

Identification of Combination Gene Sets for Glioma Classification¹

Seungchan Kim, Edward R. Dougherty,
Ilya Shmulevich, Kenneth R. Hess,
Stanley R. Hamilton, Jeffrey M. Trent,
Gregory N. Fuller, and Wei Zhang²

Department of Electrical Engineering, Texas A&M University, College Station, Texas 77840 [S. K., E. R. D.]; Departments of Pathology [E. R. D., I. S., S. R. H., G. N. F., W. Z.] and Biostatistics [K. R. H.], The University of Texas M. D. Anderson Cancer Center, Houston, Texas 77030; and Cancer Genetics Branch, National Human Genome Research Institute, NIH, Bethesda, Maryland 20892-4470 [S. K., J. M. T.]

Abstract

One goal for the gene expression profiling of cancer tissues is to identify signature genes that robustly distinguish different types or grades of tumors. Such signature genes would ideally provide a molecular basis for classification and also yield insight into the molecular events underlying different cancer phenotypes. This study applies a recently developed algorithm to identify not only single classifier genes but also gene sets (combinations) for use as glioma classifiers. Classifier genes identified by this algorithm are shown to be strong features by conservatively and collectively considering the misclassification errors of the feature sets. Applying this approach to a test set of 25 patients, we have identified the best single genes and two- to three-gene combinations for distinguishing four types of glioma: (a) oligodendroglioma; (b) anaplastic oligodendroglioma; (c) anaplastic astrocytoma; and (d) glioblastoma multiforme. Some of the identified genes, such as insulin-like growth factor-binding protein 2, have been confirmed to be associated with one of the tumor types. Using combinations of genes, the classification error rate can be significantly lowered. In many instances, neither of the individual genes of a two-gene set performs well as an accurate classifier, but the combination of the two genes forms a robust classifier with a small error rate. Two-gene and three-gene combinations thus provide robust classifiers possessing the potential to translate expression microarray results into diagnostic histopathological assays for clinical utilization.

Introduction

Current estimates suggest that there are approximately 30,000–40,000 genes in the human genome (1, 2), and subsets of those genes are expressed in different cell types and in different cellular states. The combination of expressed genes at different levels determines the overall physiology of the cell. Two primary goals of functional genomics are to screen for, from amid the massive amount of transcriptomic data generated by high-throughput cDNA microarray technology, the key genes and gene combinations that explain specific cellular phenotypes (*e.g.*, disease) on a mechanistic level and to use this data to classify diseases on a molecular level (3–7).

An important consideration is that the number of genes in such gene feature sets should be sufficiently small so as to be potentially useful for clinical diagnosis/prognosis or as candidates for functional analysis to determine whether they could serve as useful targets for therapy. A number of classification approaches have been used to exploit the class-separating power of expression data; however, the size of the gene sets (sometimes as large as 70) renders the construction of practical immunohistochemical diagnostic/prognostic panels and the experimental design for functional testing problematic (3, 8, 9).

We use a recently proposed algorithm to identify strong gene feature sets that are responsible for distinct patient groups (10). These gene sets are “strong” in the sense that the algorithm builds classifiers from a probability distribution resulting from spreading the mass of the sample points to make the classification more difficult, while maintaining sample geometry. In an effort to identify the strong feature genes among the different histological diagnoses in patients with gliomas, we applied this method, in a proof-of-principle study, to glioma tissue specimens from 25 patients with four different types of glioma: (a) GM;³ (b) AA; (c) AO; and (d) low-grade OL. After finding the sets of genes that are capable of accurately classifying the different types of glioma, we have also identified strong features (genes) that are seemingly responsible for the distinct phenotype of each type of cancer.

Gliomas are the most common malignant primary brain tumors (11, 12). These tumors are derived from neuroepithelial cells and can be divided into two principal lineages: astrocytomas and OLs. Current glioma classification schemes are based on morphological feature assessment and remain highly subjective and problematic for many atypical cases. Diagnoses are often dependent on the relative

Received 2/14/02; revised 8/30/02; accepted 9/30/02.

¹ Supported in part by the Tobacco Settlement Funds as appropriated by the Texas State Legislature, by a generous donation from the Michael and Betty Kadoorie Foundation, and by a grant from the Texas Higher Education Coordination Board (Grant 003657-0039-1999).

² To whom requests for reprints should be addressed, at Cancer Genomics Core Laboratory, Department of Pathology, Box 85, The University of Texas M. D. Anderson Cancer Center, 1515 Holcombe Boulevard, Houston, TX 77030. Phone: (713) 745-1103; Fax: (713) 792-5549; E-mail: wzhang@mdanderson.org.

³ The abbreviations used are: GM, glioblastoma multiforme; OL, oligodendroglioma; AO, anaplastic oligodendroglioma; AA, anaplastic astrocytoma; IGF2BP2, insulin-like growth factor-binding protein 2; LOO, leave-one-out.

weighting of specific morphological features by individual pathologists. We reason that by identification of robust signature gene classifiers using typical cases, the atypical cases can be classified based on the signature classifier genes in the future.

