Molecular Classification of Cancer: Unsupervised Self-Organizing Map Analysis of Gene Expression Microarray Data[1]

David G. Covell,[2] Anders Wallqvist, Alfred A. Rabow, and Narmada Thanki

National Cancer Institute-Frederick, Science Applications International Corporation-Frederick, Developmental Therapeutics Program, Screening Technologies Branch, Laboratory of Computational Technologies, Frederick, Maryland 21702

Abstract

An unsupervised self-organizing map-based clustering strategy has been developed to classify tissue samples from an oligonucleotide microarray patient database. Our method is based on the likelihood that a test data vector may have a gene expression fingerprint that is shared by more than one tumor class and as such can identify datasets that cannot be unequivocally assigned to a single tumor class. Our self-organizing map analysis completely separated the tumor from the normal expression datasets. Within the 14 different tumor types, classification accuracies on the order of ~80% correct were achieved. Nearly perfect classifications were found for leukemia, central nervous system, melanoma, uterine, and lymphoma tumor types, with very poor classifications found for colorectal, ovarian, breast, and lung tumors. Classification results were further analyzed to identify sets of differentially expressed genes between tumor and normal gene expressions and among each tumor class. Within the total pool of 1139 genes most differentially expressed in this dataset, subsets were found that could be vetted according to previously published literature sources to be specific tumor markers. Attempts to classify gene expression datasets from other sources found a wide range of classification accuracies. Discussions about the utility of this method and the quality of data needed for accurate tumor classifications are provided.

Received 9/30/02; revised 12/20/02; accepted 1/15/03.
The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.
The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.
[1] This project has been funded in whole or in part with federal funds from the National Cancer Institute, NIH, under Contract No. NO1-CO-12400.
[2] To whom requests for reprints should be addressed, at National Cancer Institute-Frederick, Science Applications International Corporation-Frederick, Developmental Therapeutics Program, Screening Technologies Branch, Laboratory of Computational Technologies, Building 1052, Room 238, Frederick, MD 21702.

Introduction

The challenge of human cancer classifications will require methodologies capable of accurately assessing over 100 tumor types and an even larger number of tumor subtypes (1). Recent efforts to develop algorithmic methods for multiclass tumor classifications based on fingerprints of tumor gene expression profiles offer considerable hope toward precise, objective, and systematic cancer diagnosis (2–8) More importantly, these successes extend beyond disease classifications by providing a basis for molecular targeted therapies (9, 10). These early stages of cancer classification methodologies have raised many questions, which fall primarily into two broad areas: one involving analytical issues related to the computational and statistical strategies; and the other related to the construction of comprehensive patient-based gene expression databases (11, 12).

This paper will focus on the analytical issues of multiclass tumor classifications based on the publicly available oligonucleotide microarray samples available at the Whitehead site.[3] Our analysis will develop an unsupervised SOM[4]-based classification scheme using 190 patient tumor samples spanning 14 common tumor types and 90 normal tissue expression profiles. Our procedure is based on a fuzzy classification scheme that attempts to assign rankings of each tissue sample to all 14 tumor classes. Fuzzy is used here to indicate cases where an exact set of rules (genes) may not always yield clear and unique separations (tumor classes) among test samples (tissues). Our results will be contrasted with previously published, highly accurate, supervised classification schemes (2). These earlier supervised schemes have used binary classification strategies trained *a priori* according to tumor class membership as well as partitioning algorithms organized by class hierarchy (11).

Our intent here is to offer an alternative methodology directed at the common goal of multiclass tumor assignments. The approach is general and can easily be implemented on any gene expression dataset across a large number of tumor classes. Noteworthy in our approach is the development of fuzzy classification tumor assignments, which can be demonstrated to identify unclassifiable tumor samples as well as characterize other samples possessing gene profiles characteristic of more than one tumor class. As our results will demonstrate, our classification accuracies match those obtained by others, suggesting that little can be gained from this alternative approach. Implicit in our procedure, however, is the absence of gene selection using *a priori* classification

[3] www-genome.wi.mit.edu/MPR.
[4] The abbreviations used are: SOM, self-organizing map; CNS, central nervous system; S2N, signal to noise; PCA, principal component analysis.

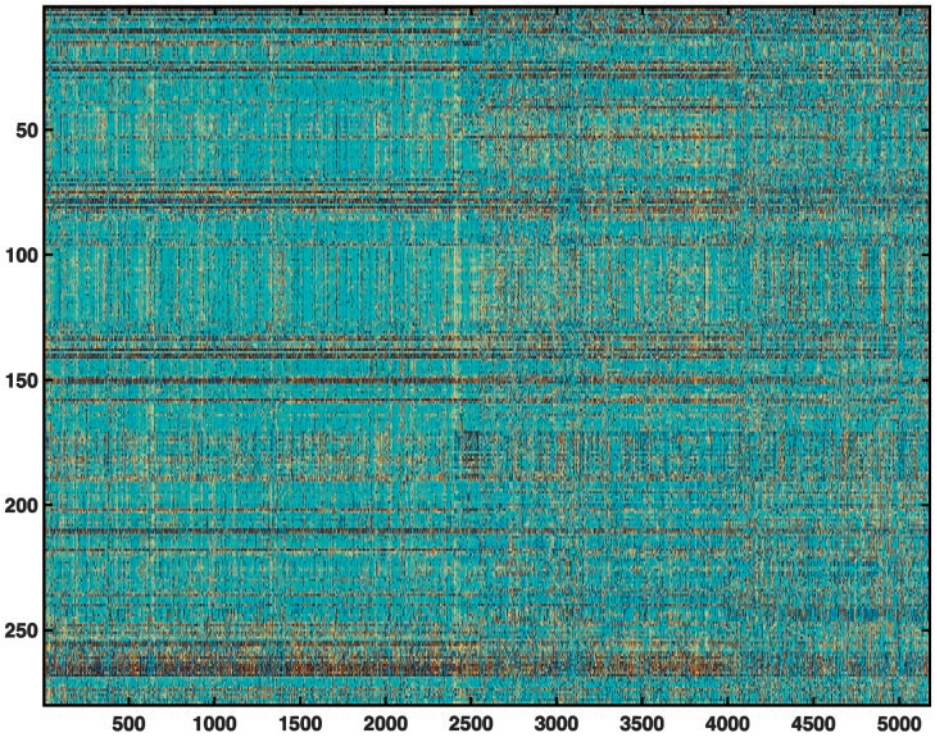

Fig. 1. Filtered, normalized dataset used in this analysis (2). The 16,063 genes on the complete dataset are filtered according to measures of variance to yield 5,183 genes. Analyzed data represent the expression values in Affymetrix's scaled average difference units, where the average difference values are calculated using Affymetrix's GeneChip software. Gene expressions are colored spectrally from *red* (highest) to *blue* (lowest) expressions.

labels. Successful classification schemes that rely on class labels for *a priori* training or supervision cannot resolve whether their success results from information contained directly within measures of gene expressions or from the procedures used to select features according to predetermined class labels. The unsupervised method proposed here finds that gene expression patterns alone can be used to obtain reasonably high tumor class predictions.

Methods

Data Filtering. The publicly available expression dataset at the Whitehead site[3] was used for our analysis. These data consist of gene measurements for 280 tissue samples based on the 16,000 Affymetrix microarray chipset. Ninety of these measurements are from normal tissues, whereas the remaining 190 datasets consist of microarray expressions from tumor tissues histologically assigned to 14 tumor classes. The data were initially filtered to exclude genes that had minimal variation for each tissue dataset. Genes with expression patterns less than 0.5 absolute deviation unit from their mean were excluded from further analysis, to yield 5,183 genes. Each of the 280 records was then normalized by subtracting the mean gene expression for this set of 5,183 genes and dividing by its variance. This normalization step was capped at 6 absolute deviation units. Fig. 1 displays the filtered dataset used in the subsequent analysis, ordered from top to bottom beginning with the tumor and ending with the normal tissue expressions. Distinctive patterns that either segregate the normal from tumor tissues or further classify tumor subsets are not obvious.

Previous reports to analyze parts of the same data, for the purpose of tumor classifications, were based on supervised clustering methods (2). Their analysis compared a variety of classification algorithms, with the finding that moderately accurate predictions could be achieved using a subset of the complete set of genes, selected using properties such as S2N (2, 3), radius-margin scaling (13), and gene shaving (14). Each of these methods based class predictions on explicit selection of the most highly informative genes, with each method having an optimal window for classification accuracy in terms of gene number. The highest classification accuracies were achieved by Ramaswamy *et al.* (2) when all 16,000 genes in their measured set were used in a one-*versus*-all classification scheme. Comparably high prediction accuracies could also be achieved in this same study when using a subset of the complete expression information for "training" their predictor to maximize accuracy. Although the method of Ramaswamy *et al.* (2) does not involve explicit gene selection before classification, selection of a gene subset is imposed indirectly by favorably weighting those genes that contribute the greatest to separation among the classification hyperplanes (2). Thus, in principle, their method is supervised based on using feature selection as an explicit component of their classification procedure.

Because virtually all genes measured from these tissues convey information about tumor class, it is reasonable to ask which set(s) of genes may prove most informative in an unsupervised scheme. Unsupervised approaches can be contrasted with supervised or trained approaches where a known answer is sought, and class predictions become a matter of identifying which portion of the data will maximize prediction accuracy. As noted above, the most accurate predictions of Ramaswamy *et al.* (2) occur when using all of the gene expression data. Supervised analysis, using only a subset of the most informative

genes, could be used to obtain highly accurate discrimination between their 14 tumor classes. The analysis of Ramaswamy *et al.* (2) noted that unsupervised approaches, especially those based on SOM analysis, lacked sufficient discrimination between tumor classes.

Using as our starting point the Whitehead's result that maximal predictions occur when using the most data, we proposed an unsupervised approach that begins with their complete gene expression dataset. These data were first filtered to exclude those genes with a low S2N ratio, with the implicit assumption that these genes contain insufficient information for tumor distinctions when using an unsupervised scheme. In contrast to the procedure of Ramaswamy *et al.* (2), utilization of all genes did not improve our prediction accuracies. This result is consistent with the premise that a gene expression within the noise range (*i.e.,* close to the group mean) does not provide additional information for tumor classifications. Filtering data using a higher cutoff than 0.5 deviation units resulted in poorer classification accuracies. This observation would suggest that our approach requires at least a minimal number of sampled genes for classification. No additional attempts were made to select an alternative cutoff and thus identify a minimal set of genes for unsupervised classifications. Our results will show that gene expressions can be used in an unsupervised classification scheme to accurately classify tumor types. A direct result of this analysis is the identification of gene subsets that are most informative for tumor classifications as well as gene expressions that are widely shared among many tumor classes.

SOM Clustering. Filtered data were clustered using SOMs (15). Since the early 1960s, research in statistical methods has produced a wide range of tools for the analysis of multidimensional data. Commonly used approaches include techniques of hierarchical clustering (16), k-means clustering (17, 18), multidimensional scaling, binary deterministic annealing (19), and SOMs (15, 20). Whereas all of these methods are aimed at the identification of pattern similarities between diversity measures, literature references report various degrees of success when using these methods for the analysis of large biological datasets. Many of these approaches begin by assignment of pairwise measures of similarity between data records using Pearson correlations and Euclidean, Mahalanobis, or Minkowski measures of similarity (16). Such pairwise measures are known to have limited power, particularly when data are contaminated with large amounts of noise, resulting in a high likelihood of random statistical correlations (16, 21), or when the data are not hierarchical. Alternative methods that deal with noisy data include PCA and the related method of singular value decomposition, where the data are reexpressed along directions that maximize the S2N ratio (22). The SOM method has been used extensively in the analysis of microarray gene expression data (3, 15). A noteworthy feature of the SOM methodology is its capacity to cluster quite noisy and often incomplete datasets (21, 23). Our prior efforts to analyze large drug screening datasets using the SOM methodology further demonstrate its capacity to handle such crude data (24).

The SOM method can be divided into two regimes: clustering in high dimensional space; and projections into a lower dimensional display space. This first step clusters data in its original high dimension space ($N = 5183$) by locating reference vectors in this high dimensional data space. Each reference vector is an "average" of all data vectors within a given cluster. These reference vectors are obtained by minimizing the deviation between the data vectors (V) and reference vectors (R):

$$\nabla R \propto \sum_i h(\|V - R\|)\|V - R\| \quad (1)$$

where ∇R is the incremental change in position of the reference vector R, V is the set of data vectors, and $\|V - R\|$ is the distance between data and reference vector. The neighborhood kernel function $h(\|V - R\|)$ weights the change in the position of the reference vectors. This neighborhood kernel collectively orders the reference vectors to locations of maximum information in data space. The form of the neighborhood kernel function exhibits a maximum when the data and reference vectors coincide and goes to zero as these vectors become more distant. Often the neighborhood kernel is a Gaussian function; however, our analysis (24) finds that the Epanechikov function [$\max(0,1-\|V - R\|^2)$] consistently yields improved clustering, and was selected for this analysis.

The form of Eq. 1 determines the position of reference vectors that best mirror the data space or, alternatively, how the reference vectors partition the data space into clusters. Regions that are rich in data vectors attract many reference vectors and, as a result, finely divide these regions of high information content. This process can be contrasted with the more conventional PCAs, where data are oftentimes reoriented, in a linear fashion, on to the space of the topmost principal components. Multivariate data may also be nonuniformly distributed across all observations (*i.e.,* genes in this case); in which case, effective means are required to partition this data into densely populated subspaces. The SOM transformation stretches these data-rich regions, thereby enhancing relevant cluster distinctions. A direct consequence of the SOM reordering of the data space is the ability to display these results in an interpretable manner. The method of display is the uniform projection of the SOM clustering in high-dimensional space to a low-dimension display space. This mapping is simple and also retains a great deal of the original high-dimensional information (See Fig. 2*A*).

SOM Results

Map Features. Map dimensions in SOM analysis determine the number of clusters for each dataset. Oftentimes this number is selected *a priori*; however, the SOM method of Kohonen (15) uses an heuristic based on the ratio of the first two principal components of a PCA. Using this procedure and the 280×5183 data vectors in the filtered dataset, map dimensions of 29 rows by 17 columns were obtained as the minimal map dimensions. We note above that the SOM methodology seeks to uniformly populate clusters with nearly equal numbers of data vectors (*i.e.,* observations). To achieve this, regions of data space with the highest information content are stretched to enhance discrimination among these data points, whereas the opposite occurs for regions of low information content. As a consequence, attempts to map data vectors to too few clusters result in heterogeneous

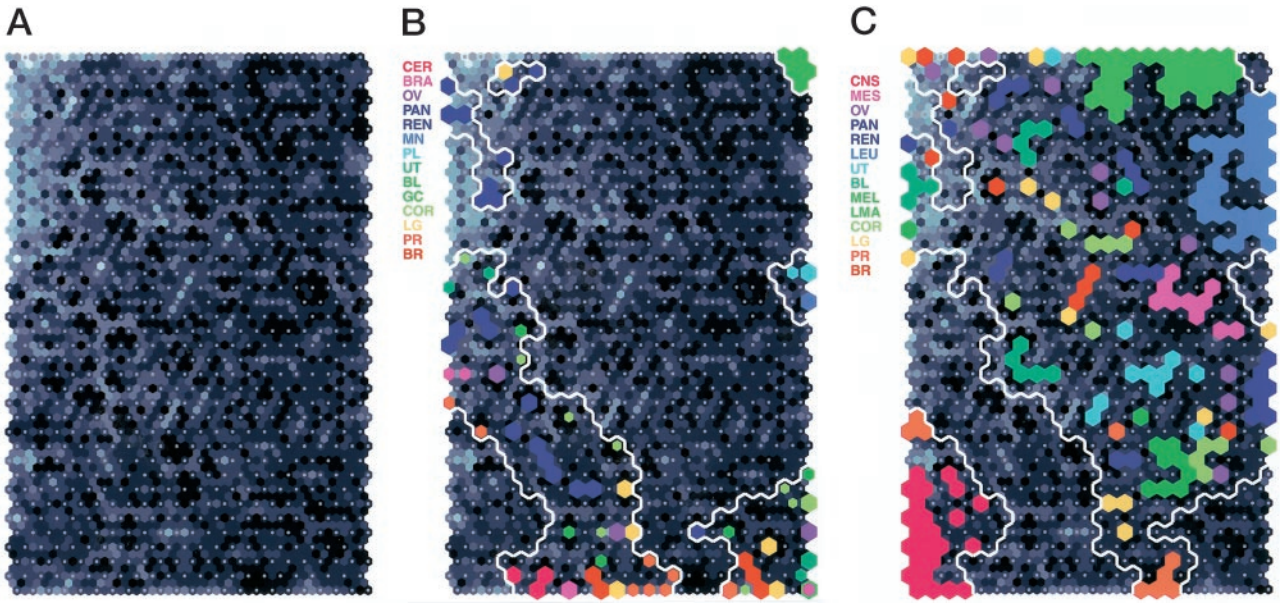

Fig. 2. *A,* SOM of filtered dataset. Map dimensions are 38 rows by 23 columns. Map colors indicate Euclidian distance between reference vectors at each map node: *black* = close; and *white* = far. *B,* SOM annotated according to the projected node locations for normal tissue gene expressions. Fourteen different tissue types were analyzed across 90 samples: *BR,* breast; *PR,* prostate; *LG,* lung; *COR,* colorectal; *GC,* germinal center; *BL,* bladder; *UT,* uterine; *PL,* peripheral lymphocytes; *MN,* normal monocytes; *REN,* renal; *PAN,* pancreas; *OV,* ovarian; *BRA,* brain; and *CER,* cerebellum. *C,* map projections for the 190 tumor samples. Fourteen different tumor types were analyzed: *BR,* breast adenocarcinoma; *PR,* prostate adenocarcinoma; *LG,* lung adenocarcinoma; *CO,* colorectal adenocarcinoma; *LMA,* lymphoma; *MEL,* melanoma; *BL,* bladder cell transitional carcinoma; *UT,* uterine adenocarcinoma; *LEU,* leukemia; *REN,* renal cell adenocarcinoma; *PAN,* pancreatic adenocarcinoma; *OV,* ovarian adenocarcinoma; *MES,* pleural mesothelioma; and *CNS,* CNS. The *thick white line* represents a boundary between the normal and tumor tissue types. The location of this boundary is not precise, and its appearance is intended as a reference landmark across all SOMs.

cluster memberships. Attempts by Ramaswamy *et al.* (2) to cluster this same data in a 5 × 5 SOM and devise accurate predictions based on this map were failures. In contrast, assignments of data vectors to maps of slightly higher than the SOM recommended dimensions enhance separation between clusters, essentially by generating a cluster without any data vectors, with a reference vector as an interpolation of its nearest map nodes. Here we have used a SOM with 38 rows and 23 columns. When compared with clustering on the recommended 29 × 17 SOM, the map expansion used here has little influence on cluster membership for the 280 tissue samples, but as we will show later in our validation steps, it has the advantage of greater flexibility for class assignments using data derived from alternative sources. An obvious criticism of this larger map is that 280 data vectors can be placed on 874 possible clusters, raising the possibility that each data vector will be placed in its own cluster. Because the SOM procedure spatially organizes data vectors on a two-dimensional map, the most similar data vectors are placed in nearby map locations. Thus a "neighborhood" analysis can be used to assign map nodes to tumor class, without the risk of assigning heterogeneous data vectors to the same group. The SOM clustering of the 280 expression datasets was refined from a convergence radius of 40 map units to a final radius of 0.001. Using C++ code, a total of 8 processing hours was required on a COMPAQ ALPHA processor to complete one SOM.

SOM Results. The SOM for this dataset appears in Fig. 2*A* as a collection of hexagonal nodes projected onto a rectangular map. Map coloring defines the Euclidian distance between reference vectors, where near and far are shown in *black* and *white,* respectively. In general, the SOM has organized these data into well-defined clusters, shown as dark and light "islands" on this map, to designate clusters of nodes with high and low similarity between reference vectors, respectively. Visualizing SOMs using a two-dimensional projection is a departure from the more conventional dendrograms provided by hierarchical clustering methods. A reasonable paradigm is to consider SOMs as dendrograms, where the clades disappear into the plane of the page, and the colors between each clade correspond to the length of branches on the dendrogram. Unlike dendrograms, which by their nature are two-dimensional projections, the SOM method is nonhierarchical in nature and provides a level of organization which, according to the hexagonal projection used here, allows visualization of up to six neighbors, as opposed to the pairwise associations available for hierarchical dendrograms.

The ability of SOMs to organize data can be evaluated in a number of ways. Correlation statistics can be used to relate the expression datasets to the reference vectors on the map. The mean and SD for the complete distribution of data to reference vector correlations is 0.96 ± 0.04. This result reflects a very high degree of match between the data and reference vectors. This high correlation also reflects map size, which, in the extreme case of a very large map, would yield perfect correlation statistics between data and reference vectors. Even with oversized maps, the nonhierarchical

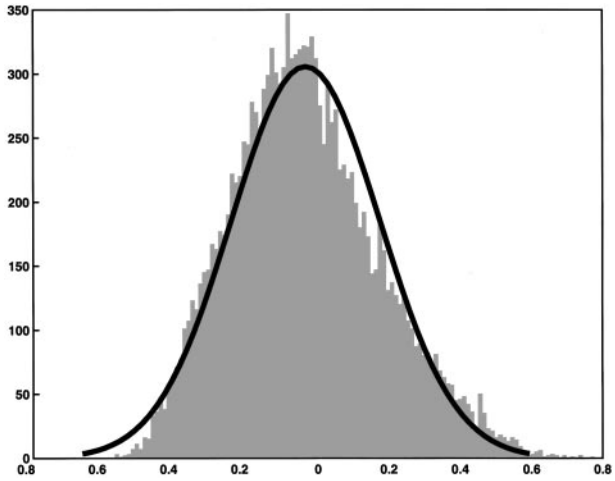

Fig. 3. Histogram of correlations between all 874 SOM reference vectors ($\bar{x} = 0.04$; $s = 0.21$). The *solid line* represents Gaussian fit to the histogram.

organization provided by the SOM analysis can be used for inspecting local neighborhoods and evaluating merits of goodness of fit within and between members of these neighborhoods. Motivation for exploiting neighborhood information is based on the diversity among the 874 reference vectors of the SOM. The mean and SD for all pairwise correlations between reference vectors for all map nodes is 0.098 ± 0.21 (see Fig. 3), with a range of correlations between -0.67 and 0.74, thus spanning over 3 SD units. This diversity between reference vectors is an indication of the differences in expression profiles for this dataset, differences that will be used later to classify tissue types.

Visual examinations of cluster members can be used to provide confidence that SOM clusters are comprised of similar expression data vectors. Fig. 4 plots two different cases of clustered expression data, one consisting of leukemia expression data and the other consisting of prostate tumor data. Within each cluster, similarities in gene expression are indicated by the vertical "banding" evident across the gene set, most notably in the cases of low (*blue*) and high (*red*) gene expression. In this example, the low and high gene expressions in the prostate tumor tissue, at gene positions 1800–2000 and 4000–4100, respectively, are not shared by the leukemia tumor set. This is a typical example of the diversity observed between different clusters, as well as the coherence of expression patterns within a given cluster.

