromyces lactis*)
Hsa.41875	U21909	mRNA
Hsa.479	T61661	Profilin I (human)
Hsa.25322	T52015	Elongation factor 1-γ (human)
Hsa.1610	H24754	Fructose-bisphosphate aldolase A (human)
Hsa.1732	H22688	Ubiquitin (human)
Hsa.9218	T93094	Annexin II (human)
Hsa.5971	T51560	40S Ribosomal protein S16 (human)

Genes in bold appear in Fig. 5 and are those that distinguish *class 4* from *1, 2,* or *3* based upon their $NBR_{4j}(F_l)$ values.

cell-cell interactions are known to play an important role in tumor cell migration, invasion, and metastasis.

Selecting markers for use in understanding colon adenocarcinoma biology, creating diagnostic tools, and highlighting targets for therapeutic intervention and drug-design requires reducing the original 1,988 genes to smaller, more manageable subsets. The aforementioned markers were defined using a fairly stringent threshold $|NBR_{ij}(F_l)| \geq 0.7$ and a small fixed number of top-ranked genes (i.e., 50). This "low-hanging fruit" approach is unable to detect genes involved in more subtle interactions.

Naive Bayes Model-Based Feature Relevance Experts

Developing a decision support system may require using a larger number of genes than an experimental investigator might be interested in pursuing. A learning system designed to discriminate between tumor and nontumor specimens should optimize specificity and generalization performance rather than minimize the number of genes proposed as being important. The consequences of a tumor specimen labeled incorrectly as nontumor (a false negative) may be more severe than overpredicting false positives.

One approach to identifying markers for prototype decision support systems is by means of a feature relevance expert. Table 3 shows the generalization performance of leave-one-out SVMs trained using 11 feature subsets. The maximum generalization performance achieved, 55, is less than the maximum possible, 62. The top 50 genes perform as well as the full repertoire of 1,988 genes. Using only the top two degrades the overall performance by only three (55 to 52). Further studies are required to assess whether, for example, any 2 in the top 10 would have the same performance as the top 2. The NBGR ranking appears to be meaningful because the performance of the top 500, 50, 25, 10, 5, and 2 subsets is consistently higher than the equivalent number of bottom-ranked genes. As the number of genes used decreases from 500 to 2, the difference in performance increases from 52 − 50 = 2 to 52 − 27 = 25. As shown elsewhere (8), there are likely to be other subsets of 50 genes that have some or no overlap with the NBGR top 50 but which have the same generalization performance.

The patients can be divided into three broad groups based on their pattern of assignments. The first group includes 34 specimens whose labels are consistently assigned irrespective of the subset used. It includes 7/9 members of *class 4*. The second group includes six specimens that are invariably inconsistently assigned. Most are members of the problematic *class 3* shown in Table 1. Additional studies are required to assess whether these six specimens are genuine outliers (*patient 30,* tumor; *patient 33,* tumor; *Patient 36,* nontumor and tumor; *Patient 34,* nontumor; *patient 8,* non-

Table 3. *The generalization performance of leave-one-out SVMs trained using 11 different NBGR-ranked feature subsets*

Patient	1,988		1,750		1,500		1,250		1,000		500		50		25		10		5		2	
1	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T
3	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T
5	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T
6	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T
7	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T
9	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T
10	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T
27	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T
29	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T
40	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T
13		T		T		T		T		T		T		T		T		T		T		T
14		T		T		T		T		T		T		T		T		T		T		T
15		T		T		T		T		T		T		T		T		T		T		T
16		T		T		T		T		T		T		T		T		T		T		T
17		T		T		T		T		T		T		T		T		T		T		T
18		T		T		T		T		T		T		T		T		T		T		T
21		T		T		T		T		T		T		T		T		T		T		T
22		T		T		T		T		T		T		T		T		T		T		T
24		T		T		T		T		T		T		T		T		T		T		T
25		T		T		T		T		T		T		T		T		T		T		T
26		T		T		T		T		T		T		T		T		T		T		T
31		T		T		T		T		T		T		T		T		T		T		T
38		T		T		T		T		T		T		T		T		T		T		T
39		T		T		T		T		T		T		T		T		T		T		T
8	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T
34	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T
36	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T
33	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T
30		T		T		T		T		T		T		T		T		T		T		T
35	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T
28	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T
4	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T
37		T		T		T		T		T		T		T		T		T		T		T
12	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T
19		T		T		T		T		T		T		T		T		T		T		T
2	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T
20		T		T		T		T		T		T		T		T		T		T		T
11	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T
32	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T
23		T		T		T		T		T		T		T		T		T		T		T
Generalization Performance	55		55		55		54		54		52 (50)		55 (33)		52 (41)		53 (37)		52 (38)		52 (27)	