Materials and Methods

Primary Glioma Tissues. All primary glioma tissues were acquired from the Brain Tumor Center tissue bank of The University of Texas M. D. Anderson Cancer Center. Tissue bank specimens were quick-frozen shortly after surgical removal and stored at -80°C . Although it is not known whether or to what extent the time delay between tumor removal and tumor freezing affects gene expression, all of the tumor tissue samples used in this study were handled in an identical fashion and experienced a similar length of delay. Thus, the tumor-harvesting procedure would have affected all samples in a similar manner and would not be expected to have contributed to the difference in gene expression patterns seen among the samples. H&E-stained frozen tissue sections are routinely prepared from all tissue bank specimens for screening purposes. All tissue specimens for cDNA array analysis were screened by a neuropathologist (G. N. F.), and the diagnoses were independently confirmed by a second neuropathologist. The glioma tissue blocks were specifically selected for densest and purest tumor, and they were all comparatively and uniformly "pure." There was minimal contamination by normal brain parenchyma and minimal variation between samples in this regard. The tumors were diagnosed according to two commonly used criteria: (a) St. Anne-Mayo (11); and (b) the recently revised WHO Classification of Tumors of the Nervous System (WHO 2000; Ref. 12). In this study, the gliomas are termed according to the St. Anne-Mayo nomenclature as low-grade OL, AO, AA, and GM.

Isolation of Total RNA and mRNA from Tissues. The tissues were ground to powder under frozen conditions, and tissue powder (0.3–1.5 g) was lysed in the lysis buffer TRI Reagent (Molecular Research Center, Cincinnati, OH). The RNA isolation was done as described previously (13).

Hybridization to the Human Atlas cDNA Expression Array Blots. The cDNA microarray containing fragments representing 597 human genes with known functions and known tight transcriptional controls (Clontech Laboratories, Inc., Palo Alto, CA) was used for our experiments, as described previously (13). After a high-stringency wash, the hybridization pattern was analyzed by autoradiography and quantified by phosphorimaging.

Development of an Algorithm for Finding Strong Feature (Gene) Sets. We desire classifiers that categorize sample tissues based on gene expression values. There are two reasons why we desire classifiers involving small numbers of genes: (a) the limited number of samples often available in clinical studies makes classifier design and error estimation problematic for large feature sets (14); and (b) small gene sets facilitate design of practical immunohistochemical diagnostic panels. Thus, we use a simple classifier and a small number of genes (at most three in this study) to form classifiers (10).

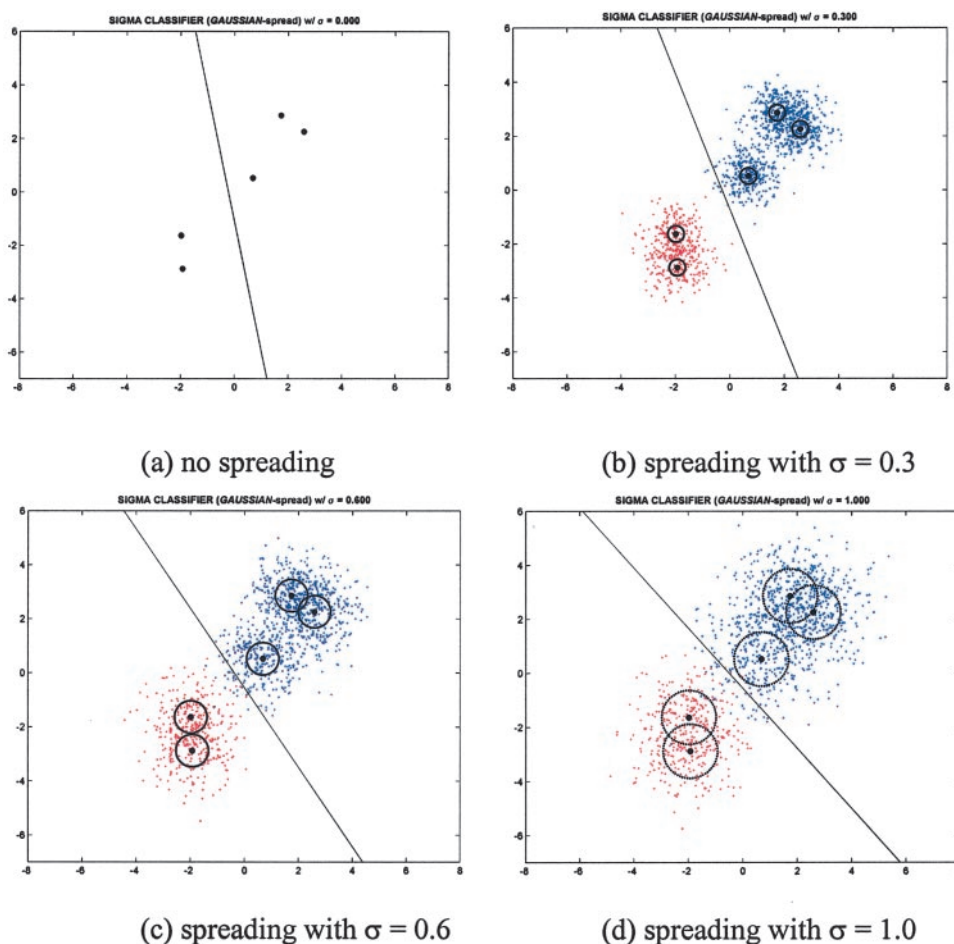
Given a set of features on which to base a classifier, two issues must be addressed: (a) design of a classifier from sample data; and (b) estimation of its error. When selecting features from a large class of potential features, the key issue is whether a particular feature set provides good classification. A key concern is the precision with which the error of the designed classifier estimates the error of the optimal classifier. When data are limited, an error estimator may be unbiased but may have a large variance and therefore may often be low. This can produce many feature sets and classifiers with low error estimates. The algorithm we use mitigates this problem by designing classifiers from a probability distribution resulting from spreading the mass of the sample points. The algorithm is parameterized by the variance of the distribution. The error gives a measure of the strength of the feature set as a function of the variance.