Normal Tissues. The SOM analysis has completely separated the normal tissue expressions (Fig. 2*B*) from those of the tumor set (Fig. 2*C*). For the normal and tumor tissues, the node locations of individual tumor classes are shown as *hexagons,* colored the same for each tumor class (*e.g.,* normal and tumor breast tissue appear as *red hexagons* in Fig. 2, *B* and *C,* normal and tumor ovarian tissues appear as *purple hexagons,* and so forth). None of the clusters are jointly occupied by normal and tumor data. SOM groupings for these normal tissues comprise two large contiguous regions running diagonally along the middle, left edge toward the bottom, middle, at the lower right corner, and four isolated regions (one pair at the top left corner and the other pair at the right edge). Evidence for grouping of similar tissue types into single regions on the map are found for normal germinal center (*GC*), pancreatic (*PAN*), cerebellum (*CER*), brain (*BRA*), monocyte (*MN*), and prostate (*PR*) tissues. The remaining tissues also display evidence for localized groupings; however, these are located in two or more regions of the map. The fact that the normal tissue expressions are located primarily at the map perimeter indicates that SOM clustering has placed these datasets at the most distant locations from the tumor expressions. This point will be used later to define tumor *versus* normal markers for tissue classes. The *boundary lines* shown in *white* represent the major branches of the hierarchically clustered reference vectors. The complete set of boundaries for 60 branches of this cluster tree will be used later to define local neighborhoods for tissue classifications. In summary, the normal expression datasets are readily distinguished from the tumor expressions based on SOM analysis.

It is interesting to compare these results with those based only on hierarchical clustering. Consistent with previously published reports on this and other datasets, our analysis, using average linkage Ward's clustering, finds that the normal tissue data are integrated within dendrogram branches containing tumor tissues and thus are not readily separable. Our observation that SOM clustering readily separates the normal and tumor groups has also been reported by Ramaswamy *et al.* (2) and lends support to our earlier premise that the nonhierarchical nature of this data precludes the effective use of hierarchical clustering methods.

Tumor Tissues. The tumor regions of the SOM comprise ~80% of the total map space (see Fig. 2*C*). Two populations of cluster groupings are found for the tumor data: tissue expressions that are clearly segregated into contiguous neighborhoods on the SOM; and others that are localized to two or more map regions. The most contiguous tumor groupings are found for lymphoma (*LMA*), leukemia (*LEU*), CNS, melanoma (*MEL*), uterine adenocarcinoma (*UT*), and mesothelioma (*MES*). The remaining tumor datasets are grouped into two or more regions on the map. An example of multiple projection sites is illustrated for bladder tumor expressions (*blue-green*), where nearly equal populations of tumor tissues are grouped into three locations, two at the upper left portion of the map and one near the map center. Tissues from renal, breast, colorectal, lung, and ovarian tumor tissues are widely scattered over the tumor region of the map. Noteworthy in these projections is a lack of substantial tissue heterogeneity within different local neighborhoods. The existence of a tumor class projected to multiple map locations is an indication that additional tumor classes, or subclasses, may exist within this data. Speculations of this type cannot be explored further with the small dataset used here.

A number of cluster groupings are apparent where normal and tumor tissues of the same type are located as nearby cluster neighbors. Examples include CNS adjacent to normal brain (*BRA*) and cerebellum (*CER*), leukemia (*LEU*) adjacent to normal monocytes (*NM*) and peripheral lymphocytes (*PL*), lymphoma (*LMA*) adjacent to normal germinal tissue (*GC*), and adjacent locations for normal and tumor prostate tissues (*PR*). These results indicate that although tissues of the same organ are found as adjacent SOM neighbors and, as such,

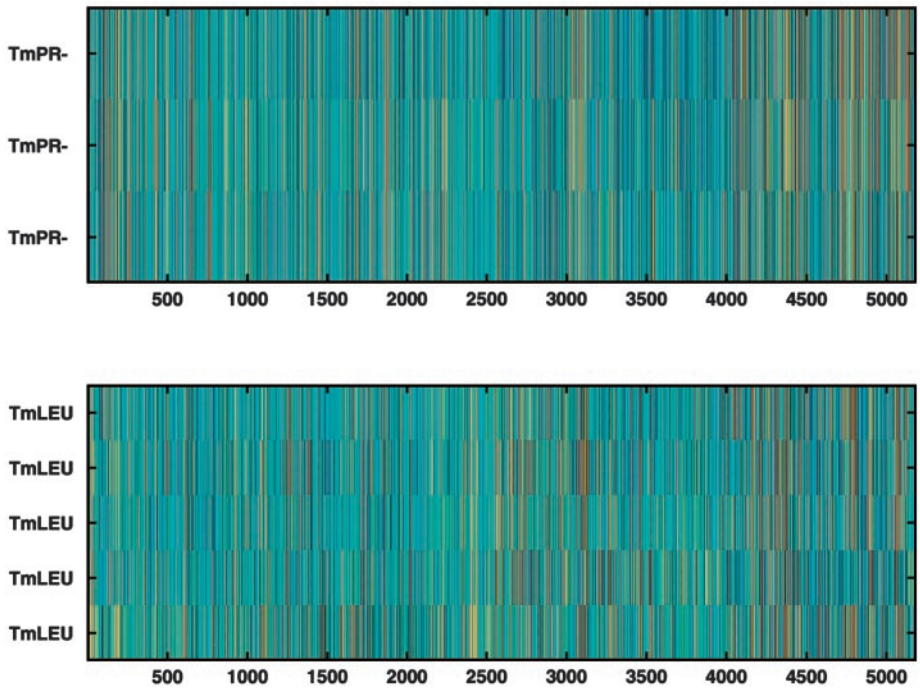

Fig. 4. Raw expression data for separate map nodes. Examples shown here are for tumor prostate (*top*) and leukemia (*bottom*) datasets. Gene expressions are colored spectrally from *red* (highest expressions) to *blue* (lowest expressions).

reflect nearly similar gene expressions, they remain as separate and distinct tumor and normal map clusters.

Classifications

Fuzzy Probabilities. The potential for tumor classifications based on measures of gene expression holds high promise for improvements in therapeutic strategies. However, genetic boundaries between tumor classes may not be sufficiently sharp for the highly accurate class predictions necessary for selecting therapies. The basis for these limitations continues to be actively pursued at the levels of basic and applied scientific research, with the hope that technological improvements in gene expression measurements and a greater understanding of the roles of unusual gene expression in cancer etiology will substantially reduce these limitations.

Assignments of gene expression data vectors into tumor or normal classifications are made by analyzing the SOM results. The basic idea uses a fuzzy classification scheme that assigns a probability of each expression dataset to all of the 14 tumor and 14 normal tissue classes (25). This method is similar in concept to that used by Ramaswamy *et al.* (2) but is based on the SOM results rather than binary correlation statistics. At each map node (*i,j*) ($1 \leq i \leq 38$, $1 \leq j \leq 23$) occupied by at least one data vector, the Euclidian distance between the data vector ($V_{i,j}$)[5] and reference vector ($R_{i,j}$) is determined.

$$d_{i,j} = \text{sqrt}(V_{i,j} - R_{i,j})^2 \qquad (2)$$

[5] *V* and *R* designate data and reference vectors, respectively (see Eq. 1). The subscript *i,j* has been added to the label to denote the coordinates of the SOM node or cluster in terms of a row (*i*) and column (*j*) designation.

By default, the smallest distance determines the map location for projection of each data vector on the 38×23 SOM. Next, a random sampling of data vectors selected from the pool of 280 data samples is constructed to determine the distribution of distances between a random population of data and the reference vector, $R_{i,j}$. Each sample set consists of 5000 randomly selected data vectors and is repeated for all map nodes ($N = 38 \times 23 = 874$) such that at each node, a distribution of Euclidian distances is generated. At the completion of this step, 874 distributions of distances between data vectors and reference vectors are established. A Z_{score} is determined at each node to establish the statistical significance (at 2 SDs from the mean) between each data vector and its projection to all map nodes.

$$Z_{scorei,j} = (d_{i,j} - \bar{x}_{i,j})/s_{i,j} \qquad (3)$$

where $\bar{x}_{i,j}$ and $s_{i,j}$ are the sample mean and variance, respectively, for the randomly sampled distribution of distances. The intention here is to assess the match between a data vector and its best fit reference vector. A small numerical value may not be small in the context of all possible matches. Thus Eq. 3 assesses the "strength" of a match according to its difference from the distribution of all such matches. This step yields 874 matrices of dimensions 38×23, where each matrix contains the Z_{score}s for each set of data vectors projected to all map locations. The magnitude of Z_{score} is a measure of goodness of fit between the data and reference vectors of the SOM and is similar in concept to the S2N calculations of Golub *et al.* (3). However, by defining a Z_{score} for all map positions (*i.e.,* clusters), information is retained about locations of the best to worst projection scores. Low and high Z_{score}s indicate low and high preference, respec-

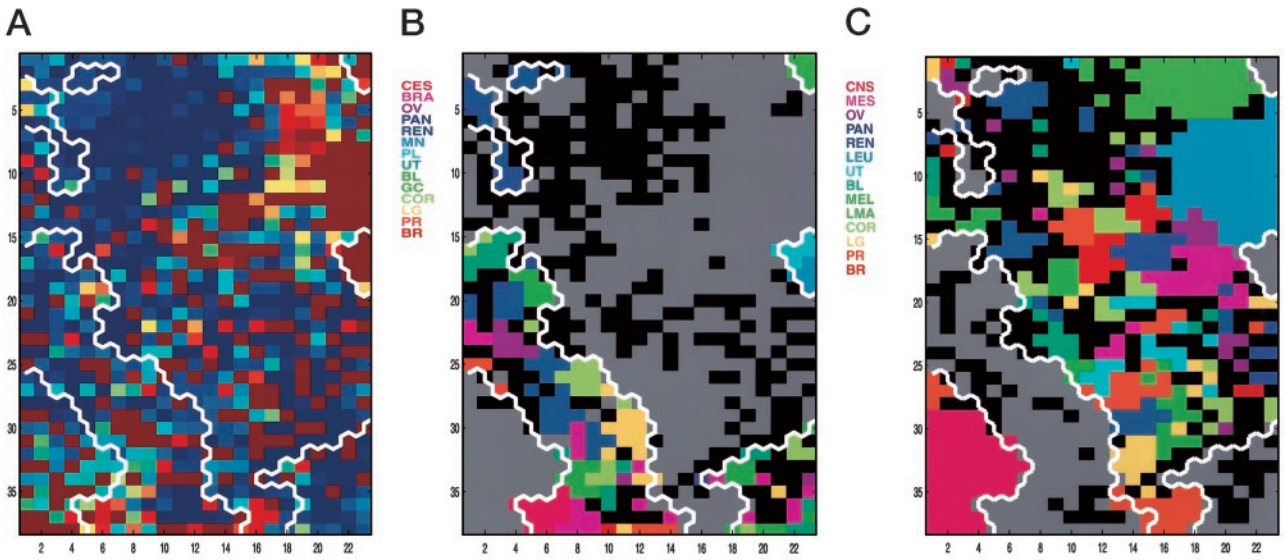

Fig. 5. SOM annotated according to tissue prediction probabilities. *A* displays the prediction probabilities at each map node, $PR_{i,j}$, colored spectrally from *red* (highest probabilities) to *blue* (lowest probabilities). Map boundaries, shown as a *solid white line*, indicate separation between normal and tumor samples. *B* and *C* identify class predictions of normal and tumor tissues for map nodes. Map nodes are colored according to tissue type and indicated along the figure edge.

tively, between a data vector and its best position on the SOM.

The next step is to assign a probability for all tumor classes to each node on the SOM. The procedure described above yields a set of Z_{score} preferences for data vectors at all map positions, regardless of the tumor type for each data vector. This set of Z_{score}s is an indication of the first, second, and third positions and so forth on the SOM where an individual data vector fits best. These Z_{score}s are then labeled according to the tumor and normal class assignments for each of the 280 data vectors. Tissue-specific Z_{score}s, designated as $PR(class)_{i,j}$, are constructed by locating the map projections for data vectors from each expression dataset and constructing the average Z_{score} for each tissue class at all map nodes.

$$PR(class)_{i,j} = \sum_{k=1}^{nclass(class)} Z_{scorei,j}(class, k)/nclass(class) \quad (4)$$

where *nclass(class)* is the number of data vectors in each of the 28 possible classes projected to map position (i,j), and $Z_{scorei,j}(class, k)$ refers to the *k*th class projection data vector at node (i,j). *PR(class)* assigns a probability score for node (i,j) to each tumor class. Because data vectors within a tissue class will rarely be identical and may often share expression profiles for a different tissue class, the probabilities, as calculated here, reflect the "fuzzy" nature of this data. It is important to stress that these probabilities reflect only the information contained in the expression data and, as such, correspond to the fact that boundaries between tumor classes are not precisely defined. The analysis presented below will document the prediction accuracies for the set of 280 data vectors as well as make comparisons with predictions from microarray datasets derived from earlier Affymetrix chips.