The rows representing "patient" have been reordered to highlight patterns across the subsets (only 22 patients have matched tumor and nontumor specimens). A test example for which the SVM assignment is inconsistent is boxed. "Generalization performance" denotes the number of true positive and true negative assignments (maximum possible performance 62). Numbers in parenthesis indicate the performance for the bottom 500, 50, 25, 10, 5 and 2 ranked genes. The 50 top-ranked genes are listed in Table 2.

tumor). The assignments for the third group of 18 specimens changes from consistent (inconsistent) to inconsistent (consistent) as the number of top-ranked genes is reduced.

Although the exact shape of the function relating performance to the number of top-ranked genes is unknown, it is possible to improve the performance by examining subsets in the 500–50 range. Table 4 shows that of all the subsets examined, the maximum generalization performance is achieved with the top 200 genes (56). The original 62 training examples were partitioned such that the 56 consistently assigned specimens (N or T in Table 4) formed the estimation set. The remaining six specimens formed the test examples. The assignments made by an SVM trained using the top 200 genes did not change, i.e., the false positive and false negative assignments support the notion that these six specimens are likely to be outliers. The results suggest that the 200 top-ranked genes from the 56 aforementioned specimens could be used to develop a prototype diagnostic tool. Further studies are required to ascertain the success of such a tool when used for large-scale colon adenocarcinoma screening studies.

Learning System: SVM vs. Supervised Naive Bayes Model

Table 5 shows how the leave-one-out learning method and naive Bayes model used to compute the NBGR ranking affect performance. SVMs are consistently better than supervised naive Bayes models. Features subsets derived from an NBGR ranking based on the number of estimated classes ($K = 4$) outperform those in which the classes are defined according to the

Table 4. *Improving the generalization performance of leave-one-out SVMs by fine tuning the number of top-ranked genes used for training*

Patient	500		400		300		200		150		100		75		50	
1	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T
3	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T
5	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T
6	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T
7	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T
9	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T
10	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T
11	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T
27	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T
28	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T
29	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T
32	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T
40	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T
13		T		T		T		T		T		T		T		T
14		T		T		T		T		T		T		T		T
15		T		T		T		T		T		T		T		T
16		T		T		T		T		T		T		T		T
17		T		T		T		T		T		T		T		T
18		T		T		T		T		T		T		T		T
19		T		T		T		T		T		T		T		T
20		T		T		T		T		T		T		T		T
21		T		T		T		T		T		T		T		T
22		T		T		T		T		T		T		T		T
23		T		T		T		T		T		T		T		T
24		T		T		T		T		T		T		T		T
25		T		T		T		T		T		T		T		T
26		T		T		T		T		T		T		T		T
31		T		T		T		T		T		T		T		T
38		T		T		T		T		T		T		T		T
39		T		T		T		T		T		T		T		T
8	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T
34	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T
36	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T
33	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T
30		T		T		T		T		T		T		T		T
2	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T
4	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T
35	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T
37		T		T		T		T		T		T		T		T
12	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T
Generalization Performance	52		55		55		56		55		55		54		55	

The general format is the same as Table 3; the top 500 and 50 results are reproduced for comparison purposes.