When the data are limited, and all of it is used to design the classifier, there are several ways to estimate the classifier error. We comment on two of these. The resubstitution estimate, ϵ_n , for a sample of size n is the fraction of errors made by the designed classifier on the sample. Typically, it is low-biased, meaning $E[\epsilon_n] \leq E[\epsilon_n]$, the expected value of the actual error. For LOO estimation, n classifiers are designed from sample subsets formed by leaving out one data point at a time. Each is applied to the left-out point, and the estimator $\hat{\epsilon}_n$ is $1/n$ times the number of errors made by the n classifiers. It is an unbiased estimator of ϵ_{n-1} , meaning that $E[\hat{\epsilon}_n] = E[\epsilon_{n-1}]$. This unbiasedness comes at a cost: the variance of the LOO estimator is greater than that of resubstitution (15).

For $\sigma \geq 0$, the algorithm we use constructs from the sample data a linear classifier $\Psi\sigma$, where σ^2 gives the variance of the distribution used to spread the data. Both $\Psi\sigma$ and its error, ϵ_σ , are computed analytically. For $\sigma = 0$, which means there is no spreading of the sample mass, ϵ_σ is equal to the resubstitution error estimate for the sample. Thus, the standard theory informs us that the variance of ϵ_0 is less than that of the LOO estimator. Moreover, model-based studies indicate that the variance of ϵ_σ decreases as σ increases. To standardize the interpretation of the results, σ is normalized relative to the variance of the data. Under this normalization, simulation studies with Gaussian distributions show ϵ_σ to be an unbiased estimator of the optimal linear classifier for $\sigma = 0.4$ and to be increasingly high-biased for increasing σ . To obtain conservative estimates of the optimal error, we take $\sigma \geq 0.4$. Moreover, for very small feature sets, we normalize by the maximum variance of the features. By being conservative, we reduce the chance that the resulting error estimate is optimistic. When considering a large number of potential feature sets in the presence of a small amount of data, the salient issue is one of data mining. Taking a conservative approach reduces the number of optimistic error estimates while at the same time selecting feature sets that perform well on a distribution that is significantly more dispersed than the actual data.

The concept of forming spread distributions from the data can be appreciated by reference to Fig. 1, which shows sample points from two classes (*red* and *blue*) based on measurements of genes $g1$ (*horizontal axis*) and $g2$ (*vertical axis*). Fig. 1a shows a linear classifier derived solely from the

Fig. 1. The concept of forming increasingly disperse distributions from the data can be appreciated in this figure, which shows sample points from two classes (*red* and *blue*) based on measurements of genes *g1* (*horizontal axis*) and *g2* (*vertical axis*). *a*, shown is a linear classifier derived solely from the sample points. *b–d*, shown are synthetic samples constructed from the original sample points by randomly adding noise of increasing variance to the original points to form larger samples that are spread about the original sample. *Dotted circles* are shown to represent a SD of spreading. A linear classifier has been derived for each synthetic sample. Increasing the variance increases the error. This method is called a Monte-Carlo simulation, but this simulation method is not used in the new method proposed. A new analytical method is developed to speed up the algorithm.



sample points. Fig. 1, *b–d*, shows samples constructed from the original sample points by deliberately adding artificial random noise of increasing variance to the original points to form larger samples that are spread about the original sample. A linear classifier has been derived for each synthetic sample. Increasing the variance increases the error. A classifier that has a small error for a large variance is desirable because its performance is more likely to be robust relative to new data. Because the implementation of this approach takes a long time if the Monte-Carlo method is used, the actual algorithm used does not use random synthetic data to find the classifier and its error but instead constructs class distributions from the sample data and then finds both the classifier and its error analytically via simple matrix operations (10).

Results and Discussions

Biologists are often interested in finding individual genes that have some influence on the system under study. In the context of classification, this approach translates into finding single-gene classifiers. Indeed, in the case of glioma classification, there appear to be cases in which single genes can provide decent classification, but certainly not always—for instance, when several genetic variations interact to result in

a phenotype. If we are interested in sets of genes that perform in a multivariate manner to provide strong classifiers, then we should look for pairs of genes that perform well and substantially better than either of the genes individually, triples of genes that perform well and substantially better than the best performing pair among the three, and so on. For any feature set, we let ϵ_σ denote the error of the optimal classifier for the feature set, and we let $\Delta(\epsilon_\sigma)$ denote the largest decrease in error for the full feature set relative to all of its subsets. The feature sets are first ranked based on the σ -error, and they are ranked again based on the improvement, $\Delta(\epsilon_\sigma)$. For multiple-gene classifiers, we will focus on feature sets with high rank in both lists. Indeed, this is our major focus: to find strong feature sets in which all genes contribute to glioma discrimination.

To aid in understanding the gene expression characteristics of the selected feature sets, all of the genes in the data set are clustered in such a way as to be close to other genes with similar expression. This is accomplished via hierarchical clustering using the Pearson correlation and average linkage. An added value to the clustering is that genes with known behavior can be used to analyze the results, and genes with unknown behavior can be placed into certain pathways for future functional testing.

Table 1 Feature sets to discriminate OL from others

Only pairwise classifiers that ranked at higher than 100th in both lists are included. Triplet-wise classifiers are included only when they are ranked at higher than 50th in both lists. For any feature set, ϵ_σ denotes the error of the optimal classifier for the feature set, and $\Delta(\epsilon_\sigma)$ denotes the largest decrease in error for the full feature set relative to all of its subsets. LOO is computed by designing n classifiers from sample subsets formed by leaving out one data point at a time, and then each classifier is applied to the left-out point, and the estimator LOO is $1/n$ times the number of errors by the n classifiers.

Gene names			ϵ_σ	$\Delta(\epsilon_\sigma)$	LOO
<i>Transducin $\beta 2$ subunit 2</i>			0.0207		0.00
<i>Transducin $\beta 1$</i>			0.0890		0.08
<i>Growth factor receptor-bound protein 2 (GRB2)</i>			0.1115		0.12
<i>Cyclin D3</i>	<i>SMARCA4</i>		0.0712	0.0986	0.04
<i>Follitropin receptor</i>	<i>Thymosin $\beta 10$</i>		0.0760	0.0921	0.04
<i>MRP</i>	<i>HSP70.1</i>		0.0761	0.0792	0.04
<i>MUC18</i>	<i>Clusterin</i>		0.0804	0.0667	0.04
<i>MUC18</i>	<i>Transducin $\beta 1$</i>	<i>GRB2</i>	0.0156	0.0213	0.00
<i>MUC18</i>	<i>Transducin $\beta 1$</i>	<i>RXR-β</i>	0.0310	0.0270	0.04
<i>MUC18</i>	<i>Transducin $\beta 1$</i>	<i>Clusterin</i>	0.0319	0.0260	0.00
<i>MUC18</i>	<i>BCL-W</i>	<i>GRB2</i>	0.0364	0.0303	0.04
<i>MUC18</i>	<i>Transducin $\beta 1$</i>	$\alpha 1$ <i>catenin</i>	0.0364	0.0215	0.04
<i>MAP kinase 10</i>	<i>SMARCA4</i>	<i>Neuronal acetylcholine Receptor $\alpha 3$</i>	0.0388	0.0500	0.00
<i>MRP</i>	<i>GRB2</i>	<i>Erythropoietin receptor</i>	0.0388	0.0274	0.04
<i>MAP kinase 10</i>	<i>Follitropin receptor</i>	<i>Neuronal acetylcholine Receptor $\alpha 3$</i>	0.0397	0.0531	0.00
<i>MUC18</i>	<i>GRB2</i>	<i>Erythropoietin receptor</i>	0.0406	0.0256	0.04
<i>Clusterin</i>	<i>ISGF3 γ</i>	<i>Erythropoietin receptor</i>	0.0431	0.0462	0.00

Table 2 Feature sets to discriminate GM from others

Pairwise classifiers are selected when they are ranked at higher than 200th in both lists, and triplet-wise classifiers are selected only when ranked at higher than 50th.

Gene names			ϵ_σ	$\Delta(\epsilon_\sigma)$	LOO
<i>TIE-2</i>			0.1315		0.12
<i>IGFBP2</i>			0.1392		0.16
<i>VEGFR1</i>			0.2113		0.16
<i>TIE-2</i>	<i>TNFSF5</i>		0.0796	0.0519	0.08
<i>IGFBP2</i>	<i>EPHA1</i>		0.0861	0.0531	0.00
<i>IGFBP2</i>	<i>CDK7</i>		0.0879	0.0513	0.08
<i>IGFBP2</i>	<i>TNFSF5</i>		0.0911	0.0481	0.04
<i>myc</i>	<i>IGFBP2</i>	<i>CDK7</i>	0.0582	0.0297	0.00
<i>IGFBP2</i>	<i>TNFSF5</i>	<i>CC chemokine receptor type 2</i>	0.0591	0.0320	0.00
<i>TIE-2</i>	<i>CXC chemokine receptor type 4</i>	<i>EPHA1</i>	0.0623	0.0322	0.04
<i>TIE-2</i>	<i>CDK7</i>	<i>JAK3</i>	0.0634	0.0302	0.04
<i>TIE-2</i>	<i>IGFBP2</i>	<i>JAK3</i>	0.0660	0.0315	0.08

Classification Analysis for Glioma Data. We applied the algorithm (10), which was described briefly in “Materials and Methods,” to a set of gene expression profile data derived from 25 human glioma surgical tissue samples. The cDNA microarray experiments were carried out to gain expression information for 597 known cellular genes.

