RESEARCH REPORT

Statistical Modeling and Visualization of Molecular Profiles in Cancer

Robert Scharpf, Elizabeth S. