



ACADEMIC  
PRESS

Available online at [www.sciencedirect.com](http://www.sciencedirect.com)

SCIENCE @ DIRECT®

Journal of Biomedical Informatics 35 (2002) 111–122

Journal of  
Biomedical  
Informatics

[www.academicpress.com](http://www.academicpress.com)

## Characteristic attributes in cancer microarrays

I.N. Sarkar,<sup>a,1</sup> P.J. Planet,<sup>b,1</sup> T.E. Bael,<sup>c</sup> S.E. Stanley,<sup>d</sup>  
M. Siddall,<sup>e</sup> R. DeSalle,<sup>e,\*</sup> and D.H. Figurski<sup>b</sup>

<sup>a</sup> Department of Medical Informatics, College of Physicians and Surgeons, Columbia University, New York, NY 10032, USA

<sup>b</sup> Department of Microbiology, College of Physicians and Surgeons, Columbia University, New York, NY 10032, USA

<sup>c</sup> Department of Internal Medicine, Columbia-Presbyterian Medical Center, New York, NY 10032, USA

<sup>d</sup> Genaissance Pharmaceuticals, New Haven, CT 06511, USA

<sup>e</sup> Division of Invertebrate Zoology, American Museum of Natural History, Central Park West & 79th Street, New York, NY 10024, USA

Received 27 February 2002

### Abstract

Rapid advances in genome sequencing and gene expression microarray technologies are providing unprecedented opportunities to identify specific genes involved in complex biological processes, such as development, signal transduction, and disease. The vast amount of data generated by these technologies has presented new challenges in bioinformatics. To help organize and interpret microarray data, new and efficient computational methods are needed to: (1) distinguish accurately between different biological or clinical categories (e.g., malignant vs. benign), and (2) identify specific genes that play a role in determining those categories. Here we present a novel and simple method that exhaustively scans microarray data for unambiguous gene expression patterns. Such patterns of data can be used as the basis for classification into biological or clinical categories. The method, termed the Characteristic Attribute Organization System (CAOS), is derived from fundamental precepts in systematic biology. In CAOS we define two types of characteristic attributes ('pure' and 'private') that may exist in gene expression microarray data. We also consider additional attributes ('compound') that are composed of expression states of more than one gene that are not characteristic on their own. CAOS was tested on three well-known cancer DNA microarray data sets for its ability to classify new microarray samples. We found CAOS to be a highly accurate and robust class prediction technique. In addition, CAOS identified specific genes, not emphasized in other analyses, that may be crucial to the biology of certain types of cancer. The success of CAOS in this study has significant implications for basic research and the future development of reliable methods for clinical diagnostic tools.

© 2002 Elsevier Science (USA). All rights reserved.

**Keywords:** Cladistics; Expression profile; Pattern recognition; T-cell; B-cell; Melanoma; Colon cancer; ALL; AML

### 1. Introduction

Gene expression profiles generated with DNA microarray technology yield immense quantities of data. Access to the wealth of valuable information hidden within these data requires computationally intensive methodologies for organization and interpretation of results.

Current methods used to search for previously unrecognized order ("class discovery" sensu Golub et al. [23]) in microarray data sets most commonly rely on

measurements of overall similarity or correlation to find natural groupings of microarray samples and genes [1,2,6,16,21,23,44]. Such techniques can organize genes or microarray samples either as a partitioned constellation of unrelated groups [6,41,44] or as a hierarchy of related individuals and groups [1,2,16,21]. The latter has often been depicted as a branching diagram or tree by using established methods from phylogenetics and systematic biology. We recently argued for the use of other systematic/cladistic methods of hierarchic organization based on analysis of discrete attributes as an alternative to measurements of overall similarity [36].

Another important objective of microarray analysis is to identify particular gene expression patterns that may contribute to the biology of a class. This goal has been

\* Corresponding author. Fax: 1-212-769-5277.

E-mail address: [desalle@amnh.org](mailto:desalle@amnh.org) (R. DeSalle).

<sup>1</sup> These authors contributed equally to this work.

pursued using multidimensional scaling plot analysis [7], correlation with idealized expression patterns [23], coupled two-way analysis [21], measurements of misclassification potential [5], artificial neural networks [26], support vector machines [9,50], principle component analysis/singular value decomposition [3,11,37], and correspondence analysis [19]. Due to the high dimensionality of the data, efforts have been made to reduce the computational intensity of analyses using either heuristic algorithms that optimize classification [50] or by using statistical projection techniques to reduce the total number of variables in the data (e.g., Principle Component Analysis/Singular Value Decomposition [3,37], Correspondence Analysis [19]). Gene expression patterns identified by these methods serve a double purpose. First, they provide starting points for future research that aims to elucidate relevant molecular pathways. Second, they can be used as diagnostic predictors to classify new microarray samples.

Here we present a novel, computationally tractable, and exhaustive approach to interpreting gene expression microarray data that is based on the simple concept that members of a particular group may share attributes that do not occur in other groups. This method, designated the Characteristic Attribute Organization System (CAOS), searches an entire data set for all discrete gene features that unambiguously characterize individual microarray samples. CAOS is based on the fundamental idea of parsimony—i.e., the best answer is the one that accounts for all of the data while making the least number of additional (ad hoc) hypotheses. A closely related technique from systematic biology, Population Aggregation Analysis (PAA; [12]), is used in conservation biology to find patterns of discrete features (attributes) that unambiguously distinguish all individuals in a population from members of other groups. Gene expression levels can be viewed as the attributes of individual DNA microarray samples. We designed CAOS to search large microarray data sets for “characteristic attributes” that are then used as diagnostic standards to classify new microarray samples.

We show that the characteristic attributes derived from even a very simple version of CAOS can be used as an accurate and robust diagnostic tool for class prediction. Analysis of three well-known cancer microarray data sets reveals the potential for CAOS to be a powerful tool. Additionally, CAOS identified potentially

relevant genes with diagnostic expression patterns that were not emphasized by other techniques.

## 2. Materials and methods

*Initial classification of microarray data sets.* The microarray data sets used to test CAOS have previously been described [2,7,23] (Table 1). The Bittner et al. [7] data set was grouped into the two classes of malignant melanomas (major and minor) elucidated in that study by phenetic clustering techniques. The Golub et al. [23] data set was divided first into acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) as determined in that study. The Alon et al. [2] data were separated into normal and colon cancer.

*Data matrices and recoding.* Continuous gene expression values in the microarray data matrices were recoded as discrete values using a binning approach [30,36]. A bin is defined as a range of scaled intensity values into which each data point could fall. We designated two bins separated by the median value for a particular gene over all microarray samples. Values above the median were recoded as the value “1;” and those below, as “0.” Other more complex binning strategies are possible [30], and we are currently investigating various techniques, as well as estimations of the amount of data that is lost during the binning process.