We designed two-class classifiers for the classification of OL from others, AO from others, AA from others, and GM from others. We limited the number of genes for each classifier to only three, and the dispersion levels (amount of spread) of samples were varied from $\sigma = 0.4$ to $\sigma = 0.8$. We focus on $\sigma = 0.6$ because it provides conservative error estimation, but not too conservative (10). Even with analytic classifier design and error estimation, due to the number of potential feature sets and the various cases

considered, the computations were done on a Beowulf-based supercomputer (16) at the Center for Information Technology at NIH.

Tables 1–4 show the feature sets identified for each classification category. The tables are constructed so that feature sets ranked high in both σ -error, ϵ_σ , and improvement, $\Delta(\epsilon_\sigma)$, of σ -error are listed. This is accomplished according to the following scheme: (a) the top three single-gene classifiers for the category are listed in each table; (b) two-gene classifiers ranked in the top N_2 pairs for both σ -error and improvement of σ -error are included (N_2 table-dependent); and (c) three-gene classifiers included in the top N_3 triples for both ϵ_σ and $\Delta(\epsilon_\sigma)$ are included (N_3 table dependent). For comparison purposes, the LOO error estimate is also shown in the tables. As expected, overall the σ -error is more con-

Table 3 Feature sets to discriminate AO from others

Pairwise classifiers are selected when they are ranked at higher than 10th in both lists, and triplet-wise classifiers are selected only when ranked at higher than 50th.

Gene names			ε_{σ}	$\Delta(\varepsilon_{\sigma})$	LOO
DNase			0.1556		0.28
TNFSF5			0.1658		0.20
RAD50			0.1659		0.20
TNFSF5	DNase X		0.0750	0.0806	0.04
Prostaglandin E2 receptor EP4	DNase X		0.0784	0.0772	0.08
GNA13	TNFSF5		0.0826	0.0832	0.08
Prostaglandin E2 receptor EP4	TNFSF5		0.0907	0.0947	0.08
RAB5A	TNFSF5		0.0909	0.0749	0.16
erbB4	Prostaglandin E2 receptor EP4		0.1012	0.0841	0.08
PKA C- α	TNFSF5	Preprotachykinin β	0.0529	0.0549	0.04
DNase X	R κ B DNA-binding protein	Preprotachykinin β	0.0534	0.0479	0.04
PKA C- α	DNA ligase IV	TNFSF5	0.0591	0.0488	0.04
DNA ligase IV	TNFSF5	Acidic fibroblast growth factor	0.0616	0.0474	0.04

Table 4 Feature sets to discriminate AA from others

Pairwise classifiers are selected when they are ranked at higher than 100th in both lists, and triplet-wise classifiers are selected only when ranked at higher than 50th.

Gene names			ε_{σ}	$\Delta(\varepsilon_{\sigma})$	LOO
CREB1			0.1018		0.20
IFN- α receptor 2			0.1208		0.12
DCC			0.1210		0.16
CREB1	RAB3A		0.0695	0.0323	0.16
CREB1	IL2R- γ		0.0745	0.0273	0.12
Cyclin E	CREB1		0.0759	0.0258	0.12
CREB1	MAP kinase 10		0.0761	0.0257	0.12
CREB1	BCL2A1		0.0761	0.0256	0.08
U-PAR	VEGFR2		0.0855	0.0704	0.12
U-PAR	FADK		0.0917	0.0641	0.17
U-PAR	HTF4		0.0935	0.0597	0.16
CREB1	BCL2A1	CD11B	0.0511	0.0250	0.04
CREB1	VEGFR2	Thymosin β 10	0.0565	0.0231	0.04
CREB1	VEGFR2	CD11B antigen	0.0584	0.0212	0.08
Cyclin E	CREB1	Thymosin β 10	0.0587	0.0173	0.08
CREB1	BCL2A1	Endothelin receptor type A	0.0588	0.0174	0.04
U-PAR	CREB1	VEGFR2	0.0592	0.0204	0.08
p55-FGR	Cyclin E	CREB1	0.0592	0.0167	0.08
CREB1	BCL2A1	IL12- α	0.0599	0.0162	0.04
CREB1	VEGFR2	Caspase 2	0.0607	0.0189	0.08
CREB1	VEGFR2	SCYB5	0.0623	0.0173	0.08

servative, so that when the σ -error is very small, usually the LOO error is also very small or zero.