Fig. 5*A* displays the class prediction probabilities ($PR_{i,j}$) for all regions of the SOM. These results represent measures of goodness of fit between the map's reference vectors and data vectors projected across the complete SOM. The *dark red regions* indicate map nodes with the highest confidence that the data vectors projected to these nodes are well placed in the context of all possible map positions. Notable in the regions of highest prediction accuracies are leukemia, CNS, normal germinal tissue, plasma monocytes, and peripheral lymphocytes. *Dark blue regions* indicate reference vectors with low *PR(class)* values, less than 0.3. These nodes are often associated with neighborhoods of data projections for different tumor classes, and, as such, no clear class distinction can be made for data vectors that project to these regions. To retain only those prediction classifications with the highest confidence, any probability below 0.3 is regarded as unclassifiable. Consequently, data vectors projected to these locations cannot reliably be assigned to a tissue class. These regions on subsequent maps are colored *black*.

Tests of Class Predictions. The probabilities, *PR* (*class*), provide a baseline for distinguishing between tissue classes. To establish a "best case scenario" for these predictions, a test was conducted to determine how well the class probabilities correctly assign each of the 280 data vectors used to generate the SOM. Supervised clustering is often based on developing rules from a training set that maximize discrimination among the tissue classes, hence the term supervised. The best training rules yield very high prediction accuracies for the training set, whereas lower accuracies are usually found for the test set. Iterations between improving training rules and evaluating test results can lead to refinements of

Table 1 Tissue predictions based on original data

Tissue	Preference	Hits	Total
Tm-BR[a]	0.72	8	11
Tm-PR	1.00	10	10
Tm-LG	0.91	10	11
Tm-COR	1.00	11	11
Tm-LMA	1.00	22	22
Tm-MEL	1.00	10	10
Tm-BL	0.91	10	11
Tm-UT	0.90	9	10
Tm-LEU	0.96	29	30
Tm-REN	1.00	11	11
Tm-PAN	0.91	10	11
Tm-OV	0.82	9	11
Tm-MES	1.00	11	11
Tm-CNS	1.00	20	20
Nm-BR[b]	0.40	2	5
Nm-PR	0.56	5	9
Nm-LG	0.71	5	7
Nm-COR	1.00	11	11
Nm-GC	1.00	6	6
Nm-BL	0.57	4	7
Nm-UT	0.83	5	6
Nm-PL	1.00	3	3
Nm-NM	1.00	2	2
Nm-REN	1.00	13	13
Nm-PAN	0.90	9	10
Nm-OV	1.00	3	3
Nm-BRA	0.80	4	5
Nm-CER	1.00	3	3

[a] Tumor (Tm): BR, breast adenocarcinoma; PR, prostate adenocarcinoma; LG, lung adenocarcinoma; COR, colorectal adenocarcinoma; LMA, lymphoma; MEL, melanoma; BL, bladder cell transitional carcinoma; UT, uterine adenocarcinoma; LEU, leukemia; REN, renal cell adenocarcinoma; PAN, pancreatic adenocarcinoma; OV, ovarian adenocarcinoma; MES, pleural mesothelioma.
[b] Normal (Nm): BR, breast; PR, prostate; LG, lung; COR, colorectal; GC, germinal center; BL, bladder; UT, uterine; PL, peripheral lymphocytes; NM, normal monocytes; REN, renal; PAN, pancreas; OV, ovarian; BRA, brain; CER, cerebellum.

the supervision rules and better prediction accuracies. In the method proposed here, the complete dataset is used for developing prediction statistics, via inclusion of randomly generated datasets, and then used for evaluating these results for prediction accuracies. As noted above, each SOM node has a tissue class probability assigned to it. For example, node (20,19) may be assigned a 0.47, 0.38, and 0.15 probability of being in the breast, prostate, and lymphoma tumor class, respectively, and zero probability for the remaining classes. A test data vector projected to map node (20,19) would thus have the greatest probability for assignment to the breast tumor class. Table 1 provides a summary of these assignments for the 280 vectors available in this study. Based on these results, 12 of the 14 tumor classes could be correctly assigned for over 90% of the tumor data vectors. The worst class assignments are for breast and ovarian tumors, where 8 of 11 and 9 of 11 cases are correctly assigned. The classifications within the normal tissues are slightly weaker when compared with the tumor datasets, with seven cases of perfect classification, and the rest being correctly assigned for 40–90% of the cases. This type of calculation is similar to a "leave-one-out" or jack-knife procedure because repeated generations of SOMs with one expression dataset removed does not substantially change the SOM when compared with using all 280 data vectors. This type of analysis, as noted above, only establishes minimal guidelines for quality of class predictions based on the same data used to define the SOM clusters. It is apparent from these results that poor tumor predictions can be anticipated for breast and ovarian tissues. In the following section, a general method for tumor class prediction is proposed and applied to additional publicly available tissue expression datasets.

Tumor Classifications

The previous section revealed that very good class assignments are possible when the "training" dataset of 280 data vectors is postprojected on the SOM clusters. Similar tests where small random perturbations were introduced in the training data vectors found that the quality of class predictions was substantially poorer than those listed in Table 1. The details of this result will not be presented here, other than to note that much of this poor prediction was related to groups of neighboring SOM clusters that contained heterogeneous mixtures of nearly equal class probabilities. In these cases, class predictions based on the highest class probability were correct for ~50% of the cases. In cases where different tumor classes appeared as neighboring clusters, it was found that a local sampling of the class probabilities improved the quality of predictions. Based on this observation, a heuristic for local sampling was developed based on weighting the class probabilities by the Euclidian distance of the local neighborhood of reference vectors:

$$PR'(class)_{i,j} = \sum_{k,l=i,j}^{neighborhood_{i,j}} PR(class)_{k,l}/(1 + D) \qquad (5)$$

where D is the Euclidian distance between the reference vectors at node (i,j) and neighboring reference vectors. The local neighborhood is established according to hierarchical clustering of the map's reference vectors and selecting a limit of 60 clusters. The procedure works as follows. The projection of a data vector to a map node, based on the lowest Euclidian distance, determines the map node (i,j). Based on this node location, the local region is sampled by calculating *PR* across the set of neighboring nodes and assigning the tumor class according to the highest probability, *PR′*, as in Eq. 5. The result of this neighborhood sampling does not change the class probability for data vectors projected to node (i,j) because D in this case is zero. By using this heuristic, nearby clusters with high class probabilities are included in the prediction assignment. Using this procedure, class assignments can be made for each node based on information about the local probability (*PR*) and the neighborhood probability (*PR′*) for all 14 tumor classes. The most reliable class assignments correspond to a cluster containing mostly tissues of one tumor class, with a surrounding neighborhood of nodes also containing mostly tissues of the same tumor class.

Table 2 lists the class prediction assignments, grouped according to each of the 14 tumor classes. Each column

Table 2 Tissue predictions based on original data[a]

	BR	PR	LG	COR	LMA	MEL	BL	UT	LEU	REN	PAN	OV	MES	CNS	
BR	88		12												BR
PR		89						11							PR
LG	12		75	13											LG
COR	40			40				20							CO
LMA					100										LY
MEL						100									ME
BL							86	14							BL
UT								100							UT
LEU									100						LE
REN						17		16		67					RE
PAN			20								60		20		PA
OV										30		30	30		OV
MES								25					75		MS
CNS														100	CS
	BR	PR	LG	COR	LMA	MEL	BL	UT	LEU	REN	PAN	OV	MES	CNS	

[a] BR, breast adenocarcinoma (n = 11, 3); PR, prostate adenocarcinoma (n = 10, 1); LG, lung adenocarcinoma (n = 11, 3); COR, colorectal adenocarcinoma (n = 11, 6); LMA, lymphoma (n = 22, 0); MEL, melanoma (n = 10, 2); BL, bladder cell transitional carcinoma (n = 11, 4); UT, uterine adenocarcinoma (n = 10, 2); LEU, leukemia (n = 30, 0); REN, renal cell adenocarcinoma (n = 11, 5); PAN, pancreatic adenocarcinoma (n = 11, 6); OV, ovarian adenocarcinoma (n = 11, 8); MES, pleural mesothelioma (n = 11, 3); CNS (n = 20, 0). Numbers in parentheses refer to the total number of tissues samples in each tumor class and the number of unclassifiable tissues in each class, respectively.

corresponds to the histologically assigned tumor class. The values in each cell represent the fraction of classifiable data vectors within each tissue type. An average number of 3 data vectors/tumor group was not sufficiently distinct to fall into one tumor class. In particular, only 6 of the 11 renal samples and 5 of the pancreatic samples were classifiable, whereas all of the lymphoma, leukemia, and CNS data could be classified. An average of 79% of the data vectors were correctly assigned,[6] with the lowest accuracies found for colorectal (40%) and ovarian (30%) tissues. Eight classes were correct for >85% of the cases, with perfect classification for five cases. Direct comparison of these results with those of Ramaswamy *et al.* (2) found a nearly equal overall prediction accuracy of ~80%, with the highest accuracies associated with the CNS, lymphoma, melanoma, and leukemia classes, and the lowest accuracy was found for the ovarian samples. The most dramatic difference in prediction accuracies was observed for the colorectal data, where the training and test class accuracies of Ramaswamy *et al.* (2) were 75% and 100%, respectively, whereas our accuracies were quite low, at 30%. This class was also among the least classifiable datasets, with six of the samples being unclassifiable. The greatest number of incorrect classifications was assigned to the uterine tumor class, where misassignments occurred for the breast, colorectal, bladder, renal, and mesothelioma classes.

Fig. 5, *B* and *C,* displays the SOM locations according to the normal and tumor tissue classifications, selected for the tumor type with the highest value of $PR'_{i,j}$ as derived using Eq. 5, with the values of $PR_{i,j}$ as shown in Fig. 5*A*. The same color scheme used to identify tissue classification types in Fig. 2 is applied here. Inspection of these figure reveals the close correspondence expected between the map locations for different tissue types and their classification probabilities. In some tissues, coherent gene expressions within a tumor class define a contiguous region on the map, whereas for other tissues, there is a clear indication for subclusters of gene expression, as indicated by different map locations. Notable in this latter case are multiple map locations for prostate and bladder tissues, each with adjacent regions for their normal and tumor tissue types.

Fig. 6, *A* and *B,* displays the map boundaries obtained from hierarchical clustering of the reference vectors of the SOM. Each clade defines the most similar reference vectors, based on Euclidian distance. On average, ~ 14 ± 9 map nodes appear in each clade branch, and the average correlation coefficient between all data vectors within a clade was 0.86 ± 0.09. This high correlation indicates a very high degree of similarity within clade members.

Our prediction values for $PR'_{i,j}$ are based on local sampling of nodes within the boundary containing the projection node (i,j). Evident from these figures is the fuzzy nature of this data, as revealed by boundaries containing multiple tissue types (*i.e.,* color). Most of the boundaries in the central portion of the map have the possibility of multiple types of tissue classifications. In contrast, the lymphoma, melanoma, and CNS classes at the map corners are homogeneous. Fig. 7 provides example indications of the difficulties associated with discrimination between some tissue classes. Fig. 7*B* displays the normalized data vectors for a clade that consists of only leukemia expressions, whereas Fig. 7*A* and Fig. 7*C* display different clades containing mixtures of tumor classes. All of these groups are characterized by within clade data correlations above 0.75. Moreover, gene expressions are visibly different between each group. Whereas the results for Fig. 7*B* provide a "signature" for leukemia, the "signature" in Fig. 7, *A* and *C,* is shared by most of the tumor and some of the normal tissue classes.[7] Whereas our classifi-

[6] Based on the average of the diagonal values in Table 2.