*Characteristic attributes.* The new matrix of recoded data was then searched for binned gene expression states (“attributes”) that unambiguously distinguished members of different classes from each other (“characteristic attributes”). Characteristic attributes are defined as binned expression values (i.e., ‘1’ or ‘0’) for a particular gene that occur in microarray samples of one class and are never observed outside that class. “Pure” characteristic attributes are found in every member of a group; “private” characteristic attributes are found only in some members of a group. Using CAOS-ag99 (“CAOS-aggregator 1999”), an application written in C++, each gene was evaluated on the basis of its gene expression states as a possible characteristic attribute. CAOS-ag99 finds and records binned expression values that only appear in one of the two classes and assigns such characteristic attributes a weighted rank ( $R_w$ ) (Fig. 2A).  $R_w$  is the ratio of the number of samples in the data set that exhibit the characteristic attribute ( $S_E$ ) to the total

Table 1  
Numbers of characteristic attributes for the three data sets

Dataset	Bittner et al. (275 genes)				Alon et al. (2000 genes)				Golub et al. (7129 genes)			
	Minor [12]		Major [19]		Cancer [40]		Normal [22]		ALL [27]		AML [11]	
	Pure	Private	Pure	Private	Pure	Private	Pure	Private	Pure	Private	Pure	Private
Simple	0	75	0	5	0	84	0	11	1	885	1	80
Compound	33	6028	12	1529	0	64,416	2	17,560	602	1,592,211	1680	177,127

number of samples in the group ( $S_G$ ) (Eq. (1)). This value is then scaled to 10 so that it is comparable to other ranking systems. An  $R_w$  of 10 indicates a pure characteristic attribute, whereas values less than 10 indicate private characteristic attributes with different relative abundances

$$R_w = \left( \frac{S_E}{S_G} \right) \times 10. \quad (1)$$

CAOS-ag99 also searches for combinations of bin values that occur *together* only in members of a predefined group or class (“compound” characteristic attributes). In this procedure genes are combined in the matrix to create a new set of attributes that is the combination of the bin values coded as 00, 01, 10, and 11. The theoretical upper limit ( $U$ ) for the number of compound characteristic attributes (Eq. (2)) is calculated from the total number of genes in the microarray ( $N$ ) and the order ( $D$ ) of the compound attribute, which is defined as the integer number of genes that are combined to make the compound attribute

$$U = \frac{N!}{D!(N-D)!} \quad \text{such that, } D > N(2). \quad (2)$$

By definition, the combination of any simple characteristic attribute with any other attribute will always yield a compound characteristic attribute. Because such compound attributes yield no new information, we excluded them from this analysis by requiring that each compound characteristic attribute be composed of only those attributes that were not characteristic at any lower order. For example, all simple (first-order compound) characteristic attributes were excluded from the identification of second-order compound attributes. Weighted ranks ( $R_w$ ) were determined for compound attributes with the same algorithm used for simple attributes.

*Class prediction/voting.* Characteristic attributes (pure and private; simple and compound) identified by CAOS-ag99 were then used to create a set of diagnostic rules to classify microarray samples (Fig. 2B). The diagnostic rule set was constructed and implemented using CAOS-rr01 (Characteristic Attribute Organization System—rule reader 2001’), a script written in Perl5. The script first tests unclassified samples for the presence or absence of the listed characteristic attributes.

CAOS-rr01 then sums the  $R_w$  values for every characteristic attribute found in the unclassified sample (foundRw). This sum is compared to the sum of all possible  $R_w$  values (possibleRw) for each particular group ( $G$ ) to eliminate bias towards groups with greater numbers of characteristic attributes. This ratio,  $A$ , is the adjusted score (Eq. (3))

$$A_G = \frac{\sum \text{foundRw}}{\sum \text{possibleRw}}. \quad (3)$$

The unclassified sample is classified into the group for which the  $A$  score is higher. When the  $A$  score is within 5% of the other score(s) (e.g.,  $A_1/A_2 < 0.95$ ), the classification is regarded as ambiguous.

*Robustness.* To test the robustness of CAOS, we used a Leave-One-Out-Cross-Validation (LOOCV) jackknifing technique similar to that presented by Golub et al. [23]. A single microarray sample was removed from the data set, and the remaining samples were searched with CAOS-ag99 to identify characteristic attributes. Using diagnostic rules generated by CAOS-rr01, we attempted to classify the withheld sample. This procedure was repeated for every sample in the data set, and the result was monitored for correct or incorrect classification.

Standard measurements of classification were calculated for each iteration. True Positive (TP) values were calculated as the number of samples correctly assigned to the group to which each belongs (e.g., if examining AML, AML sample being classified as AML). True Negative (TN) values were the number of samples correctly identified as not belonging to the group including those classified as ambiguous (e.g., if examining AML, the number of ALL classified as ALL). False Positive (FP) values were the number of samples incorrectly assigned to a group (e.g., if examining AML, ALL being classified as AML). False Negative (FN) values were the number of samples that were incorrectly placed outside of the group including those classified as ambiguous (e.g., if examining AML, AML being classified as ALL).

The overall *accuracy* of CAOS-based classifications was calculated as the total number of correctly classified samples (TPs for both group 1 and 2) divided by the total number of samples including those regarded as ambiguous. *Sensitivity* is  $TP/(TP+FN)$ . *Specificity* is  $TN/(TN+FP)$ . The *ambiguity* of all the classifications was calculated as the percentage of classifications that were not classifiable (either when  $A_1/A_2 < 0.95$  or when  $A_1 = A_2 = 0$ ).

An important feature of CAOS is that all possible characteristic attributes are considered regardless of rank. Even low-ranking characteristic attributes may be needed to identify members of subsets with few representatives in the original (training) data set. Ideally all characteristic attributes would be used, but we found that we could increase the accuracy of diagnosis in some cases by excluding some characteristic attributes, suggesting insignificant variation in the sample data—i.e., “noise.” To attempt to filter out noise while retaining valuable low ranking private characteristic attributes, we introduced the concept of corroboration, which simply requires a characteristic attribute to be supported by its appearance in a specified number of other samples before it is included in the diagnostic rules. The number of samples required to corroborate a characteristic attribute (corroboration coefficient; Crob) can be

increased to create very well corroborated but potentially less sensitive diagnostic rule sets, or decreased to create potentially noisy (perhaps nonspecific) rule sets that should be more sensitive.

To maximize sensitivity with the least amount of corroboration, we repeated the LOOCV jackknifing procedure while excluding characteristic attributes that fell below a series of “corroboration rank cutoffs.” Corroboration rank cutoffs (CRC) were determined by adjusting the corroboration coefficient (Crob) according to the number of samples in a particular class ( $S_G$ ) (Eq. (4))

$$CRC = Crob \times \left( \frac{10}{S_G} \right). \quad (4)$$

Crob was varied as an integer from 1 to  $q - 1$  (where  $q$  is the number of samples in the smaller class). The “optimal corroboration rank cutoff values” (OCRC) were then determined as the value at which the most microarray samples were accurately diagnosed with the least amount of corroboration (lowest Crob) required.