To illustrate interpretation of the tables, consider discrimination of OL in Table 1. When selecting multivariate classifiers, we have removed all classifiers that include transducin β 2 subunit 2 from the list because this gene itself has discriminating power so great that no matter what gene (even a noninformative gene) is used with it, the pairwise σ -error is very low (at least as low as for the gene itself). Because of our desire to avoid this kind of redundancy in the tables, there are gene sets omitted from the two- or three-gene lists that possess smaller σ -errors than those shown in the table. For instance, in Table 1, the σ -error for the top-listed two-gene set is substantially

greater than for any pair involving transducin β 2 subunit 2, simply because adjoining genes to transducin β 2 subunit 2 produce a σ -error less than that of transducin β 2 subunit 2 itself. The complete performance lists for both error and improvement in error can be found in the supplementary information.⁴

The advantage of reporting the results in the way we have is that multivariate discriminatory power is revealed. This is clearly demonstrated in Table 1 with regard to cell surface glycoprotein MUC18. The gene does not appear on the

⁴ Supplementary data is available at *Molecular Cancer Therapeutics* Online (<http://mct.aacrjournals.org>).

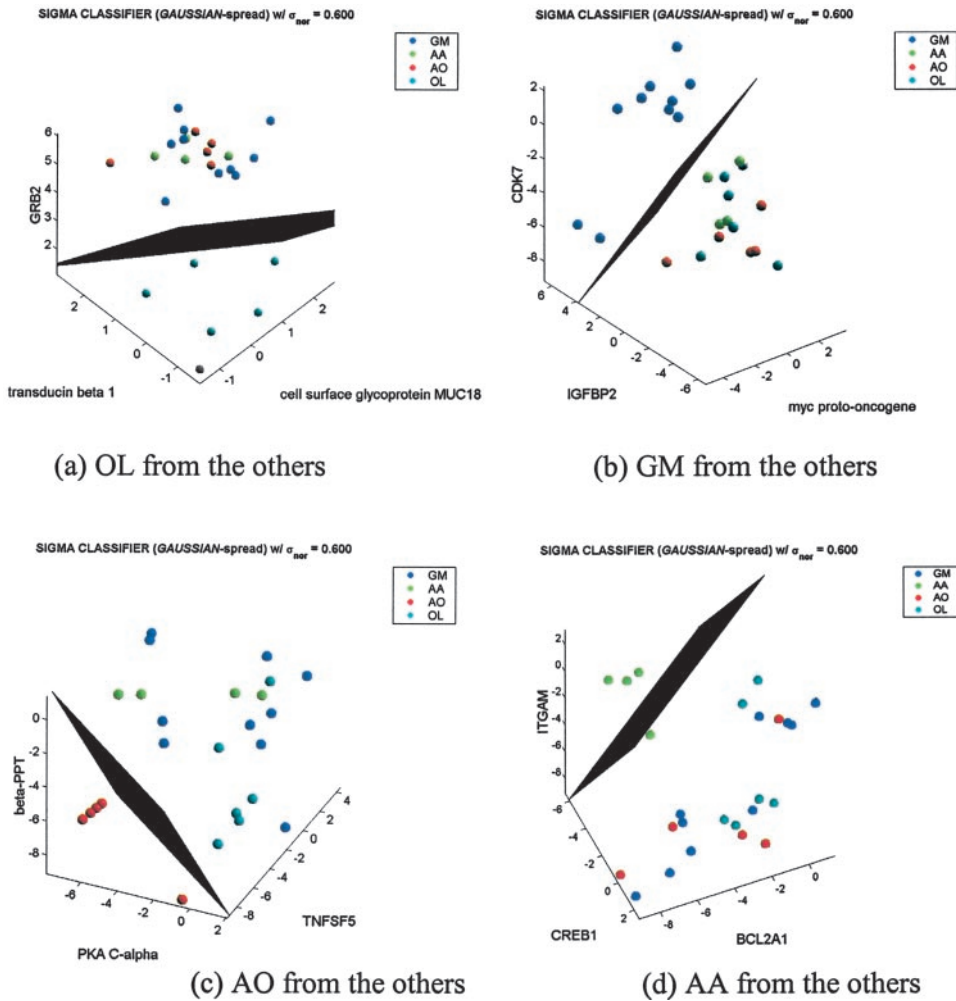


Fig. 2. Multivariate discriminators for glioma classifications (examples): a, a strong multivariate (three-gene) discriminator of OL from other types of glioma. OL shows relatively low expression in all three genes. b, discriminator of GM from other types. GM shows relatively high expression of all three genes shown. c, discriminator of AO from others. d, discriminator of AA from others. Note that there is a clear separation between AO and others even though the hyperplane doesn't discriminate them perfectly. This is because of the nature of the algorithm. The algorithm tries to find the best discriminator that is efficient not only on the data set given but also on prediction. This is confirmed when LOO error is computed. The LOO error for the feature set and the data are not 0 but 0.04 (1 of 25). This is important because it again shows that the designed classifier does not over-fit the data. Scale on each axis represents a log₂-transformed normalized intensity, log₂ (intensity/median intensity of an array).

single-gene list, indicating that its σ -error exceeds 0.1115; however, it appears with clusterin (CLU) in the two-gene list and both with and without clusterin (CLU) in the three-gene list. The substantial improvement in each case demonstrates the significant contributions of the genes within each gene set.