Table 5. *The generalization performance of two different learning systems trained using feature subsets derived from two different NBGR rankings*

Learning Method	NBGR Ranking	100	50	25	10	5	2
SVM	*K*-class unsupervised	55	55	52	53	52	52
SVM	$K = 2$ supervised	42	46	48	47	38	20†
Supervised naive Bayes model	*K*-class unsupervised	42	42	45	48	49	35
Supervised naive Bayes model	$K = 2$ supervised	34	36	34	33	29	33

"Learning method" indicates whether the leave-one-out models were SVMs or supervised naive Bayes models (see *Naive Bayes Model-Based Feature Relevance Expert,* in METHODS). "NBGR Ranking" denotes whether the NBGR values used to rank the 1,988 genes were based on an unsupervised (*K*-class unsupervised) or supervised ($K = 2$ supervised) naive Bayes model. The 6 feature subsets examined contained the 100, 50, 25, 10, 5 and 2 top-ranked genes. †No model could be found when the tumor specimen from *patient 24* was the test example, so the maximum generalization performance possible in this instance is 61 not 62.

Table 6. *The K + 2 supervised NBGR top 50 genes*

Gene ID		Gene Annotation
Hsa.42738	H55933	Calcineurin B subunit isoform 1 (*H. sapiens*)
Hsa.3969	R39465	Factor 1 mRNA, complete cds
Hsa.8192	R39465	Macrophage colony stimulating factor-1 precursor (*H. sapiens*)
Hsa.2463	R85482	QM protein (human)
Hsa.33699	U14973	Protein (CRP) gene, exons 5 and 6
Hsa.23249	R02593	For c-sis gene (clone pSM-1)
Hsa.848	T51496	Twitch skeletal muscle/cardiac muscle troponin C gene, complete cds
Hsa.33982	H80240	And polyadenylation specificity factor mRNA, complete cds
Hsa.8010	T65938	Binding inhibitor (DBI) mRNA, complete cds
Hsa.30310	T55131	HLA class II histocompatibility antigen, DR-1 beta chain (human)
Hsa.35201	T72863	SKD1 protein (*Mus musculus*)
Hsa.3065	H86060	Lysosomal protective protein precursor (HUMAN)
Hsa.14595	X63432	Thyroid receptor interactor (TRIP1) mRNA, complete cds
Hsa.35741	H20709	60S ribosomal protein L7A (human)
Hsa.479	**U14971**	Profilin I (human)
Hsa.41164	T52342	Integral membrane protein, calnexin, (IP90) mRNA, complete cds
Hsa.32404	L28809	(human)
Hsa.43284	T63508	Aflatoxin B1 aldehyde reductase (*Rattus norvegicus*)
Hsa.2910	H09263	Trans-acting transcriptional protein ICP0 (Herpes simplex virus)
Hsa.538	T49423	Histone, class B mRNA, complete cds
Hsa.2529	H79852	For tyrosine hydroxylase type 3
Hsa.38205	J02763	C substrate, 80-kDa protein, heavy chain (human); contains TAR1 repetitive element
Hsa.11712	**R22197**	Calpactin I light chain (human)
Hsa.28162	T59954	RD protein (human)
Hsa.32358	H80240	UNC-33 protein (*Caenorhabditis elegans*)
Hsa.6048	T95018	Platelet-activating factor acetylhydrolase 45-kDa subunit (*Bos taurus*)
Hsa.37254	H86060	Protein 42 (human)
Hsa.24279	T63484	PP1-γ catalytic subunit (human)
Hsa.8831	R02593	ER lumen protein retaining receptor 1 (*H. sapiens*)
Hsa.34416	M11799	(huc) mRNA, complete cds
Hsa.27560	T61609	14-3-3-like protein GF14 omega (*Arabidopsis thaliana*)
Hsa.1731	T62220	γ3 heavy chain disease OMM protein mRNA
Hsa.896	T51574	P37879 lysyl-tRNA synthetase
Hsa.25536	T48041	Phospholipase A_2, membrane associated precursor (human)
Hsa.35528	T96832	Inhibin β A chain precursor (*M. musculus*)
Hsa.1258	H54676	JC2042 SUI1 translation initiation factor
Hsa.36657	R86975	Of *Drosophila* discs large protein, isoform 2 (hdlg-2) mRNA, complete cds
Hsa.3280	T63258	Calcium/calmodulin-dependent protein kinase type II delta chain (*R. norvegicus*)
Hsa.2359	T57619	Merozoite surface antigens precursor (*Plasmodium falciparum*)
Hsa.41247	T88723	For RNA polymerase II associated protein RAP74
Hsa.15115	R36455	ATP synthase γ chain, mitochondrial precursor (human)
Hsa.45499	T61602	IG kappa chain precursor V-III region (human)
Hsa.19143	T58861	Polymerase II subunit hsRPB7 mRNA, complete cds
Hsa.1672	U21909	For protein kinase C-γ (partial)
Hsa.168	T61661	60S Ribosomal protein L7 (human)
Hsa.587	T52015	CD63 antigen (human)
Hsa.3086	H24754	Farnesyl pyrophosphate synthetase (human)
Hsa.3026	H22688	Synthase subunit B, brain isoform (human)
Hsa.4907	T93094	Protein kinase CEK1 (*Schizosaccharomyces pombe*)
Hsa.209	T51560	General negative regulator of transcription subunit 1 (*Saccharomyces cerevisiae*)

Genes in bold are present in the *K*-class unsupervised NBGR top 50 shown in Table 2.

known tumor or nontumor labels ($K = 2$). Profilin I and calpactin I light chain are the only genes in common to the $K = 2$ supervised NBGR top 50 (Table 6) and the K-class unsupervised NBGR top 50 (Table 2). Since they are highlighted as being important by two independent ranking schemes, these genes may be noteworthy markers. Biologically, they suggest that regulation of cell morphology via control of cell adhesion and cytoskeletal molecules could be important factors in understanding colon adenocarcinoma biology.

The results indicate that ranking and selecting markers that distinguish tumor from nontumor specimens is best achieved by estimating the number of underlying experiment profile vector classes rather than assuming the presence of $K = 2$ classes suggested by the observed phenotype. Given a method for generating (disjoint) classes of profile vectors such as gene shaving (19), other K-class unsupervised NBGR rankings could be determined. The performance of feature subsets derived from such NBGR-based rankings as well as alternative methods for calculating feature relevance are areas for future research. For example, the Bayesian technique known as automatic relevance determination (ARD) (MacKay DJC and Neal RM, unpublished observations) uses labeled data to compute a regularization coefficient for each feature; large values signify variables that are less relevant to the decision. These coefficients could be used to rank genes.

DISCUSSION

Here, a modular framework for the analysis of molecular profile data and domain knowledge was proposed as a method for understanding basic mechanisms and developing decision support systems for diagnosis, prognosis, and monitoring. Specific generative (graphical models) and discriminative (SVMs) methods were suggested as techniques for addressing tasks associated with certain modules. Published sixty-two 1,988-feature experiment profile vectors from colon adenocarcinoma tissue specimens labeled as tumor or nontumor were analyzed using a combination of an unsupervised (naive Bayes model) and supervised (SVM) learning methods. Putative tumour subtypes were identified, "tumor" or "nontumor" labels were assigned to new specimens, and six potentially mislabeled specimens were detected. The profile vector classes discovered and characterized by the naive Bayes model were used as the basis for feature selection. SVMs trained using feature subsets derived from these rankings had the same or better generalization performance than the full repertoire of 1,988 genes. Approximately 90 biologically plausible marker genes were pinpointed for use in understanding the etiology of colon adenocarcinoma, defining targets for therapeutic intervention, and developing diagnostic tools.

A more thorough interpretation of and explanation for the results would be possible if information such as the sizes, sites, and disease stages of cancer for the tumors and patient histories were available. Given such information, the gene expression measurements could be correlated with potential clinical outcomes such as radiosensitivity and response of the tumor to chemotherapy. The strategy utilized here is sufficiently general that it can be applied to other transcription profiling studies as well as other types of molecular profile data.