Xiong et al. [50] have recently demonstrated that more versatile heuristic algorithms can be used to maximize classification accuracy. Significantly, they have reported classification accuracy levels higher than those reported here for the Alon et al. data set. We are currently exploring similar optimizing algorithms that could be applied to CAOS classification.

**Receiver Operating Characteristic (ROC) curves.** ROC curves were determined from the efficacy of CAOS-based analysis on the entire data set using varying training set sizes. Training set sizes were 3,  $q/2$ , and  $q - 1$  samples from each group ( $q$  is the number of elements in the smaller group). We performed 200 replicates at each training set size value. The specificity and sensitivity of each test was determined and plotted for each replicate. We generated ROC curves using the SPSS statistical package [42].

**Testing an independent data set.** The Golub et al. training data set was used to determine a set of characteristic attributes to distinguish between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). The LOOCV jackknifing technique, as described above, was performed at each corroboration coefficient. The optimal corroboration coefficient (OCRC) was determined as above to exclude non-optimal characteristic attributes. CAOS-rr01 was then used to classify an independent set of 34 samples as provided in the supplementary data of Golub et al.

**Comparison of potentially biologically relevant attributes.** We compared our CAOS-based rankings to the prediction strengths of genes identified by the three cancer studies analyzed here. The “prediction strengths” from the Golub et al. and Alon et al. data sets were normalized to a scale of 0–10. To determine the statis-

tical similarity of the sets of ranked genes, we used a Wilcoxon Signed Rank Test, which evaluates the statistical significance of the difference between two ordered lists [42].

### 3. Results

#### 3.1. Characteristic attributes—pure, private, simple, and compound

Individual members of any group may share attributes that unambiguously distinguish them from members of other groups. We refer to these attributes as “characteristic attributes.” Here we consider the scaled intensities of hybridization (i.e., gene expression levels) as the attributes of individual microarrays. Our approach defines distinct patterns of gene expression states that unambiguously characterize members of a specific group.

We defined two different types of characteristic attributes: “pure” and “private” (Fig. 1). Pure characteristic attributes are present in every member of a specified class and *absent* from every individual outside the class (e.g., a gene that is on in every cancerous cell and off in every normal cell). These are the most sought after genes in any method. Because such genes are extremely rare, we also defined private characteristic attributes, which

	sPu (1)		sPr (0)		cPu (1,1) or (0,0)		cPr (1,0)												
1	1	1	1	0	1	1	1	1	0	1	1	1	1	1	1	1			
2	1	1	0	0	1	1	0	1	0	0	0	0	0	0	1	1	0		
3	1	0	1	0	1	1	1	0	1	1	1	1	0	1	0	1	0	0	
4	1	1	0	0	1	1	0	1	0	0	1	0	1	0	1	0	1	1	
5	1	1	1	0	1	1	1	1	1	0	0	1	0	1	0	1	1	0	
6	1	0	0	0	1	1	0	0	0	1	1	1	0	0	1	0	1	0	0
7	0	1	1	0	1	1	0	1	1	0	1	0	1	0	1	0	1	1	0
8	0	0	1	0	0	1	0	0	1	1	0	1	1	0	0	0	1	0	0
9	0	1	1	0	1	1	1	1	0	0	1	0	1	1	0	1	0	1	1
10	0	1	1	0	0	1	0	1	1	0	0	0	1	0	1	0	1	1	0
11	0	0	1	0	1	1	1	0	0	1	1	1	1	0	1	0	0	0	0
12	0	1	1	0	0	1	0	1	1	0	0	1	1	0	1	1	1	1	1
	sPu (0)		sPr (0)		cPu (1,0) or (0,1)		cPr (1,0)												

Fig. 1. Four types of characteristic attributes. Characteristic attributes are depicted in a hypothetical alignment of binned microarray data. Each row represents a single microarray sample. Each column represents a particular gene. Simple attributes are taken from only one column. Simple Pure (sPu) characteristic attributes are binned gene expression values (0 or 1) that are found in all members of one group and never outside that group. Simple Private (sPr) characteristic attributes are found in some members of one group and never outside that group. Compound characteristic attributes are found by assessing each possible combination of two genes. Compound Pure (cPu) characteristics are combinations of gene expression values (11, 01, 10, or 00) that are found in all members of a group and are never found together in samples outside that group. Compound Private (cPr) characteristic attributes are combinations of gene expression values that are found in some samples in one group and never outside that group.

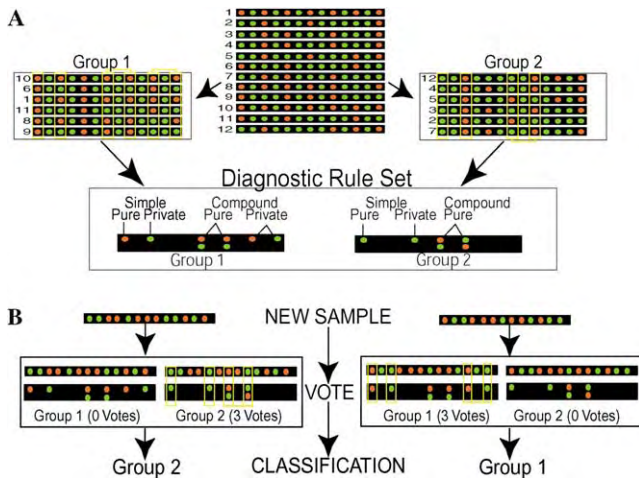


Fig. 2. CAOS is a two-step process. Panel A: Finding characteristic attributes in microarray data. Data are shown as red and green spots representative of increased expression or decreased expression results, respectively. These data are first grouped using some a priori criterion. Next, characteristic attributes are determined by CAOS-ag99. These characteristic attributes form the basis for a set of diagnostic rules. Panel B: Diagnostic rules can then be used to classify novel samples by a voting process in which the new sample is placed in the group for which it has the highest vote total.

are present in some, but not all, members of a class, and *absent* from every individual outside that class. If a sample has either type of characteristic attribute it can be unambiguously classified. Our analysis discards attributes that occur in both groups.

We differentiated between “simple” characteristic attributes and “compound” characteristic attributes (Fig. 1). Compound characteristic attributes are formed when two or more simple attributes are associated with

each other only in one group and never outside that group. This idea expands the concept of microarray attributes to include combinations of gene expression states. It allows for the inclusion of a new class of characteristic attributes composed of multiple attributes that alone are *not* characteristic—i.e., we excluded any attributes found to be individually characteristic from participating in compound characteristic attributes. In this preliminary study, we limited our analysis to compound characteristic attributes formed from two individual attributes only. In theory, compound characteristic attributes could consist of three or more attributes. Like simple characteristic attributes, compound characteristic attributes can be either “pure” or “private.”