There are other instances where the improvement of classification error is sufficient to warrant inclusion in a table. In Table 2, even though IGFBP2 is by itself a decent discriminator, when it is combined with others, such as ephrin type A receptor 1 (EPHA1), the error is significantly improved. The σ -error decreases by more than 0.05, from 0.1392 (data not shown) to 0.0862. The improvement for the LOO error is more significant, from 0.16 (4 of 25) to 0 (0 of 25). Because of this, feature sets including IGFBP2 are shown in the table. We recently studied IGFBP2 expression in 256 cases of gliomas of different grades using tissue microarray and found that IGFBP2 is overexpressed in 80% of GBMs (Ref. 17; data not shown). Further testing with suitable antibodies will be able to test whether combination of IGFBP2 and EPHA1 will provide more accurate classifications. Some of these multivariate discriminators are shown in Fig. 2.

Clustering. In Fig. 3, the global clustering map is shown on the *left* for the hierarchical clustering analysis outlined in “Materials and Methods.” Four clusters are interesting with regard to discriminating OL and GM. Most genes found to singly discriminate OL from other types of glioma appear in the first cluster extracted on the *right*, and they are under-expressed in OL. Most genes found to classify GM from other types lie in the other three extracted clusters. In the first cluster, most of the genes are overexpressed in GM and underexpressed in AO and OL; in the second cluster, they are overexpressed in GM and underexpressed in OL; and in the third cluster, they are slightly underexpressed in GM.

Most genes identified as singly but only marginally classifying AO from the others are not clustered together as well as in the OL and the GM cases, nor are those classifying AA from the others. We find this interesting because this is consistent with the fact AO and AA represent more heterogeneous characteristics of the cancer. This supports the usefulness of the multivariate approach. Had we tried only a univariate approach to identify a singleton discriminator, we would not have found feature sets that can discriminate these two classes from others.

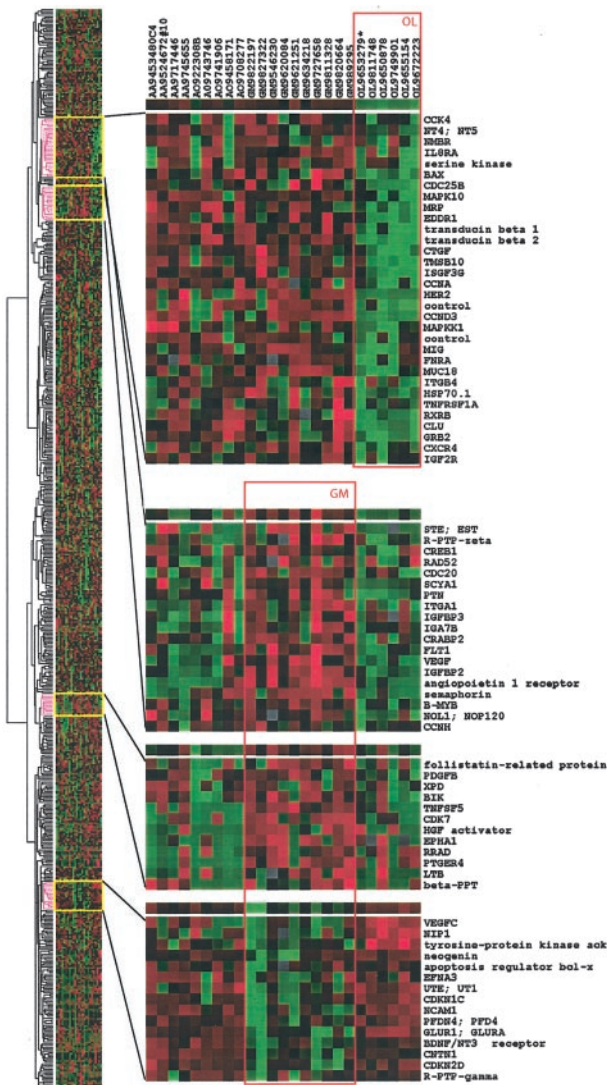


Fig. 3. Hierarchical clustering of gene expression profile on glioma data set. *Top panel*, a cluster of genes that discriminate OL from other types of glioma. Most of the genes are underexpressed in OL. *Second panel*, a cluster of genes that classify GM from other types where most of the genes are overexpressed in GM and underexpressed in AO and OL. *Third panel*, a cluster of genes that classify GM from other types where most of the genes are overexpressed in GM and underexpressed in OL. *Bottom panel*, a cluster of genes that classify GM from other types where most of the genes are slightly underexpressed in GM.

Gliomas are very complex cancers involving different growth characteristics and cell lineage features (12). Because the original clone of tumor cells may exist at any stage of cell differentiation and may have different transformation events, the boundaries between tumor grades and tumor lineages can be blurred. This is reflected in the current morphologically based tumor classification schemes that often mix cell lineage features with tumor growth characteristics. The results are frequently subjective, and disagreements among pathologists regarding the identity of the tumor are not uncommon. The gene expression activities yielded by the study of molecular biology and genomic biology may provide

a more objective method to classify diseases. This belief is based on the assumption that cell phenotypes have genotypic origins. Recent successes in subclassification of neoplasms within a disease group using gene expression profiles (3–7) provide support for such a belief.

Thus, the issue is how to best identify the strong feature genes that are closely linked to specific phenotypes from among the thousands of genes in gene expression profiles and how to determine whether this information really aids classification of tumors. There are many technical challenges in the path to accomplishing the task of finding the key links.