[7] Note that these groupings are based on hierarchical clustering of reference vectors and using a cutoff of 60 branches. At this level of classification, the tumor and normal samples are comingled.

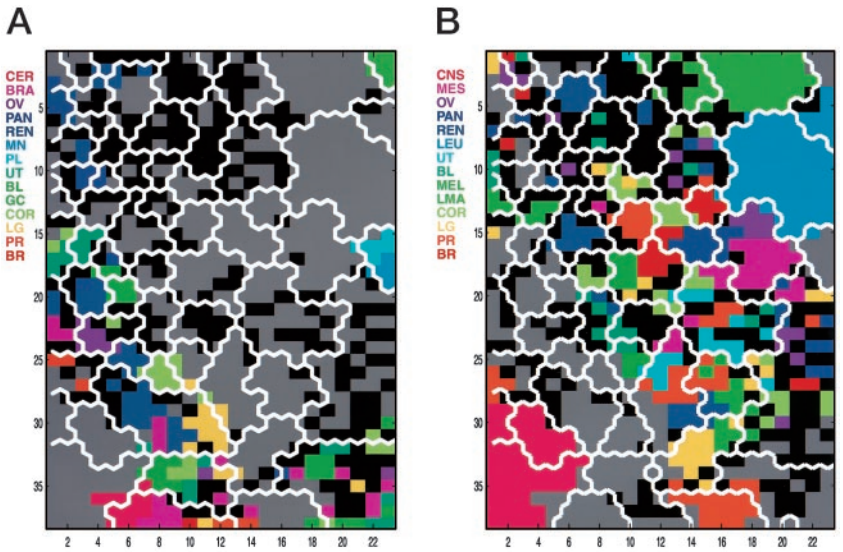

Fig. 6. Class predictions as described in Fig. 5, *B* and *C,* but annotated according to the 60 clades determined by hierarchical clustering of the SOM reference vectors. An average of 14 ± 9 reference vectors appear in each clade, and the average within clade correlation coefficient between data vectors is 0.86 ± 0.09.

cation results are based only on the highest probabilities, in reality, classification probabilities are calculated for all tumor classes, and thus alternative classifications are readily available with this approach.

Differentially Expressed Genes

Because the selection scheme for the cancer classifier genes is derived based only on their expression differential and not on any preconceived selection scheme, it is of interest to investigate these genes as potential marker genes or at least verify that they have been experimentally determined as being expressed for a particular cancer type. We have limited the investigated genes to only those genes that show the most markedly differential gene expression, *i.e.,* the genes that have a differential expression values with a $Z_{score} \geq 2$. The comparison is between successfully predicted cancer tissue types *versus* the pooled normal tissues. We examine only the cases where a gene is down-regulated in normal tissue and up-regulated in neoplasia. Using this scheme, there are, on the average, 39 marker genes for each tumor class.

The marker genes for the different tumor classes as well as for the tumor *versus* normal tissue show minimal overlap. The greatest overlap occurs between the leukemia and lymphoma dataset, which has 10 common genes out of a total of 151 selected genes, constituting a maximum overlap of 7%. More typically, 62% of all pairwise comparisons show no overlap between marker genes, indicating that the selected set is well differentiated and constitutes a nearly unique set of genes for each tumor class. It is this set that constitutes an experimentally verifiable expression profile characteristic of the cancer type. Although each single gene might not be a true unique marker for that particular cancer type, the entire profile could provide enough information to classify the tumor.

Classification of the two groups of pooled tumor tissues *versus* pooled normal tissues defines unique gene sets consistent with their distinctly different SOM projections and suggests a means for separation of normal tissue *versus* cancerous tissue from expressions alone. Marker genes that have been verified in the literature as being expressed in cancer and not overexpressed in normal tissue are listed in Table 3. As an example, among these genes we find prothymosin α (*PTMA*), which is associated with cell proliferation, to be overexpressed and used as a marker in both lung (26) and breast (27) cancer. The genes that span the classification set for the separation of tumor *versus* normal tissue are not the same as those that are found as predictors for each individual cancer type. These gene profiles could then be considered general cancer markers, not indicative of a single specific cancer type. This implies that marker genes for a general condition of cancer can be considered separately from individual cancer types. Genes that are down-regulated in cancer include *GSN* (gelsolin), which has been found to act as a tumor suppressor in breast cancer (28); *TGFBR3* (transforming growth factor β receptor III, β-glycan), another tumor suppressor that inhibits angiogenesis and tumor growth in human cancer cells (29); and *GDP1* (glycerol-3-phosphate dehydrogenase 1), which is found to be diminished in tumors (30). This illustrates the ability of the differential gene expression to accurately reflect a true gene expression/function in the investigated tissues.

For the case of highly differentially expressed genes among the tumor tissue classes, subsets of literature-verified tumor-specific genes can be found. Our analysis focuses on only genes identified as up-regulated in tumor tissue and down-regulated in normal tissue. This does not necessarily imply that an up-regulated gene is unique for that cancer type, just that up-regulation of that gene has been observed experimentally in that specific tumor type. These genes are given in Table 3 for each investigated cancer type. In bladder carcinoma, we find that the observed up-regulation of topoisomerase I (31), which relaxes supercoiled DNA, corresponds to an enhanced tumor/normal differential gene expression. An up-regulation of topoisomerase has also been correlated with drug resistance to camptothecin in bladder

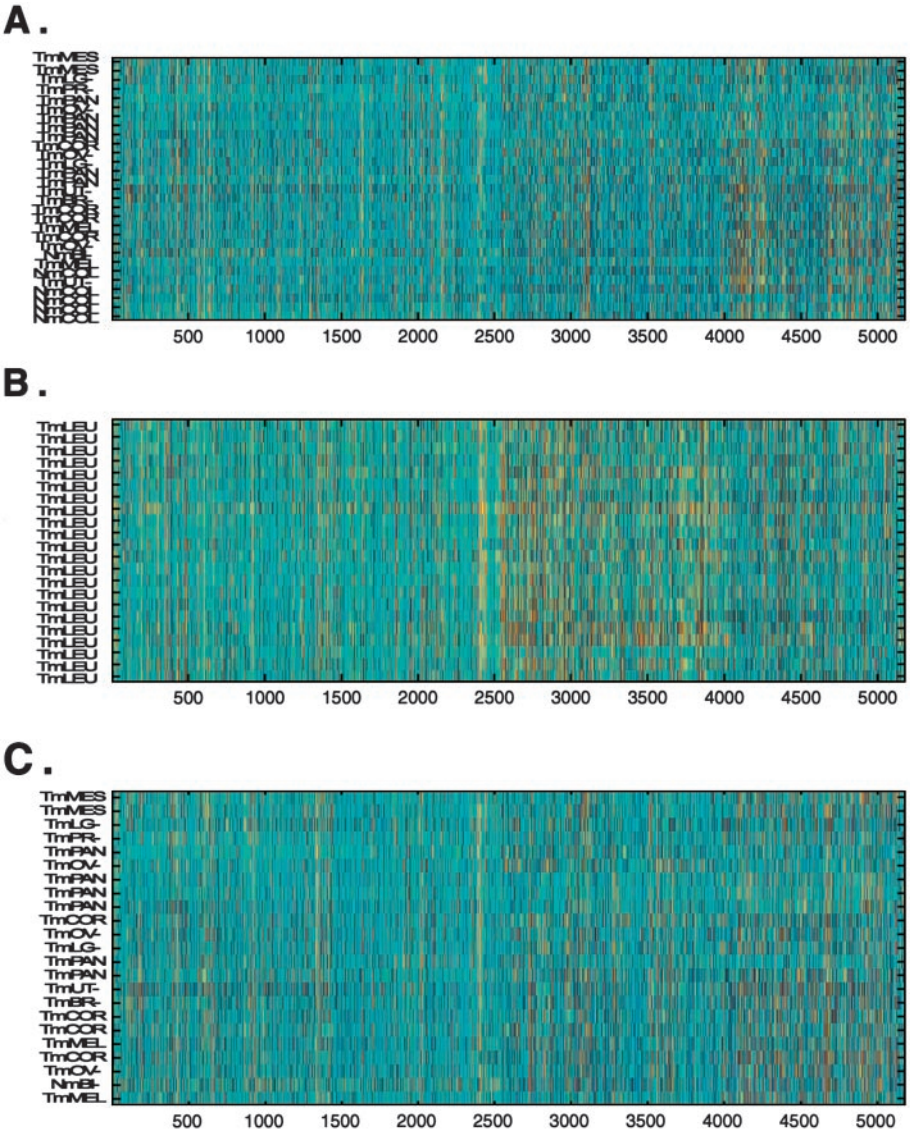

Fig. 7. Normalized data vectors for three clades. *B* consists of only leukemia data vectors, whereas *A* and *C* display separate clades containing mixtures of tumor and normal tissue classes. All of these groups are characterized by within-clade correlations of >0.75.

cancer cells (32). Few genes are selected for the breast cancer group, and among these, only the *IL1RL1* (interleukin 1 receptor-like 1) gene, which is thought to be part of the regulation machinery involved in protumorigenic activities, could be related to human breast cancer (33, 34). For CNS cancer, there are more genes selected, and we get a higher number of verifiable hits, among them *BSG* (basigin, CD147, EMMPRIN), stimulator of matrix metalloproteinase synthesis in fibroblasts, which had high mRNA expression levels in gliomas compared with nonneoplastic brain (35). *TEGT* (testis enhanced gene transcript, BAX inhibitor 1) was also found to be overexpressed in gliomas (36). Vascular endothelial growth factor (*VEGF*) is found to be overexpressed in many human cancers, where it plays a central role in tumor angiogenesis. It is also found in highly vascularized and infiltrative brain tumors (37, 38). The non-uniqueness of this classifier gene for CNS makes it a more general marker for many types of cancers. It is necessary but not sufficient that the identified markers are expressed in a particular cancer type. Thus, the predicted differential gene expression profile does not carry only single cancer markers, but it is the profile itself that defines the classifier. The two markers found to be correlated with colorectal cancer, *CYP2A6* (cytochrome P450, subfamily IIA polypeptide 6), which has been shown to have a strong relationship to colorectal cancer risk (39), and *KRT20* (cytokeratin 20), which is of prognostic value (40), are not general cancer markers but tend to be specific for colorectal cancer. *ADA* (adenosine deaminase) has been proposed as a novel marker for chronic lymphocytic leukemia by Martin *et al.* (41). This gene is also identified by our selection scheme as a differentially expressed gene typical for leukemia. Among other genes that we have identified as being highly expressed in leukemia are *CXCR4* [chemokine (C-X-C motif), receptor 4, fusin], *FLI1* (Friend leukemia virus integration 1), and *HMG1* [high-mobility group (nonhistone chromosomal) protein 1]. *CXCR4* has been identified as being highly ex-

Table 3 Selected genes with differential expression[a]