The results reiterate the importance of controlling and optimizing the experimental techniques used to obtain and handle in vivo specimens because of their impact on the information that can be extracted. The aforementioned markers implicate the microenvironment, cell-matrix interactions, cell-cell communication, and the immune system as key factors that differentiate nontumor from tumor colon adenocarcinoma tissue specimens. It remains to be seen whether transcription profiles derived from cell lines or cultures would have highlighted the role of tissue biology in this disorder. It will be necessary to compare tumor and nontumor specimens with those from individuals having no record of adenocarcinoma. Arrays containing the full complement of human genes and not just the selected set employed here are likely to reveal additional marker genes. For the purpose of developing and using a robust decision support system, it is critical that collection and preparation of all specimens conform to a standardized procedure in order to minimize heterogeneity in the cell types assayed.

Learning biologically realistic networks from data even with the aid of domain experts remains a challenging task. This stems from the nature and quality of the available data, theoretical issues of learning models with large numbers of noisy variables, and efficient implementations of the modeling methods. For example, mRNA transcript and protein levels are not necessarily correlated (18). Clearly, genetic networks inferred from molecular profile data alone will be insufficient to understand many aspects of the behavior of cells and tissues. Nonetheless, the results in this and related work (8, 35) suggest that the framework and techniques proposed here have the potential for creating robust decision support systems and learning plausible networks. The successful integration of discriminative and generative methods in the analysis of molecular sequence data (21, 34) augers well for their application to molecular profile data.

This work was supported by the Director, Office of Science, Office of Biological and Environmental Research, Life Sciences Division under US Department of Energy Contract DE-AC03-76SF00098. The data are available upon request.

Present addresses: E. J. Moler, Chiron Corp., 4560 Horton St., Emeryville, CA 94608; M. L. Chow, Gene Logic Inc., 2001 Center St., Berkeley, CA 94704.

REFERENCES

1. **Agata Y, Matsuda E, and Shimizu A.** Two novel Krüppel-associated box-containing zinc-finger proteins, KRAZ1 and KRAZ2, repress transcription through functional interaction with the corepressor KAP-1 (TIF1/KRIP-1). *J Biol Chem* 274: 16412–16422, 1999.