We wrote a computer application (CAOS-ag99) that searches microarray data sets for characteristic attributes and assigns ranks based on the percentage of microarray samples that share that particular attribute. Three well-studied cancer DNA microarray data sets were used to test the effectiveness of CAOS. CAOS-ag99 identified all the types of characteristic attributes that we defined. All applications were built and executed on a Dual-Processor 400 MHz Sun Enterprise 3000 Server with 1 GB of System Memory running SunOS 5.8. The application was able to perform exhaustive searches for simple characteristic attributes in approximately 1 min (Bittner et al. data set) to about 10 min (Golub et al. data set). Searches for compound attributes varied from 10 min (Bittner et al, data set) to approximately 3 h (Golub, et al. data set).

While simple pure characteristic attributes are rare or absent, compound pure characteristic attributes are

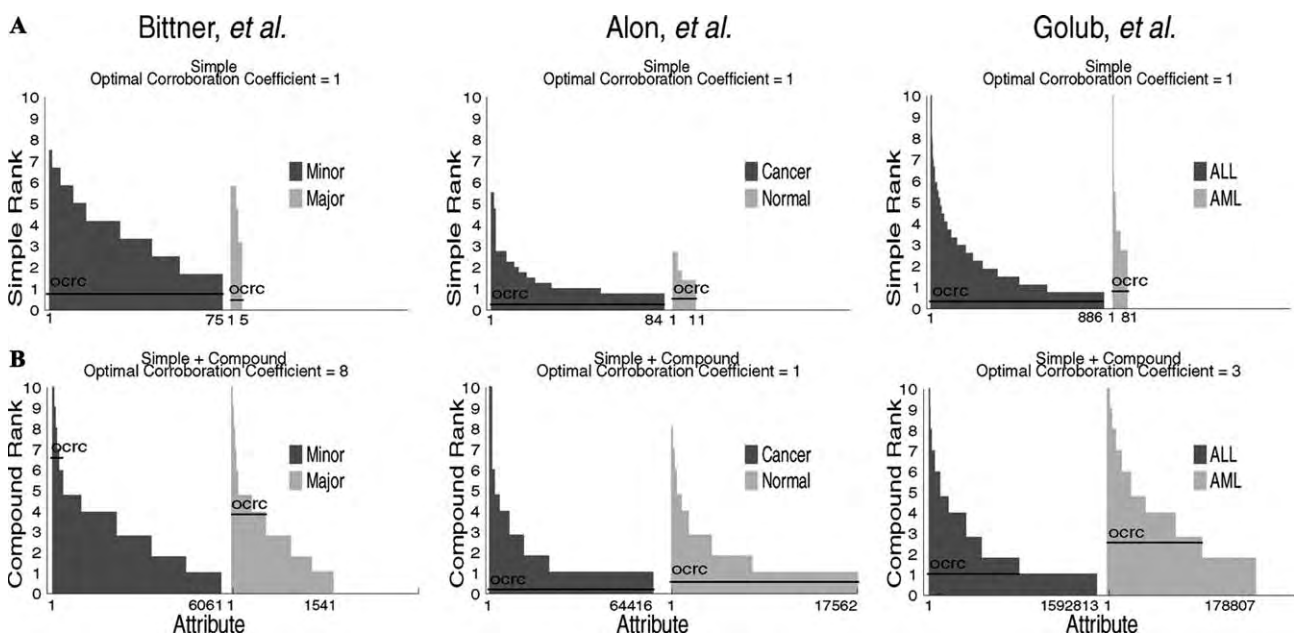


Fig. 3. Distribution of ranked genes for both simple (A) and compound (B) characteristic attributes. The histograms plot rank values ( $R_w$ ) for each characteristic attribute in each group, in decreasing order. Horizontal lines indicate the optimal corroboration rank cutoff values (OCRC).



relatively frequent (Table 1). CAOS identified compound pure characteristic attributes in at least one group from each of the three data sets, whereas simple pure characteristic attributes were found only in the Golub et al. data set. In addition, rank distributions and the total number of characteristic attributes were different from one group to another and from one data set to another (Fig. 3). Such differences would be expected when comparing groups with differing levels of homogeneity. For instance, heterogeneous groupings composed of distinct subgroups would be expected to yield many private characteristic attributes with few pure characteristic attributes. Differences in the total number of characteristic attributes would be expected depending on the overall similarity of the groups being compared. Two very similar types of cancer might yield few distinguishing characteristic attributes, whereas two very different types may yield many.

### 3.2. Classification with characteristic attributes

We hypothesized that an optimal classification scheme would be based on the least ambiguous set of criteria possible. Characteristic attributes, because they represent unambiguous distributions of attribute data, should be ideal as the basis for such a classifier. By definition, rules derived from the training set will always be 100% specific with regard to the samples in the training set—i.e., they will never misclassify a sample. Also, if at least one characteristic attribute exists for every sample, then CAOS-based diagnosis will always result in the classification of all samples in the training set with 100% sensitivity. This was the case in all three data sets studied here. Importantly, diagnostic rules from compound characteristic attributes alone and the combination of simple and compound characteristic attributes always classified samples from the training set with 100% sensitivity. Likewise, rules from simple characteristic attributes alone were able to classify with 100% sensitivity in the Golub et al. and Bittner et al. data sets. In the Alon et al. data set, some samples contained no simple characteristic attributes at all, giving a sensitivity of 94%. These samples relied entirely on compound characteristics for classification.

To test the robustness of CAOS-based diagnosis and its efficacy on samples drawn from outside the training data set, we used a LOOCV jackknifing technique (Table 2) and found very high levels of accuracy. By excluding characteristic attributes with ranks below certain optimized thresholds (OCRCs), we found that we could increase the overall accuracy of diagnosis. To find the optimal corroboration rank cutoff value, we did multiple LOOCV replicates and varied the number of samples that were required to possess, and therefore corroborate, a particular characteristic attribute in order for it to be kept in the set of diagnostic rules. Corro-

ration rank cutoffs were optimized for each data set by maximizing accuracy and then minimizing the amount of corroboration and ambiguous classification (Fig. 3).

We tested the diagnostic accuracy of CAOS-based classification using the independent sample test set presented in Golub et al. [23]. In that study, the prediction technique confidently classified 29 of 34 samples, resulting in 85% sensitivity with 100% specificity. In comparison, CAOS-based diagnostics were able to classify all 34 samples, to yield 100% sensitivity with 100% specificity.

We hypothesized that the specificity and sensitivity of CAOS-based diagnosis would increase as more samples were included in training sets to construct diagnostic rules. To test this idea, we used a bootstrapping technique to gauge the success of classification in each data set using randomly chosen training subsets of different sizes. ROC curves showed a consistent trend towards higher specificity and sensitivity with larger initial data training sets (Fig. 4). These results indicate that diagnostic rule sets can be improved by increasing the size of the training set. Furthermore, the results suggest that use of very large training sets will allow CAOS to create highly reliable diagnostic rule sets for clinical diagnosis.