The first major roadblock is the small sample size issue inherent to microarray-based classification efforts (14). Contributing to this are the limited numbers of human tissues for study and the cost of such gene expression profiling projects. Because classifiers are designed from observed expression vectors that have randomness arising from biological and experimental variability, the design, performance evaluation, and application of classifiers must take this randomness into account, especially when the number of samples (tissue specimens) is small, which is the case in most human tissue-based microarray experiments.

Algorithms are therefore needed to identify robust classifiers from very limited data sets. Three criteria have to be met for an algorithm to be considered strong. First, given a set of variables, a classifier from the sample data should provide good classification over the general population. Second, the algorithm should be able to estimate the error of a designed classifier when data are limited. Third, given a large set of potential variables, the algorithm should be able to select a set of variables as inputs to the classifier.

Taking these issues into consideration, we used a recently developed method to find both strong classifiers and strong features (10). This algorithm considers the inherently variable or “high-noise” nature of microarray measurements. Using this algorithm, we have identified robust classifier gene sets containing one to three genes that distinguish each type of glioma from the other three. This provides guidance for the development of pathological assays using a reasonable number of markers for clinical use.

In a broader context, the approach applied in this study can be used to identify genes that contribute to the major differences between any two groups of samples analyzed, in the process of which some less understood phenotypes might be identified. For example, we might find strong feature gene sets that distinguish cancers with high metastatic potential from cancers with little or no metastatic potential or gene sets that identify cancers that will be sensitive to specific therapies *versus* those that will be resistant and continue to grow unabated through therapy. Current histology-based classification and grading systems can do neither of these. Identification of such strong feature genes may not only provide markers for diagnosis and disease management but may also provide novel potential targets for drug development. Cancers have complex features, but we cannot target all of these features for treatment. A method that could identify the strong features, both genotypically and phenotypically, would provide an ideal route to the heart of the

problem. Future studies will tell whether the currently used algorithm or an improved one will achieve this goal.

Acknowledgments

We thank Drs. Edward B. Suh and Robert L. Martino for providing the computational resource of the Beowulf clustered supercomputer at the Center for Information Technology of NIH for the heavy computation of the algorithm. We thank Beth Notzon for editorial assistance.

References

- Hogenesh, J. B., Ching, K. A., Batalov, S., Su, A. I., Walker, J. R., Zhou, Y., Kay, S. A., Schultz, P. G., and Cooke, M. P. A comparison of the Celera and Ensembl predicted gene sets reveals little overlap in novel genes. *Cell*, 106: 413–415, 2001.
- Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., *et al.* The sequence of the human genome. *Science (Wash. DC)*, 291: 1304–1351, 2001.
- 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.
- 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. D., 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.
- 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., and Sondak, V. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature (Lond.)*, 406: 536–540, 2000.
- Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Kallioniemi, O. P., Wilfond, B., Borg, A., and Trent, J. Gene expression profiles in hereditary breast cancer. *N. Engl. J. Med.*, 244: 539–548, 2001.
- Perou, C. M., Sorlie, T., Eisen, M. B., van de 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.
- Ben-Dor, A., Bruhn, L., Friedman, N., Nachman, I., Schummer, M., and Yakhini, Z. Tissue classification with gene expression profiles. *J. Comput. Biol.*, 7: 559–583, 2000.
- 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.
- Kim, S., Dougherty, E. R., Junior, B., Chen, Y., Bittner, M. L., and Trent, J. M. Strong feature sets from small samples. *J. Comput. Biol.*, 9: 129–148, 2002.
- Daumas-Duport, C., Scheithauer, B. W., and O'Fallon, J. Grading of astrocytomas. A simple and reproducible method. *Cancer (Phila.)*, 62: 2152–2165, 1988.
- Kleihues, P., and Cavenee, W. K. (eds.). *Pathology and Genetics of Tumours of the Nervous System*, 2nd ed. (WHO Classification of Tumours of the Nervous System). New York: Oxford University Press, 2000.
- Fuller, G. N., Rhee, C. H., Hess, K. R., Caskey, L. S., Wang, R., Bruner, J. M., Yung, W. K., and Zhang, W. Reactivation of insulin-like growth factor-binding protein 2 expression during glioblastoma transformation revealed by parallel gene expression profiling. *Cancer Res.*, 59: 4228–4232, 1999.
- Dougherty, E. R. Small sample issues for microarray-based classification. *Comparative and Functional Genomics*, 2: 28–34, 2001.
- Devroye, L., Györfi, L. and Lugosi, G. *A Probabilistic Theory of Pattern Recognition*. New York: Springer, 1996.
- Sterling, T. L., Salmon, J., Becker, D. J., and Savarese, D. F. *How to Build a Beowulf: A Guide to the Implementation and Application of PC Clusters*. Cambridge, MA: The MIT Press, 1999.
- Wang, H. M., Wang, H., Zhang, W., and Fuller, G. N. Tissue microarrays: applications in neuropathology research, diagnosis, and education. *Brain Pathol.*, 12: 95–107, 2002.