Cancer type	Selected marker genes
Bladder	*ACTN3*, *ASCL1*, *OCA2*, *PDGFRA*, ***TOP1***
Breast	***IL1RL1***, *JM4*, *PDE6C*
CNS	*ACK1*, *AES*, *AGRN*, *AP3B2*, *APCL*, *BACE*, ***BSG***, *CBX3*, ***CCND2***, *CDK2AP1*, *CPNE1*, *D1S155E*, *DBN1*, *DCTN4*, *DNCH1*, *EPB41L2*, *FKBP1A*, *FUS*, *GABBR1*, *GFAP*, *GSTA4*, *HRMT1L2*, *KIDINS220*, *KIF3B*, *KIF5C*, *KLF4*, *LANCL1*, *MAGE-E1*, ***MAPK8IP3***, *MYO10*, *NDRG2*, *NLGN2*, *NONO*, *NUDT3*, *PAI-RBP1*, *PCBP4*, *PEG3*, *PGRMC1*, *PPARAL*, *PPP2R1A*, *PURA*, *QKI*, *RANBP9*, *RBM8A*, *RTN4*, *S164*, *SAFB*, *SFPQ*, *SLC6A1*, *SNL*, *SOX4*, *STMN3*, ***TEGT***, *THRA*, *TRIM28*, *TRIM37*, *TRIO*, *TTYH1*, *UBL3*, ***VEGF***, *ZFR*, *ZIN*
Colorectal	*8D6A*, *AF093680*, *B3GAT1*, *CANX*, *CBL*, *CKTSF1B1*, *CMKLR1*, *CNK1*, ***CYP2A6***, *DELGEF*, *DLEU1*, *DRIM*, *GDF11*, *HAB1*, *HOXB1*, *HPCL2*, *HTRA3*, *IL20RA*, *KCNMA1*, *KIDINS220*, ***KRT20***, *LUC7L*, *MASP1*, *MRF-1*, *MRPS26*, *MTMR8*, *OSBPL1A*, *PDCD2*, *PHRET1*, *PMS2L8*, *PURB*, *RIN*, *RUVBL1*, *SARCOSIN*, *SH3GL3*, *SLB*, *TRRAP*
Leukemia	*ACTR2*, ***ADA***, *ANAPC5*, *ARHGDIB*, *ARPC3*, *ARRB2*, ***BCL7A***, *CHD1*, *CLIC1*, ***CXCR4***, *CYBA*, *D1S155E*, *DNAJB6*, *DUT*, *EEF1B2*, *EIF3S2*, *EIF3S5*, *EIF3S6*, *EIF4A1*, *EIF4B*, *EMP3*, ***FBL***, *FKBP1A*, ***FLI1***, *FXYD5*, *G8*, *GA17*, *GAB2*, *GMFG*, *H1FX*, *H2AFY*, *H2BFA*, ***H3F3B***, *H4FG*, *HDCMC04P*, ***HMG1***, ***HMG14***, *HNRPA1*, *HNRPD*, *HNRPH3*, *HRMT1L2*, *IGBP1*, *ILF2*, *LAPTM5*, *NAP1L1*, *NONO*, *NSEP1*, *NUP50*, *P114-RHO-GEF*, *PA2G4*, *PABPC1*, *PCBP2*, *PSMA6*, ***PSME1***, *PTP4A2*, *RANBP9*, ***RAP2B***, *ROD1*, *RPL10*, *RPL36*, *RYBP*, *SFPQ*, *SFRS11*, *SFRS6*, *TAPBP*, *TRGC2*, *UNRIP*, ***ZFP36L2***
Lung	*HCGIX*, *PKP3*, *PML*, ***SFTPA1***
Lymphoma	*ACTR2*, *ACTR3*, *ANAPC5*, *ARHF*, *ARHGDIB*, *ARPC3*, ***BATF***, *CCNDBP1*, ***CD37***, ***CD53***, *CDW52*, *CLP*, *CTSS*, *EIF3S6*, *ETS1*, *HLA-DPA1*, *HLA-DPB1*, *HLA-DQA1*, ***IGBP1***, ***MAT2B***, *MLLT7*, *MS4A6A*, *NCF1*, *NSEP1*, *NUP50*, *PSMB9*, ***PTTG1***, *SBB103*, *SCYA19*, ***SCYA5***, *SSNA1*, *STOML2*, *TRB*
Melanoma	*ATIC*, *BICD1*, *DOC-1R*, *EGLN2*, *GARS*, *JRK*, ***MCAM***, *MFGE8*, *MKPX*, ***NFYC***, *NME1*, ***OA1***
Mesothelioma	*54TM*, *AGRN*, *AP47*, *APOL3*, *BART1*, *BF*, *BRD4*, *C1R*, *C1S*, *C3*, *CEBPD*, *COL1A1*, *DAB2*, *EFEMP1*, *FSTL1*, *FXYD5*, *GAS6*, *GLS*, *GPX1*, *GSA7*, *HTGN29*, *IFITM1*, *ISYNA1*, *ITGB4*, *KRT18*, *KRT8*, *LDHA*, *LGALS3BP*, *LXN*, *MT2A*, *NRD1*, *PAI-RBP1*, *PRDX5*, *PSMB9*, *PTRF*, *RAB31*, *RAP2B*, *RBBP1*, *RBM8A*, *RPS26*, *S100A10*, *SAA1*, *SEC61A1*, *SECTM1*, *SPARC*, *SRRM2*, *TEAD4*, *TEM8*, *TYROBP*
Ovarian	—
Pancreas	*BMP5*, *E2F4*, *GP2*, *HTR6*, ***IGF1R***, *JM4*, ***MUC5AC***, *PRTN3*, *TNFRSF8*
Prostate	***ACPP***, *APM2*, *ATBF1*, *C8FW*, *CD9*, *CEBPD*, *CIRBP*, *CNN1*, ***CTSD***, *DMN*, *FOXP1*, *H2AFO*, ***HPN***, *HTPAP*, ***IDH1***, *IGFBP5*, *JUNB*, ***KLK3***, *LTF*, *MSMB*, *MYLK*, ***NDRG1***, *ODC1*, *PDE9A*, *RANBP2L1*, *RPN2*, *SERP1*, *SPON2*, *SPOP*, *TMP21*, *TRA1*, *TRGC2*, *TSC22*, *WRCH-1*, *XBP1*, *ZFP36*
Renal	***AKR1B1***, *ANGPTL4*, *ARAF1*, *DAZAP2*, *DNAJB1*, ***FGF3***, *GSTTLp28*, *H4FE*, *HAX1*, *HSPA9B*, ***IGFBP3***, *MAGEA3*, *MDS001*, *NEDD8*, *PARL*, *PGAM1*, *PGM1*, *PGM5*, *PPGB*, *R32184_1*, *SLC6A3*, *SNRPB*, *STC1*, ***VIL2***, *VNN1*, *ZF*
Uterine	*APLP2*, *B4GALT1*, *B7*, *BICD1*, *BSCL2*, *C1QR1*, *C3*, *CARM1*, *CLDN3*, *DPP3*, *GOSR1*, *GPX1*, ***GSTP1***, *HOXB5*, *IFI27*, *IFITM1*, *ISYNA1*, *JM4*, *LGALS3BP*, *LSM2*, *MCAM*, *MFAP2*, *MGC2835*, *MOG1*, *MSX1*, *NME1*, *NOSIP*, *PSME2*, *PTK7*, *PVRL2*, *R32184_1*, *RES4-25*, *RNASE6PL*, *RPN2*, *RUVBL2*, *SBB103*, *SOX4*, *SSR2*, *TJP3*, *TMEM4*, *TMSB10*, *ZNF161*
Tumor/normal	***APLP2***, *BICD1*, *C1QR1*, *CNOT2*, ***DGCR5***, ***DGSI***, *HRMT1L2*, *HSKM-B*, *HTRA3*, ***LAG3***, *PA2G4*, *PARL*, ***PTMA***, *RANBP16*, *SBB103*, *TJP3*

[a] Selected genes that showed differential gene expression by being up-regulated in tumor cells and down-regulated in normal cells. Gene expressions that could be verified via literature searches are marked in bold. The gene abbreviations refer to the HUGO name of the gene.

pressed and partakes in extramedullary invasion in childhood acute lymphoblastic leukemia (42). *FLI1* overexpression has been linked to the etiology of a number of virally induced leukemias (43). *HMG1* was investigated by Cabart *et al.* (44), who found that expression of *HMG1* was higher in malignant leukemia cell lines compared with lymphocytes and granulocytes. *SFTPA1* (surfactant, pulmonary-associated protein A1) identified by us as being a differentially expressed marker for lung cancer was used by Zamecnik and Kodet (45) to distinguish primary and metastatic lung carcinomas from a broad range of nonpulmonary tumors. We identify a range of genes that have also been corroborated in the literature as being overexpressed in lymphoma. These include *BATF* (basic leucine zipper transcription factor, ATF-like), which has been found to be expressed in both lung and Raji Burkitt's lymphoma (46); *CD37* (CD37 antigen), a tetraspanin-expressing gene found in Burkitt's lymphoma (47); *IGBP1* [immunoglobulin (CD79A)-binding protein 1] found in T-lymphoblastic lymphoma as a cell marker (48); and *PTTG1* (pituitary tumor-transforming 1), which was found to be overexpressed in human T-lymphoma cell lines as well as in samples from patients with different kinds of hematopoietic malignancies (49). Among the melanoma genes, we find *MCAM* (melanoma cell adhesion molecule), which is a cell-cell adhesion receptor highly expressed by melanoma cells but not normal melanocytes (50). None of the identified

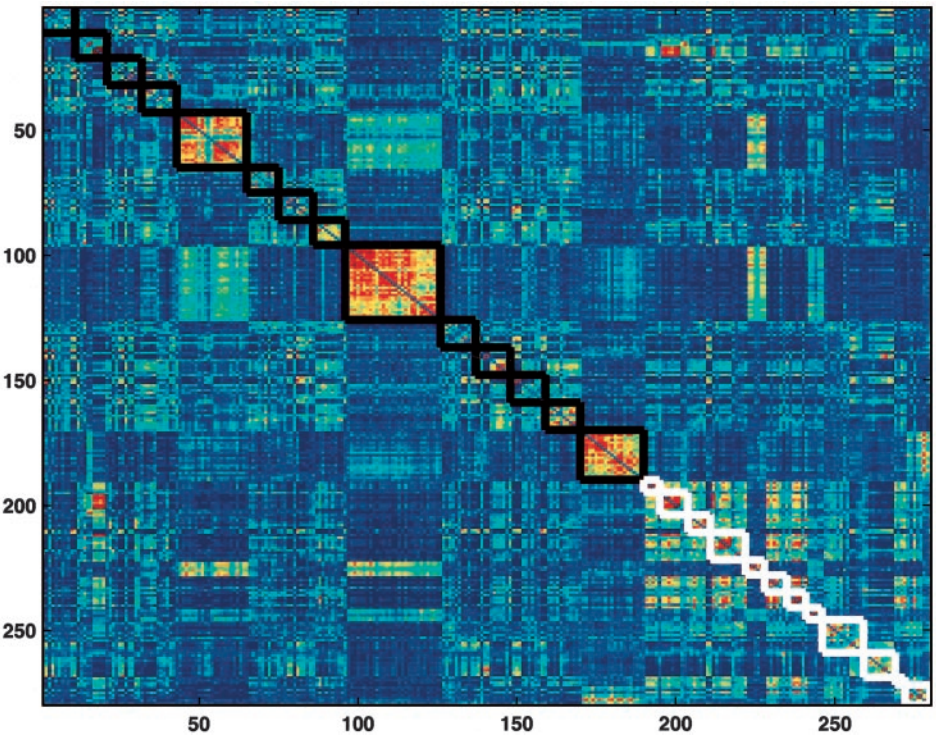

Fig. 8. Plot of correlation coefficients for the set of 1139 genes having the greatest magnitude of differential expression when compared with the pooled normal set. Data are ordered from tumor (1–190) to normal (191–280) tissue types. Axis positions corresponding to tumor and normal classes appear as *blocks along the diagonal,* with the tumor subset shown in *black lines,* and the normal subset shown in *white lines*. The blocks are ordered from the *top left* to *bottom right* as tumor (breast, prostate, lung, colorectal, lymphoma, melanoma, bladder, uterine, leukemia, renal, pancreatic, mesothelioma, and CNS) followed by normal (breast, prostate, lung, colon, germinal center, bladder, uterine, peripheral lymphocytes and monocytes, kidney, pancreas, ovarian, and brain + cerebellum). The most positive and negative correlations are shown in *red* and *blue,* respectively.

genes in mesothelioma or ovarian cancer could be verified in the literature as being overexpressed, pointing to the relative scarceness of information on differential gene expression in cancer. In pancreatic cancer we identified *MUC5AC* (mucin 5, subtypes A and C, tracheobronchial/gastric) as being a potential marker gene; commensurate with this notion, Ho *et al.* (51) found that this gene was not expressed in normal pancreas but was expressed in tumors. Among the identified differentially expressed genes in prostate tumors, the nonspecific *NDRG1* (N-myc downstream regulated, Cap43) gene appears. This marker gene has been found to be expressed at low levels in normal tissue and overexpressed in a variety of cancers, including lung, brain, liver, prostate, breast, renal and melanoma (52). A more specific prostatic marker gene is the human prostatic acid phosphatase (*ACPP*). This gene was shown to be significantly elevated in neoplastic tissue compared with benign prostatic hyperplasia (53) and is listed in Table 3 as being identified here as a potential prostatic cancer marker gene. *HPN* (hepsin), a transmembrane serine protease, has also been associated as a marker for prostate cancer (54). *Kallikrein 3* [prostate-specific antigen (*PSA*)], the most widely used indicator for prostate cancer, was also identified by our procedure as a potential marker gene that is up-regulated in tumors and down-regulated in normal tissue. *IGFBP3* (insulin-like growth factor-binding protein 3) is one potential marker given in Table 3 that has been shown to be markedly increased in renal carcinoma tissues compared with normal kidney samples (55). For uterine cancer, we identified *GSTP1* (glutathione *S*-transferase π) as being a potential marker, which has been corroborated in the literature by Osmak *et al.* (56), who found significantly higher glutathione *S*-transferase activity in tumor tissue compared with corresponding normal tissue. The general literature verification rate of up-regulated genes in tumor tissue is around 8%, indicating that the selected markers from the classification scheme identify many more genes that could be experimentally verified as possible novel biomarkers for neoplasia.