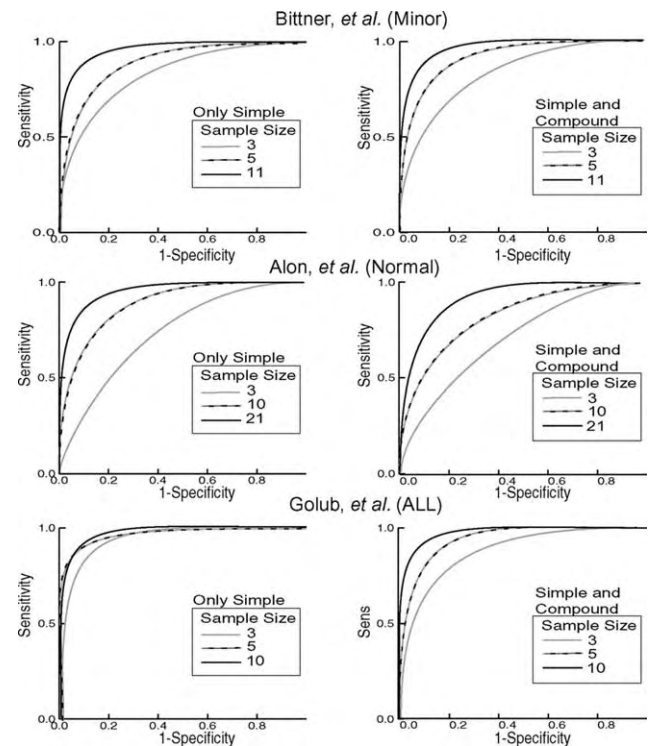


Fig. 4. Effect of training set size on sensitivity and specificity. Shown are ROC curves that were generated, at three different sample sizes, using a bootstrap methodology (see Section 2). The positive group for sensitivity and specificity calculations is noted in parentheses. Simple characteristic attributes alone and simple and compound characteristic attributes in combination were used for classification of each of the data sets. Note that there is a general tendency toward higher sensitivity and specificity as the training sample size is increased.

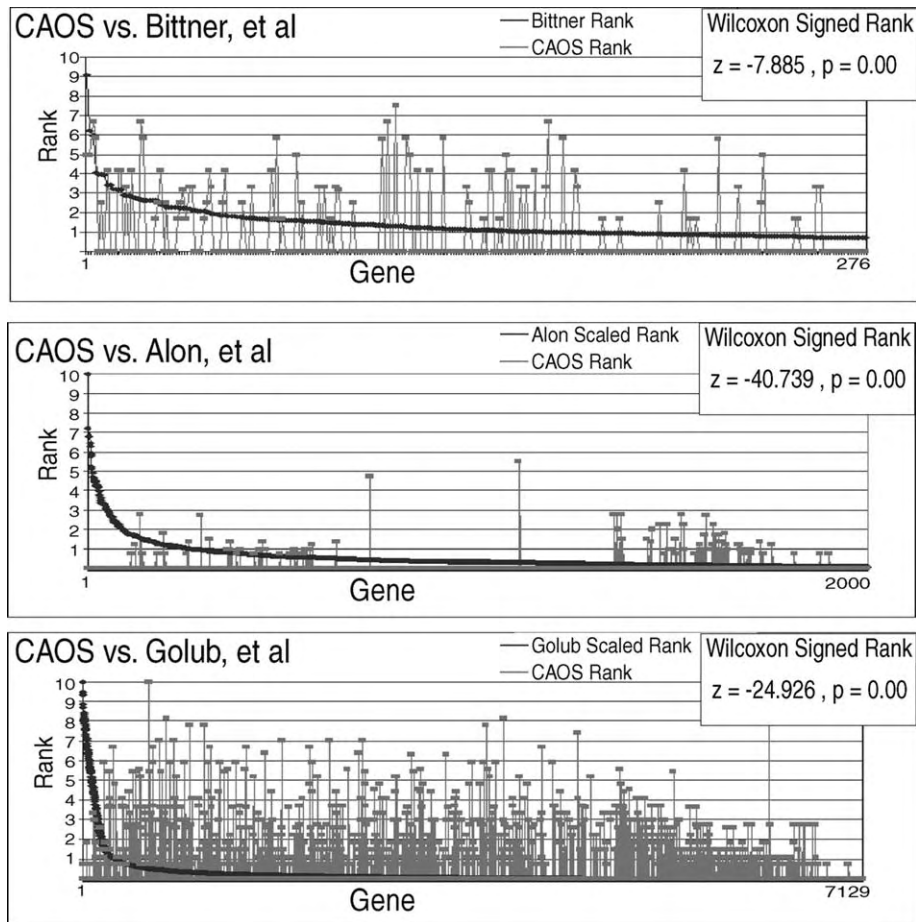


Fig. 5. Comparison of genes ranked by CAOS and other methods. Graphs depict each ranked gene from the three previous studies [2,7,23] in order from highest to lowest rank (blue) compared to the rank of the same genes found using CAOS (pink). The  $z$ -score and  $p$ -value, as reported by a Wilcoxon Signed Rank Test, are shown in the boxes.

Because of the simplicity of this algorithm, such analysis is easy to accomplish.

### 3.3. Comparison of gene identification and ranking

Does CAOS identify genes that are missed by other techniques? To address this question, we compared rank distributions of the characteristic genes and expression levels found by CAOS to those found by other techniques (Fig. 5). In every case, rank distributions were clearly different. Several genes ranked very low by other techniques were given very high ranks by CAOS and vice versa. A Wilcoxon Signed Rank Test validated this disparity as statistically significant (Fig. 5).

Because characteristic attributes are by definition unique to a category, the genes they represent in microarray samples may have functional relevance to the nature of the category. We examined our highest ranked genes to determine if there was any correlation with the cancers being diagnosed. Several of the genes are known or suggested to play a role in cancer biology including CD45, tenascin C, Annexin II, MART-1, IL-6, NCAM,

4-Ptase II, DGCR6, CD9, and INF- $\alpha$  among others (see below). Many of the high-ranking characteristic attributes are ESTs from genes whose functions are unknown.

## 4. Discussion

Characteristic attributes are commonly used in medical diagnosis. A patient's specific signs and symptoms, or combinations thereof, are routinely evaluated to identify a disease and determine prognosis. Microarray data present an opportunity to expand this type of analysis to include many gene expression events as signs of a disease. In fact, use of microarray technology is being seriously considered as a viable option for patient diagnosis [38].

We developed CAOS to: (1) create diagnostic rules for classification, and (2) highlight particular features with potential biological relevance. CAOS is an efficient, automated, and exhaustive technique that examines large amounts of data to identify characteristic attributes of a specific group and then uses them to

classify data from new samples (Fig. 2). We applied this technique to cancer gene expression data and found that it performed as an accurate and robust method for classification and analysis. In addition, we found that CAOS consistently identified genes of known importance and others of potential relevance to malignancies.