2. **Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, and Levine AJ.** 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, 1999.
3. **Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, and Sherlock G.** Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Gen* 25: 25–29, 2000.
4. **Bendik I, Schraml P, and Ludwig CU.** Characterization of MAST9/Hevin, a SPARC-like protein, that is down-regulated in non-small cell lung cancer. *Cancer Res* 58: 626–629, 1998.
5. **Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Ares M Jr, and Haussler D.** Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Natl Acad Sci USA* 97: 262–267, 2000.
6. **Bult CJ, Krupke DM, Sundberg JP, and Eppig JT.** Mouse tumor biology database (mtb): enhancements and current status. *Nucleic Acids Res* 28: 112–114, 2000.
7. **Cheeseman P and Stutz J.** Bayesian classification (AutoClass): theory and results. In: *Advances in Knowledge Discovery and Data Mining,* edited by Fayyad UM, Piatetsky-Shapiro G, Smyth P, and Uthurusamy R. AAAI Press/MIT Press, 1996. [The software is available at http://ic-www.arc.nasa.gov/ic/projects/bayes-group/autoclass/index.html]
8. **Chow ML, Moler EJ, and Mian IS.** Identifying marker genes in transcription profiling data using a mixture of feature relevance experts. *Physiol Genomics* In press.
9. **D'Haeseleer P, Wen X, Fuhrman S, and Somogyi R.** Linear modeling of mRNA expression levels during CNS development and injury. In: *Pacific Symposium on Biocomputing,* 1999, p. 41–52.
10. **Eisen MB, Spellman PT, Brown PO, and Botstein D.** Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* 95: 14863–14868, 1998.
11. **Epstein CB and Butow RA.** Microarray technology: enhanced versatility, persistent challenge. *Curr Opin Biotechnol* 11: 36–41, 2000.
12. **Franchetti P and Grifantini M.** Nucleoside and non-nucleoside IMP dehydrogenase inhibitors as antitumor and antiviral agents. *Curr Med Chem* 6: 599–614, 1999.
13. **Freije JM, MacDonald NJ, and Steeg PS.** Nm23 and tumour metastasis: basic and translational advances. *Biochem Soc Symp* 63: 261–271, 1998.
14. **Friedman N, Linial M, Nachman I, and Pe'er D.** Using Bayesian networks to analyze expression data [Online]. Stanford University. http://robotics.stanford.edu/people/nir/publications.html [2000].
15. **Gazit G, Lu J, and Lee AS.** Deregulation of GRP stress protein expression in human breast cancer cell lines. *Breast Cancer Res Treatment* 54: 135–146, 1999.
16. **Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov J, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfeld CD, and Lander ES.** Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286: 531–537, 1999. [The data are available at http://waldo.wi.mit.edu/MPR/cancer_class.html]
17. **Goss PJ and Peccoud J.** Quantitative modeling of stochastic systems in molecular biology by using stochastic Petri nets. *Proc Natl Acad Sci USA* 95: 6750–6755, 1998.
18. **Gygi SP, Rochon Y, Franza BR, and Aebersold R.** Correlation between protein and mrna abundance in yeast. *Mol Cell Biol* 19: 1720–1730, 1999.
19. **Hastie T, Tibshirani R, Eisen M, Brown P, Ross D, Scherf U, Weinstein J, Alizadeh A, Staudt L, and Botstein D.** Gene shaving: a new class of clustering methods for expression arrays [Online]. Stanford University. http://www-stat.stanford.edu/~hastie/Papers/ [2000].
20. **Hsu PI, Chow NH, Lin XZ, Yang HB, Chan SH, and Lee PC.** Prognostic significance of ferritin expression in colorectal adenocarcinoma. *Anticancer Res* 15: 1087–1093, 1995.
21. **Jaakkola TS, Diekhans M, and Haussler D.** A discriminative framework for detecting remote protein homologies. *J Computational Biol* 7: 95–114, 2000.
23. **Jensen VF.** *An Introduction to Bayesian Networks.* London: UCL Press, 1996.
24. **Joachims T.** Making large-scale SVM learning practical. In: *Advances in Kernel Methods: Support Vector Learning,* edited by Schölkopf B, Burges C, and Smola A. MIT Press, 1999. [The software is available at http://www-ai.cs.uni-dortmund.de/SOFTWARE/SVM_LIGHT/svm_light.eng.html]
25. **Jordan MI (Editor).** *Learning in Graphical Models.* Dordrecht, Netherlands: Kluwer Academic, 1998.
26. **Kauffman S.** *The Origins of Order. Self-organization and Selection in Evolution.* Oxford: Oxford University Press, 1993.
27. **Kim WH, Lee BL, Jun SH, Song SY, and Kleinman HK.** Expression of 32/67-kDa laminin receptor in laminin adhesion-selected human colon cancer cell lines. *Br J Cancer* 77: 15–20, 1998.
28. **Kitayama J, Nagawa H, Tsuno N, Osada T, Hatano K, Sunami E, Saito H, and Muto T.** Laminin mediates tethering and spreading of colon cancer cells in physiological shear flow. *Br J Cancer* 80: 1927–1934, 1999.
29. **Lee JG, Sahagun G, Oehlke MA, and Lieberman DA.** Serious gastrointestinal pathology found in patients with serum ferritin values $\leq$ 50 ng/ml. *Am J Gastroenterol* 93: 772–776, 1998.
30. **Liang S, Fuhrman S, and Somogyi R.** Reveal, a general reverse engineering algorithm for inference of genetic network architectures. In: *Pacific Symposium on Biocomputing,* 1998, p. 18–29.
32. **Matsuno H, Doi A, Nagasaki M, and Miyano S.** Hybrid Petri net representation of gene regulatory network. In: *Pacific Symposium on Biocomputing*, 2000, vol. 5, p. 338–349.
33. **Mewes HW, Heumann K, Kaps A, Mayer K, Pfeiffer F, Stocker S, and Frishman D.** MIPS: a database for genomes and protein sequences. *Nucleic Acids Res* 27: 44–48, 1999.
34. **Mian IS and Dubchak I.** Representing, and reasoning about protein families: combining generative and discriminative methods derived from different projections of a family. *J Computational Biol* In press.
35. **Moler EJ, Radisky DC, and Mian IS.** Integrating naive Bayes models with external knowledge to examine copper and iron ion homeostasis in *S. cerevisiae*. *Physiol Genomics* 4: 127–135, 2000.
36. **Murphy K and Mian IS.** Modelling gene expression data using dynamic Bayesian networks [Online]. University of California, Berkeley. http://www.cs.berkeley.edu/~murphyk/publ.html [1999].
37. **Navarro E, Espinosa L, Adell T, Tora M, Berrozpe G, and Real FX.** Expressed sequence tag (EST) phenotyping of HT-29 cells: cloning of ser/thr protein kinase EMK1, kinesin KIF3B, and of transcripts that include Alu repeated elements. *Biochim Biophys Acta* 1450: 254–264, 1999.
38. **Nelson PS, Plymate SR, Wang K, True LD, Ware JL, Gan L, Liu AY, and Hood L.** Hevin, an antiadhesive extracellular matrix protein, is down-regulated in metastatic prostate adenocarcinoma. *Cancer Res* 58: 232–236, 1998.
39. **Nicchitta CV.** Biochemical, cell biological and immunological issues surrounding the endoplasmic reticulum chaperone grp94/gp96. *Curr Opin Immunol* 10: 103–109, 1998.
40. **Pearl J.** *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann, 1988.
41. **Prasad SC, Soldatenkov VA, Kuettel MR, Thraves PJ, Zou X, and Dritschilo A.** Protein changes associated with ionizing radiation-induced apoptosis in human prostate epithelial tumor cells. *Electrophoresis* 20: 1065–1074, 1999.
42. **Raychaudhuri R, Stuart JM, and Altman RB.** Principal components analysis to summarize microarray experiments: application to sporulation time series. In: *Pacific Symposium on Biocomputing,* 2000, vol. 5, p. 452–463.
43. **Ryu DD and Nam DH.** Recent progress in biomolecular engineering. *Biotechnol Progress* 16: 2–16, 2000.
44. **Sanchez JC, Schaller D, Ravier F, Golaz O, Jaccoud S, Belet M, Wilkins MR, James R, Deshusses J, and Hoch-**