Selection of those genes that are most different from the pooled normal set reduces the 5183 genes in the initial filtered set to 1139 genes. Fig. 8 displays the correlations between these genes for all tissue types. Tissue types are ordered from tumor (1–190) to normal (191–280), and the most positive and negative correlations are shown in *red* and *blue,* respectively. Examination of the normal tissue expressions (at the *bottom right corner* of this figure) for these 1139 genes shows positive correlations among the normal tissue expressions for all but the normal germinal and pancreatic tissues and portions of the normal kidney tissues, where their expressions are mostly negatively correlated with the pooled normal set. Otherwise the pooled normal tissues comprise a relatively homogeneous set of gene expressions for this reduced set of genes. Clearly evident from this figure are the shared tissue expressions for lymphoma and leukemia, shown as the off-diagonal blocks of positive correlated genes appearing in *yellow*. It is interesting to note that these tumor tissues share positive correlations with the normal germinal tissues as well as monocytes and peripheral lymphocytes. Evident within the tumor tissues are strong positive gene correlations; the most evident are observed for the tissue types with the largest number of observations (lymphoma, leukemia, and CNS). As noted above, these coherent

Table 4 Predictions using earlier datasets[a]

	N[b]	UC	BR	PR	LG	COR	LMA	MEL	BL	UT	LEU	REN	PAN	OV	MES	CNS	%C
LEU_{ind}[c]	34	7	6	1	0	0	5	1	2	0	**10**	0	0	0	1	1	37
LEU_{tr}[d]	38	1	3	0	0	0	3	2	3	0	**26**	0	0	0	0	0	70
CNS_A[e]	42	4	1	1	0	0	14	4	1	0	2	0	0	0	4	**11**	29
CNS_B	34	6	2	0	0	0	5	0	1	0	4	0	1	0	3	**12**	43
CNS_C	60	13	7	1	1	0	12	3	0	0	5	0	0	0	3	**15**	32
LMA_1[f]	77	15	5	5	3	2	**20**	6	3	1	9	0	0	0	6	2	32
LMA_2	58	14	5	2	1	1	**14**	4	4	1	5	0	0	0	4	3	32

[a] All data were downloaded from the Whitehead site (www-genome.wi.mit.edu/MPR). Analyzed data represent the expression values in Affymetrix's scaled average difference units, where the average difference values are calculated using Affymetrix's GeneChip software.
[b] Entries in row legend correspond to: N, number of samples; UC, number of unclassifiable cases; BR, breast adenocarcinoma; PR, prostate adenocarcinoma; LG, lung adenocarcinoma; COR, colorectal adenocarcinoma; LMA, lymphoma; MEL, melanoma; BL, bladder cell transitional carcinoma; UT, uterine adenocarcinoma; LEU, leukemia; REN, renal cell adenocarcinoma; PAN, pancreatic adenocarcinoma; OV, ovarian adenocarcinoma; MES, pleural mesothelioma; %C, percentage of correct predictions.
[c] Leukemia training (Ref. 3).
[d] Leukemia test set (Ref. 3).
[e] Raw CNS data appear as three downloadable files, designated A, B, and C (Ref. 57).
[f] Dataset 1 consists of 77 data records of diffuse large B-cell and follicular lymphoma classes. Dataset 2 consists of 58 diffuse large B-cell lymphoma cases (Ref. 58).

gene expressions within each tissue type provide sets of marker genes for tissue classifications. The fuzzy nature of these expressions, indicative of strong off-diagonal patterns of positive and negative correlated genes, reveals the considerable number of shared gene expressions between tissue types. It is the complete set of gene expressions that, apparently, provides sufficient information for classifications.

Discussion

Our analysis demonstrates that an unsupervised SOM-based clustering strategy can be used to classify tissue samples from an oligonucleotide microarray patient database. Classification accuracies on the order of ~80% correct can be achieved with this method, a level of classification accuracy equivalent to that achieved on this same dataset using two different methods (2, 11). Our method is based on the likelihood that a test data vector may have a gene expression fingerprint that is shared by more than one tumor class and as such can identify datasets that cannot be unequivocally assigned to a single tumor class. Datasets with nearly uniform probabilities for more than one tumor class are regarded as unclassifiable. Although we cannot provide a basis for similar gene expressions between two different tumor classes, it is not unreasonable that cross-contaminations within and between neoplastic and nonneoplastic tissues must exist (2).

Our results are further used to identify sets of differentially expressed genes within each tumor class. Genes with the greatest difference from the pooled normal gene expressions yield a set of 1139 genes. This subset of genes is further divided by tumor class into 14 mostly nonoverlapping gene sets. Examination of these groups finds tumor class-specific evidence for ~10% of these genes. The remaining sets of genes appear to be largely unexplored with respect to their role in specific cancers and their possibilities as molecular targets for therapy.

Studies of the type documented here must be highly validated before algorithmic approaches can be considered on an equal footing with subjective approaches. Efforts to validate our method were conducted by classifying previously published datasets from the Whitehead group. Examinations were completed for leukemia (3), CNS (57), and lymphoma (58) datasets obtained with the earlier versions of the 6,500 and 12,500 Affymetrix chips. The results are listed in Table 4. In general, only ~30–40% of the 5183 genes in our filtered set were also found on these smaller chips. Despite this limited number of genes, an average of 40% of these data was correctly classified. The highest accuracies were found for the leukemia training set, with 26 of the 38 data vectors having correct assignments. In general, however, the prediction accuracies were 50% poorer than those reported in Table 2. Establishing the reasons for these differences is difficult; however, the limited number of genes in each test dataset most likely contributes to the poorer classification accuracies.

A recent study by Su *et al.* (4) provides additional data for tissue classifications. They propose a molecular classification scheme using support vector machines that analyzed 175 carcinoma datasets obtained from the U95A Affymetrix chip of ~12,000 genes. Only 2,249 genes of their published data are in the filtered set of 5,183 genes used in our analysis. Analysis of these data using our prediction scheme finds that 72 of their 174 samples are considered unclassifiable according to our criteria. A portion of these samples occurs for tumor classes that do not exist in the Whitehead set and thus had no equivalent tumor class. Among the classifiable datasets, all of them were found to fall into six tumor classes: prostate cancer; lymphoma; uterine cancer; leukemia; renal cancer; and mesothelioma. Of these classifications, 13 of the 26 prostate tissues were correctly assigned. In addition, 11 of the 26 lung adenocarcinoma and squamous cell carcinomas were classified as mesothelioma, whereas 8 of the 26 ovarian carcinomas were classified in our uterine cancer class. Comparisons beyond these few cases yield little correspondence between histologically assigned tumor class and classification prediction. Our general conclusion from these results is that test cases based on data generated from another laboratory's database provide rea-

sonable cross-validations only for a few cases. The inability to achieve, with our method and their data, the very high prediction accuracies found by Su *et al.* (4) raises concerns about conducting examinations of gene expression datasets derived from different sources. Although acceptable prediction accuracies were found in our study and that of the Whitehead group, using, respectively, unsupervised and supervised approaches, extensions of these methodologies to other datasets must be further evaluated. Previous attempts to compare expression datasets from different laboratories found their concordance to be quite low (59, 60). It is not unreasonable to assume that such difficulties will limit the development of a "universal" database and a general scheme for tumor class predictions.

Acknowledgments

We thank Sridhar Ramaswamy for generous sharing of gene expression data and helpful discussions on the topic of classification.

References

1. Hanahan, D., and Weinberg, R. The hallmark of cancer. Cell, *100:* 57–71, 2000.

2. Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C-H., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J. P., Poggio, T., Geraldand, W., Loda, M., Lander, E. S., and Golub, T. R. Multiclass cancer diagnosis using tumor gene expression signatures. Proc. Natl. Acad. Sci. USA, *98:* 15149–15154, 2001.

3. Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science (Wash. DC), *286:* 531–537, 1999.

4. Su, A. I., Welsh, J. B., Sapinoso, L. M., Kern, S. G., Dimitrov, P., Lapp, H., Schultz, P. G., Powell, S. M., Moskaluk, C. A., Frierson, H. F., Jr., and Hampton, G. M. Molecular classification of human carcinomas by use of gene expression signatures. Cancer Res., *61:* 7388–7393, 2001.

5. Bittner, M., Meltzer, P., Chen, Y., Jiang, Y., Seftor, E., Hendrix, M., Radmacher, M., Simon, R., Yakhini, Z., Ben-Dor, A., Sampas, N., Dougherty, E., Wang, E., Marincola, F., Gooden, C., Lueders, J., Glatfelter, A., Pollock, P., Carpten, J., Gillanders, E., Leja, D., Dietrich, K., Beaudry, C., Berens, M., Alberts, D., Sondak, V., Hayward, N., and Trent, J. Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature (Lond.), *406:* 536–540, 2000.

6. Perou, C. M., Sorlie, T., Eisen, M. B., van der Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S. X., Lonning, P. E., Borresen-Dale, A-L., Brown, P. O., and Botstein, D. Molecular portraits of human breast tumours. Nature (Lond.), *406:* 747–752, 2000.

7. Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA, *95:* 14863–14868, 1998.

8. Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S., and Golub, T. R. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc. Natl. Acad. Sci. USA, *96:* 2907–2912, 1999.

9. Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson, J., Jr., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. E., Armitage, J. O., Warnke, R., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O., and Staudt, L. M. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature (Lond.), *403:* 503–511, 2000.

10. Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C., and Meltzer, P. S. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat. Med., *7:* 673–679, 2001.

11. Yeang, C-H., Ramaswamy, S., Tamayo, P., Mukherjee, S., Rifkin, R., Angelo, M., Reich, M., Lander, E., Mesirov, J. P., and Golub, T. R. Molecular classification of multiple tumor types. Bioinformatics, *17* (Suppl.)*:* S316−S322, 2001.

12. Allwein, E., Shapire, R., and Singer, Y. Reducing multiclass to binary: a unifying approach for margin classifiers. *In:* Proceedings of the Seventeenth International Conference on Machine Learning, pp. 9–12, 2000.

13. Chapelle, O., Vapnik, V.,Bousquet, O. and Mukherjee, S. Machine Learning, *46* (1–3): 131–159, 2002.

14. Hastie, T., Tiibshirani, R., Eisen, M. B., Alizadeh, A., Levy, R., Staudt, L., Chan, W. C., Botstein, D., and Brown, P. Gene shaving as a method for identifying distinct sets of genes with similar expression patterns. Genome Biol., *1:* research002.1−research003.21, 2000.

15. Kohonen, T. Self-Organizing Maps. Germany: Springer Verlag, 1995.

16. Sneath, P. H. A., and Sokal, R. R. Numerical Taxonomy. San Francisco: W. H. Freeman and Company, 1973.

17. Becker, R. A., Chambers, J. M., and Wilks, A. S. A Language and System for Data Analysis. Murray Hill, NJ: Bell Laboratories Computer Information Services, 1981.

18. Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J., and Church, G. M. Systematic determination of genetic network architecture. Nat. Genet., *22:* 281–285, 1999.

19. Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., and Levine, A. J. 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, 1998.

20. Toronen, P., Kolehmainen, M., Wong, G., and Castren, E. Analysis of gene expression data using self-organizing maps. FEBS Lett., *451:* 142–146, 1999.

21. Giuliani, A., Colosimo, A., Benigni, R., and Zbilut, J. On the constructive role of noise in spatial systems. Physics Letters A, *247:* 47–52, 1998.

22. Keskin, O., Bahar, I., Jernigan, R. L., Beutler, J. A., Shoemaker, R. H., Sausville, E. A., and Covell, D. G. Characterization of anticancer agents by their growth-inhibitory activity and relationships to mechanism of action and structure. Anticancer Drug Discovery, *15:* 79–98, 2000.

23. Berry, M. W., Dumais, S. T., and O'Brien, G. W. Using linear algebra for intelligent information retrieval. SIAM Rev., *37:* 573–595, 1995.

24. Rabow, A. A., Shoemaker, R. H., Sausville, E. A., and Covell, D. G. Mining the National Cancer Institute's tumor-screening database: identification of compounds with similar cellular activities. J. Med. Chem., *45:* 818–840, 2002.

25. Hand, D. A. Discrimination and Classification. New York: John Wiley and Sons, 1981.

26. Sasaki, H., Nonaka, M., Fujii, Y., Yamakawa, Y., Fukai, I., Kiriyama, M., and Sasaki, M. Expression of the prothymosin-α gene as a prognostic factor in lung cancer. Surg. Today, *31:* 936–938, 2001.

27. Magdalena, C., Dominguez, F., Loidi, L., and Puente, J. L. Tumour prothymosin α content, a potential prognostic marker for primary breast cancer. Br. J. Cancer, *82:* 584–590, 2000.

28. Dong, Y., Asch, H. L., Ying, A., and Asch, B. B. Molecular mechanism of transcriptional repression of gelsolin in human breast cancer cells. Exp. Cell Res., *276:* 328–336, 2002.

29. Bandyopadhyay, A., Zhu, Y., Malik, S. N., Kreisberg, J., Brattain, M. G., Sprague, E. A., Luo, J., Lopez-Casillas, F., and Sun, L. Z. Extracellular domain of TGFβ type III receptor inhibits angiogenesis and tumor growth in human cancer cells. Oncogene, *21:* 3541–3551, 2002.

30. Balabanov, S., Zimmermann, U., Protzel, C., Scharf, C., Klebingat, K. J., and Walther, R. Tumour-related enzyme alterations in the clear cell type of human renal cell carcinoma identified by two-dimensional gel electrophoresis. Eur. J. Biochem., *268:* 5977–5980, 2001.

31. Monnin, K. A., Bronstein, I. B., Gaffney, D. K., and Holden, J. A. Elevations of DNA topoisomerase I in transitional cell carcinoma of the

urinary bladder: correlation with DNA topoisomerase II-α and p53 expression. Hum. Pathol., *30:* 384–391, 1999.

32. Kotoh, S., Naito, S., Yokomizo, A., Kumazawa, J., Asakuno, K., Kohno, K., and Kuwano, M. Increased expression of DNA topoisomerase I gene and collateral sensitivity to camptothecin in human cisplatin-resistant bladder cancer cells. Cancer Res., *54:* 3248–3252, 1994.

33. Paciotti, G. F., and Tamarkin, L. Interleukin-1 directly regulates hormone-dependent human breast cancer cell proliferation *in vitro*. Mol. Endocrinol., *2:* 459–464, 1988.

34. Miller, L. J., Kurtzman, S. H., Anderson, K., Wang, Y., Stankus, M., Renna, M., Lindquist, R., Barrows, G., and Kreutzer, D. L. Interleukin-1 family expression in human breast cancer: interleukin-1 receptor antagonist. Cancer Investig., *18:* 293–302, 2000.

35. Sameshima, T., Nabeshima, K., Toole, B. P., Yokogami, K., Okada, Y., Goya, T., Koono, M., and Wakisaka, S. Expression of emmprin (CD147), a cell surface inducer of matrix metalloproteinases, in normal human brain and gliomas. Int. J. Cancer, *88:* 21–27, 2000.

36. Schmits, R., Cochlovius, B., Treitz, G., Regitz, E., Ketter, R., Preuss, K. D., Romeike, B. F., and Pfreundschuh, M. Analysis of the antibody repertoire of astrocytoma patients against antigens expressed by gliomas. Int. J. Cancer, *98:* 73–77, 2002.

37. Chaudhry, I. H., O'Donovan, D. G., Brenchley, P. E., Reid, H., and Roberts, I. S. Vascular endothelial growth factor expression correlates with tumour grade and vascularity in gliomas. Histopathology, *39:* 409–415, 2001.

38. Och, W., Mariak, Z., Smolka, M., Badowski, J., and Koziorowski, W. Vascular endothelial growth factor expression in cerebral neoplasms. Neurol. Neurochir. Pol., *35:* 1071–1079, 2001.

39. Nowell, S., Sweeney, C., Hammons, G., Kadlubar, F. F., and Lang, N. P. CYP2A6 activity determined by caffeine phenotyping: association with colorectal cancer risk. Cancer Epidemiol. Biomark. Prev., *11:* 377–383, 2002.

40. Rosenberg, R., Hoos, A., Mueller, J., Baier, P., Stricker, D., Werner, M., Nekarda, H., and Siewert, J. R. Prognostic significance of cytokeratin-20 reverse transcriptase polymerase chain reaction in lymph nodes of node-negative colorectal cancer patients. J. Clin. Oncol., *20:* 1049–1055, 2002.

41. Martin, M., Aran, J. M., Colomer, D., Huguet, J., Centelles, J. J., Vives-Corrons, J. L., and Franco, R. Surface adenosine deaminase. A novel B-cell marker in chronic lymphocytic leukemia. Hum. Immunol., *42:* 265–273, 1995.

42. Crazzolara, R., Kreczy, A., Mann, G., Heitger, A., Eibl, G., Fink, F. M., Mohle, R., and Meister, B. High expression of the chemokine receptor CXCR4 predicts extramedullary organ infiltration in childhood acute lymphoblastic leukaemia. Br. J. Haematol., *115:* 545–553, 2001.

43. Truong, A. H., and Ben-David, Y. The role of Fli-1 in normal cell function and malignant transformation. Oncogene, *19:* 6482–6489, 2000.

44. Cabart, P., Kalousek, I., Jandova, D., and Hrkal, Z. Differential expression of nuclear HMG1, HMG2 proteins and H1(zero) histone in various blood cells. Cell Biochem. Funct., *13:* 125–133, 1995.

45. Zamecnik, J., and Kodet, R. Value of thyroid transcription factor-1 and surfactant apoprotein A in the differential diagnosis of pulmonary carcinomas: a study of 109 cases. Virchows Arch., *440:* 353–361, 2002.

46. Dorsey, M. J., Tae, H. J., Sollenberger, K. G., Mascarenhas, N. T., Johansen, L. M., and Taparowsky, E. J. B-ATF: a novel human bZIP protein that associates with members of the AP-1 transcription factor family. Oncogene, *11:* 2255–2265, 1995.

47. Ferrer, M., Yunta, M., and Lazo, P. A. Pattern of expression of tetraspanin antigen genes in Burkitt lymphoma cell lines. Clin. Exp. Immunol., *113:* 346–352, 1998.

48. Thomas, R., Smith, K. C., Gould, R., Gower, S. M., Binns, M. M., and Breen, M. Molecular cytogenetic analysis of a novel high-grade canine T-lymphoblastic lymphoma demonstrating co-expression of CD3 and CD79a cell markers. Chromosome Res., *9:* 649–657, 2001.

49. Dominguez, A., Ramos-Morales, F., Romero, F., Rios, R. M., Dreyfus, F., Tortolero, M., and Pintor-Toro, J. A. hpttg, a human homologue of rat pttg, is overexpressed in hematopoietic neoplasms. Evidence for a transcriptional activation function of hPTTG. Oncogene, *17:* 2187–2193, 1998.

50. Satyamoorthy, K., Muyrers, J., Meier, F., Patel, D., and Herlyn, M. MelCAM-specific genetic suppressor elements inhibit melanoma growth and invasion through loss of gap junctional communication. Oncogene, *20:* 4676–4684, 2001.

51. Ho, J. J., Crawley, S., Pan, P. L., Farrelly, E. R., and Kim, Y. S. Secretion of MUC5AC mucin from pancreatic cancer cells in response to forskolin and VIP. Biochem. Biophys. Res. Commun., *294:* 680–686, 2002.

52. Cangul, H., Salnikow, K., Yee, H., Zagzag, D., Commes, T., and Costa, M. Enhanced expression of a novel protein in human cancer cells: a potential aid to cancer diagnosis. Cell Biol. Toxicol., *18:* 87–96, 2002.

53. Li, S. S. Human prostatic acid phosphatase and prostate specific antigen: protein structure, gene organization, and expression in neoplastic and benign tissues. Kaohsiung J. Med. Sci., 12, 441–447, 1996.

54. Ernst, T., Hergenhahn, M., Kenzelmann, M., Cohen, C. D., Bonrouhi, M., Weninger, A., Klaren, R., Grone, E. F., Wiesel, M., Gudemann, C., Kuster, J., Schott, W., Staehler, G., Kretzler, M., Hollstein, M., and Grone, H. J. Decrease and gain of gene expression are equally discriminatory markers for prostate carcinoma: a gene expression analysis on total and microdissected prostate tissue. Am. J. Pathol., *160:* 2169–2180, 2002.

55. Hintz, R. L., Bock, S., Thorsson, A. V., Bovens, J., Powell, D. R., Jakse, G., and Petrides, P. E. Expression of the insulin like growth factor-binding protein 3 (IGFBP-3) gene is increased in human renal carcinomas. J. Urol., *146:* 1160–1163, 1991.

56. Osmak, M., Babic, D., Abramic, M., Ambriovic, A., Milicic, D., Eljuga, D., and Vukovic, L. Glutathione *S*-transferase activity as an early marker for malignant tumors of corpus uteri. Neoplasma, *44:* 324–328, 1997.

57. Pomeroy, S. L., Tamayo, P., Gaasenbeek, M., Sturla, L. M., Angelo, M., McLaughlin, M. E., Kim, J. H., Goumnerovak, L. C., Black, P. M., Lau, C., Allen, J. C., Zagzag, D., Olson, J. M., Curran, T., Wetmore, C., Biegel, J. A., Poggio, T., Mukherjee, S., Rifkin, R., Califano, A., Stolovitzky, G., Louis, D. M., Mesirov, J. P., Lander, E. S., and Golub, T. M. Prediction of central nervous system embroynal tumour outcome based on gene expression pro-filing. Nature (Lond.), *415:* 436–442, 2002.

58. Shipp, M. A., Ross, K. N., Tamayo, P., Weng, A. P., Kutok, J. L., Aguiar, R. C. T., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G. S., Ray, T. S., Koval, M. A., Last, K. W., Norton, A., Lister, T. A., Mesirov, J., Neuberg, D. S., Lander, E. S., Aster, J. C., and Golub, T. R. Diffuse large cell profiling and supervised machine learning. Nat. Med., *8:* 68–74, 2002.

59. Wallqvist, A., Rabow, A. A., Shoemaker, R. H., Sausville, E. A., and Covell, D. G. Establishing connections between microarray expression data and chemotherapeutic cancer pharmacology. Mol. Cancer Ther., *1:* 311–320, 2002.

60. Kuo, W. P., Jenssen, T. K., Butte, A. J., Ohno-Machado, L., and Kohane, I. S. Analysis of matched mRNA measurements from two different microarray technologies. Bioinformatics, *18:* 405–412, 2002.