#### 4.1. Development of the CAOS idea

CAOS classification is based on the idea that a new sample can be assigned to the group in which it causes the least amount of disruption to the observed pattern of unambiguously distributed attributes. Consider an unknown sample *S* that could be classified into group *A* or group *B*. We perform CAOS analysis on *A* vs. *B* and find that there is a set of characteristic attributes that can classify all members of *A* and *B* with 100% sensitivity and specificity. If we then add *S* to *A* and rerun the CAOS analysis, many characteristic attributes are deleted and several others emerge. If we add *S* to *B* and rerun the CAOS analysis only a few characteristic attributes are added and deleted. Therefore, it is least disruptive, or least contradictory, to place *S* in *B*. The fundamental basis of the CAOS classification operation presented in this paper is to place a new sample in the group with which it shares more well corroborated characteristic attributes. CAOS-based classification attempts to minimize the number of ad hoc hypotheses required to place the sequence in one of the two groups—i.e., it is the most parsimonious classification. Thus, CAOS based analysis can be supported, examined, and criticized, by the logic, techniques, and rich literature that support parsimony analysis in systematic biology [17]. Furthermore, CAOS also serves as a starting point for introducing other techniques that have been tested on discrete attributes, such as Maximum Likelihood [20] and Bayesian analysis [28], to microarray studies.

The use of attribute data is routine in evolutionary and conservation biology [10,12,22,46]. Davis and Nixon's [12] PAA method focuses on finding characteristic attributes that distinguish one population of organisms from another. These absolute types of characteristics are what we term as "pure." Pure characteristic attributes in CAOS possess the strongest diagnostic potential because they can account for every member of a group. They are also the most obvious starting points in a search for biological function.

We adopted the term "private" which is used in population genetics to designate attributes that occur only in one group but are not shared by all the members of that group [4,43]. The existence of private characteristic attributes in a group could signal the presence of a subgroup whose members all possess an attribute, or it may simply represent the loss of some constraint that

allows variation in one group and not the other. Hierarchical systematic grouping techniques may be able to distinguish between these two interpretations.

We have extended the idea of pure and private attributes to include characteristic attributes that are formed by combining multiple attributes (compound characteristic attributes). Such attributes can be thought of as associating with each other only in a certain group or as co-varying. This linkage may be indicative of some functional constraint or interaction between the attributes.

To this end we created an application that searched microarray data-sets, using an exhaustive algorithm that examines each binned gene state individually (for simple characteristic attributes) and then in pairs (for compound attributes). While the application used for the present study was not written with the intent for high-speed, the entire analysis of all three of the data sets was completed in less than 12 h on a Dual-Processor 400 MHz Sun Enterprise 3000 Server with 1 Gigabyte of System Memory running SunOS 5.8. This result is significant, as search algorithms and parallel processing systems can be used to substantially increase speed, especially as no computationally intensive calculations are required. Once the diagnostic rule sets are established, it takes less than a minute to classify a sample.

#### 4.2. Cancer DNA microarray data and CAOS

We expect that CAOS will be particularly effective in the classification of cancer microarray data for the following reasons. First, in prognostic or diagnostic categories with large amounts of heterogeneity or overlapping gene expression, simple characteristics may be the most useful class predictors. Second, any set of groupings or categories can be defined for analysis, which makes this supervised technique extremely versatile and applicable both to existing medical categories and those that are newly discovered. Third, CAOS allows detection of characteristic attributes that define subgroups. Private characteristic attributes will be identified even if they exist only in a small percentage of members of a group. Thus, if multiple subtypes of cancer exist in a predefined category, then CAOS would detect each of their characteristic gene expression patterns as private characteristic attributes. Such characteristic attributes would be missed by techniques that depend on overall similarity or on uniformity of gene expression within a group.

Fourth, CAOS has the potential to identify characteristic gene expression patterns that have relevance to cancer. Pure characteristic attributes might represent genes that are specifically induced or repressed. Private characteristic attributes may represent genes that are induced or expressed in a cancer subtype. Alternatively, they may represent genes of the cancer group that are

freed from a regulatory constraint. Compound characteristic attributes have the potential to identify genes that are interacting or co-regulated in certain types of cancer. CAOS therefore has the potential to indicate regulatory cascades and important multigene interactions.

#### 4.3. Biological relevance of characteristic attributes

Conclusions about the biological relevance of specific gene expression patterns identified within large databases must be drawn with caution. The correlation of a gene expression attribute with a specific category may be a byproduct of the laboratory isolation technique, the bioinformatic approach, sample contamination, or simple chance. For example, CD45, a marker for leukocytes such as colonic mucosa-associated T-cells, is found by CAOS to be expressed at relatively higher levels in a small subset of colon cancer specimens in the Alon et al. data set. This may represent a true biologic process or contamination of lymph tissue in some biopsy specimens. Nevertheless, in each of the three data sets analyzed, CAOS readily found genes with characteristic expression attributes, many of which are known or suspected to be important in specific disease processes.

CAOS highlighted some genes not emphasized by other analyses of the same data sets (Fig. 5). One fundamental reason for this difference is that CAOS requires each characteristic attribute to unambiguously classify every sample in which it appears. CAOS will only identify genes that have expression states unique to one group. Thus, CAOS also detects characteristic gene expression states that only occur in a fraction of the cells in a particular class. As a result, CAOS is able to differentiate between characteristic and ambiguous expression patterns even if only a small number of cells have a different expression state.

Bittner et al. [7] used multidimensional scaling plots to identify genes that differentiated between two clusters of biopsied melanomas, referred to as major and minor. This approach identified a subset of genes relating to cell motility that was differentially expressed in the major cluster of biopsy samples. Many of these genes were also given high ranks by CAOS. Tenascin C, which ranked 12th (rank score of 4.2) by Bittner et al. and 18th ( $R_w$  3.2) by CAOS has increased expression in some malignant melanomas and may affect cell motility [45]. Annexin II and MART-1 were both among the top 20 genes ranked by Bittner et al. and ranked highly by CAOS as private characteristic attributes of the minor cluster of biopsy samples. MART-1 is expressed in the majority of melanomas and Annexin II appears to play a role in immune response to melanomas [29,51].

CAOS also identified several genes whose products are known to be important in melanoma biology but were not given high priority in the Bittner et al. analysis.

For example, interleukin 6 (IL-6) is secreted by metastatic melanoma cells [13], and it has been shown to support autocrine or paracrine growth [33]. Serum IL-6 has been associated with disease progression and decreased survival [13]. CAOS found that lower expression of IL-6 was a private characteristic attribute of the major cluster. IL-6 ranks 30th ( $R_w$  4.2) by CAOS and was ranked as the 173rd (score 0.99) gene by Bittner et al.

Another example is the human neural cell adhesion molecule (hNCAM) whose gene has characteristic expression in the minor cluster and was ranked 31st ( $R_w$  4.2) by CAOS and 174th (score 0.99) by Bittner et al. Tumor cell adhesion molecules are known to be associated with development of metastatic behavior in melanoma [25]. Specifically, hNCAM is highly expressed in aggressive uveal melanomas [32].