strasser D. Translationally controlled tumor protein: a protein identified in several nontumoral cells including erythrocytes. *Electrophoresis* 18: 150–155, 1997.
45. **Somogyi R and Sniegoski CA.** Modeling the complexity of genetic networks: understanding multigenetic and pleiotrophic regulation. *Complexity* 1: 45–63, 1996.
46. **Sotiropoulos A, Gineitis D, Copeland J, and Treisman R.** Signal-regulated activation of serum response factor is mediated by changes in actin dynamics. *Cell* 98: 159–169, 1999.
47. **Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, and Futcher B.** Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol Biol Cell* 9: 3273–3297, 1998. [The data are available at http://cellcycle-www.stanford.edu]
48. **Streit M, Schmidt R, Hilgenfeld RU, Thiel E, and Kreuser ED.** Adhesion receptors in malignant transformation and dissemination of gastrointestinal tumors. *Rec Results Cancer Res* 142: 19–50, 1996.
49. **Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, and Golub TR.** Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci USA* 96: 2907–2912, 1999.
50. **Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, and Church GM.** Systematic determination of genetic network architecture. *Nat Genet* 22: 281–285, 1999.
51. **Toronen P, Kolehmainen M, Wong G, and Castren E.** Analysis of gene expression data using self-organizing maps. *FEBS Lett* 451: 142–146, 1999.
52. **Vapnik V.** *Statistical Learning Theory*. New York: Wiley, 1998.
53. **Weaver DC, Workman CT, and Stormo GD.** Modeling regulatory networks with weight matrices. In: *Pacific Symposium on Biocomputing,* 1999, p. 112–123.
54. **Williams K, Chubb C, Huberman E, and Giometti CS.** Analysis of differential protein expression in normal and neoplastic human breast epithelial cell lines. *Electrophoresis* 19: 333–343, 1998.