CAOS identified some genes with a potential role in melanoma. The gene that ranks first ( $R_w$  7.5) in the CAOS analysis and ranked 40th (score 1.3) by Bittner et al. encodes inositol polyphosphate-4-phosphatase, type II (4-Ptase II). This gene is expressed as a private characteristic attribute in the minor cluster. Inositol phospholipids may play an important role in cell signal transduction. A closely related protein, 4-Ptase I, has been shown to play a role in GATA-1 transcription factor regulation [48]. Another example is DGCR6, which has a role in cell migration and motility in the neuronal crest cells [14] the progenitors of melanocytes. The DGCR6 gene is ranked 26 ( $R_w$  4.2) by CAOS and 142nd (score 1.1) in Bittner's analysis.

Golub et al. [23], used a statistical method that ranked a gene's importance in distinguishing between AML and ALL based on its correlation with idealized expression patterns (neighborhood analysis) and noted that many of the high ranking genes are important in leukemia biology. CAOS also identified genes for several markers of leukemia lineage, including expression of Ig Dxp heavy chain (a portion of the immunoglobulin heavy chain, IgH) and CD9. IgH clonal rearrangement is characteristic of ALL and it can be found in AML, where it may indicate a poor prognosis [27]. CD9, found in 90% of B-lineage ALL, is also expressed in acute promyelocytic leukemia [24], and acute basophilic leukemia [15].

Other gene expression patterns found to distinguish between ALL and AML may give insight into the biology of leukemia. The interferon  $\alpha$ -21 gene was expressed at a higher level in a subset of ALL samples as compared to AML samples. This may mean that interferon  $\alpha$ -21 gene expression is constrained in all AML samples, but not all ALL samples. Interferon  $\alpha$  (INF- $\alpha$ ) induces cell cycle arrest and apoptosis in chronic myelogenous leukemia (CML) [47]. The constrained expression of the INF- $\alpha$  gene in AML as compared to ALL may indicate that INF- $\alpha$  plays a role in the AML as well as CML. Conversely, the gene for prostaglandin Ep3 Receptor,

Alt. Splice 8 was expressed at a lower level in a subset of ALL samples. The E-series prostaglandins are regulators of B lymphocyte function and have been shown to inhibit growth in some B lymphoma cell lines [18,35].

CAOS also highlighted genes of interest relating to the T-cell/B-cell distinction in ALL. C-yes-1 was expressed at a lower level in some B-cell samples relative to T-cell samples. C-yes-1 is a proto-oncogene mapped to the same chromosome band as bcl-2 (18q21.3) [34], and is expressed in canine lymphoid neoplasm [31]. Several genes for cell adhesion molecules such as Cadherin-6 and cellular adhesion regulatory molecule (CMAR), as well as several known tyrosine kinases, were also found to be expressed as private characteristic attributes in T or B cell ALL.

CAOS-based analysis of the Alon et al. [2] data set revealed several genes with established roles in colon cancer. The P-cadherin gene is expressed at a higher level in a portion of the colon cancer samples. Cadherins are a family of cell-to-cell adhesion molecules known to play an important role in many solid tumors including colon, prostate, breast, and lung cancer. P-cadherin up-regulation may represent early neoplastic transformation in glandular mucosa and adenomatous polyps [40].

Protein kinase C- $\zeta$  type (PKC- $\zeta$ ) has characteristic expression in the colon cancer category. The PKC family of isoenzymes plays an important role in cell signaling and tumor promotion [8]. The PKC- $\zeta$  gene is expressed in azoxymethane-induced tumors treated with nonsteroidal anti-inflammatory drugs in trials using animal tumor models [39]. Other genes highlighted by CAOS in the colon cancer samples include those for HMG-1, which is expressed in gastric and colorectal adenocarcinoma [49], and several tyrosine and serine-threonine protein kinases with as yet poorly defined roles in cancer.

## 5. Significance and future prospects

The results show that CAOS has considerable promise as a new method for microarray analysis and has the potential to identify relevant genes missed by other types of analysis. Even in this simple form, CAOS is a highly accurate method for classification. Because the sensitivity and specificity of CAOS-based diagnosis increases with more information, the use of even larger data sets will allow CAOS to identify diagnostic rules that far outperform those used in this study. Additionally, representation of the data by more than two discrete bins may reveal further expression patterns. By using just two bins, we may have obscured informative gene patterns by missing more subtle expression patterns beyond just 'on' or 'off.' It is our hope that this preliminary examination of the potential of the CAOS algorithm for basic research and clinical diagnosis will stimulate further refinement and development.

## Acknowledgments

We thank Scott Kachlany and the Columbia University DMI-Data Mining group (in particular, George Hripscak). INS is supported by National Library of Medicine Medical Informatics Training Grant LM07079-09. P.J.P. is supported on a Columbia University Training Grant.

## References

- [1] Alizadeh AA, Eisen MB, Davis RE, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling [see comments]. *Nature* 2000;403:503–11.
- [2] Alon U, Barkai N, Notterman DA, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci USA* 1999;96:6745–50.
- [3] Alter O, Brown PO, Botstein D. Singular value decomposition for genome-wide expression data processing and modeling. *Proc Natl Acad Sci USA* 2000;97:10101–6.
- [4] Barton NH, Slatkin M. A quasi-equilibrium theory of the distribution of rare alleles in a subdivided population. *Heredity* 1986;56:409–15.
- [5] Ben-Dor A, Bruhn L, Friedman N, Nachman I, Schummer M, Yakhini Z. Tissue classification with gene expression profiles. *J Comput Biol* 2000;7:559–83.
- [6] Ben-Dor A, Shamir R, Yakhini Z. Clustering gene expression patterns. *J Comput Biol* 1999;6:281–97.
- [7] Bittner M, Meltzer P, Chen Y, et al. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* 2000;406:536–40.
- [8] Blobe GC, Obeid LM, Hannun YA. Regulation of protein kinase C and role in cancer biology. *Cancer Metastasis Rev* 1994;13:411–31.
- [9] Brown MP, Grundy WN, Lin D, et al. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Natl Acad Sci USA* 2000;97:262–7.
- [10] Cracraft J. Species concept and speciation analysis. *Curr Ornithol* 1983:159–87.
- [11] Crescenzi M, Giuliani A. The main biological determinants of tumor line taxonomy elucidated by a principal component analysis of microarray data. *FEBS Lett* 2001;507:114–8.
- [12] Davis JI, Nixon KC. Populations, genetic variation, and the delimitation of phylogenetic species. *Syst Biol* 1992;41:421–35.
- [13] Deichmann M, Benner A, Waldmann V, Bock M, Jackel A, Naher H. Interleukin-6 and its surrogate C-reactive protein are useful serum markers for monitoring metastasized malignant melanoma. *J Exp Clin Cancer Res* 2000;19:301–7.
- [14] Demczuk S, Thomas G, Aurias A. Isolation of a novel gene from the DiGeorge syndrome critical region with homology to *Drosophila* gdl and to human LAMC1 genes. *Hum Mol Genet* 1996;5:633–8.
- [15] Duchayne E, Demur C, Rubie H, Robert A, Dastugue N. Diagnosis of acute basophilic leukemia. *Leuk Lymphoma* 1999;32:269–78.
- [16] Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* 1998;95:14863–8.
- [17] Farris JS. The logical basis of phylogenetic analysis. *Adv Cladistics* 1983;2:7–36.
- [18] Fedyk ER, Ripper JM, Brown DM, Phipps RP. A molecular analysis of PGE receptor (EP) expression on normal and

- transformed B lymphocytes: coexpression of EP1, EP2, EP3 $\beta$  and EP4. *Mol Immunol* 1996;33:33–45.
- [19] Fellenberg K, Hauser NC, Brors B, Neutzner A, Hoheisel JD, Vingron M. Correspondence analysis applied to microarray data. *Proc Natl Acad Sci USA* 2001;98:10781–6.
- [20] Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol* 1981;17:368–76.
- [21] Getz G, Levine E, Domany E. Coupled two-way clustering analysis of gene microarray data [In Process Citation] *Proc Natl Acad Sci USA* 2000;97:12079–84.
- [22] Goldstein PZ, DeSalle R, Amato G, Vogler AP. Conservation genetics at the species boundary. *Conserv Biol* 2000;14:120–31.
- [23] Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 1999;286:531–7.
- [24] Guglielmi C, Martelli MP, Diverio D, et al. Immunophenotype of adult and childhood acute promyelocytic leukaemia: correlation with morphology, type of PML gene breakpoint and clinical outcome. A cooperative Italian study on 196 cases. *Br J Haematol* 1998;102:1035–41.
- [25] Johnson JP. Cell adhesion molecules in the development and progression of malignant melanoma. *Cancer Metastasis Rev* 1999;18:345–57.
- [26] Khan J, Wei JS, Ringner M, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat Med* 2001;7:673–9.
- [27] Kyoda K, Nakamura S, Matano S, Ohtake S, Matsuda T. Prognostic significance of immunoglobulin heavy chain gene rearrangement in patients with acute myelogenous leukemia. *Leukemia* 1997;11:803–6.
- [28] Lewis PO. Phylogenetic systematics turns over a new leaf. *Trends Ecol Evol* 2001;16:30–7.
- [29] Li K, Adibzadeh M, Halder T, et al. Tumour-specific MHC-class-II-restricted responses after in vitro sensitization to synthetic peptides corresponding to gp100 and Annexin II eluted from melanoma cells. *Cancer Immunol Immunother* 1998;47:32–8.
- [30] Manduchi E, Grant GR, McKenzie SE, Overton GC, Surrey S, Stoekert Jr. CJ. Generation of patterns from gene expression data by assigning confidence to differentially expressed genes. *Bioinformatics* 2000;16:685–98.
- [31] Mina RB, Tateyama S, Miyoshi N, Uchida K, Yamaguchi R, Ohtsuka H. Amplification of a c-yes-1-related oncogene in canine lymphoid leukemia. *J Vet Med Sci* 1994;56:773–4.
- [32] Mooy CM, Luyten GP, de Jong PT, et al. Neural cell adhesion molecule distribution in primary and metastatic uveal melanoma. *Hum Pathol* 1995;26:1185–90.
- [33] Mouawad R, Antoine EC, Khayat D, Soubrane C. Effect of endogenous interleukin-6 on Fas (APO-1/CD95) receptor expression in advanced melanoma patients. *Cytokines Cell Mol Ther* 2000;6:135–40.
- [34] Ohno H, Fukuhara S, Takahashi R, et al. c-yes and bcl-2 genes located on 18q21.3 in a follicular lymphoma cell line carrying a t(14;18) chromosomal translocation. *Int J Cancer* 1987;39:785–8.
- [35] Phipps RP, Lee D, Schad V, Warner GL. E-series prostaglandins are potent growth inhibitors for some B lymphomas. *Eur J Immunol* 1989;19:995–1001.
- [36] Planet PJ, DeSalle R, Siddall M, Bael T, Sarkar IN, Stanley SE. Systematic analysis of DNA microarray data: ordering and interpreting patterns of gene expression. *Genome Res* 2001;11:1149–55.
- [37] Raychaudhuri S, Stuart JM, Altman RB. Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pac Symp Biocomput* 2000; 455–66.
- [38] Rosenwald A, Wright G, Chan WC, et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *N Engl J Med* 2002;346:1937–47.
- [39] Roy HK, Bissonnette M, Frawley Jr. BP, et al. Selective preservation of protein kinase C- $\zeta$  in the chemoprevention of azoxymethane-induced colonic tumors by piroxicam. *FEBS Lett* 1995;366:143–5.
- [40] Sanders DS, Perry I, Hardy R, Jankowski J. Aberrant P-cadherin expression is a feature of clonal expansion in the gastrointestinal tract associated with repair and neoplasia. *J Pathol* 2000;190:526–30.
- [41] Sherlock G. Analysis of large-scale gene expression data. *Curr Opin Immunol* 2000;12:201–5.
- [42] SPSS. SPSS for Windows, version 9.0. Chicago, IL: SPSS Inc.; 1998.
- [43] Takahata N, Slatkin M. Private alleles in a partially isolated population. II. Distribution of persistence time and probability of emigration. *Theor Popul Biol* 1986;30:180–93.
- [44] Tamayo P, Slonim D, Mesirov J, et al. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci USA* 1999;96:2907–12.
- [45] Tuominen H, Pollanen R, Kallioinen M. Multicellular origin of tenascin in skin tumors—an in situ hybridization study. *J Cutan Pathol* 1997;24:590–6.
- [46] Vogler A, DeSalle R. Diagnosing units of conservation management. *Conserv Biol* 1993:354–63.
- [47] Voutsadakis IA. Interferon- $\alpha$  and the pathogenesis of myeloproliferative disorders. *Med Oncol* 2000;17:249–57.
- [48] Vyas P, Norris FA, Joseph R, Majerus PW, Orkin SH. Inositol polyphosphate 4-phosphatase type I regulates cell growth downstream of transcription factor GATA-1. *Proc Natl Acad Sci USA* 2000;97:13696–701.
- [49] Xiang YY, Wang DY, Tanaka M, et al. Expression of high-mobility group-1 mRNA in human gastrointestinal adenocarcinoma and corresponding non-cancerous mucosa. *Int J Cancer* 1997;74:1–6.
- [50] Xiong M, Fang X, Zhao J. Biomarker identification by feature wrappers. *Genome Res* 2001;11:1878–87.
- [51] Zeuthen J, Dzhandzhugazyan K, Hansen MR, Kirkin AF. The immunogenic properties of human melanomas and melanoma-associated antigens recognized by cytotoxic T lymphocytes. *Bratisl Lek Listy* 1998;99:426–34.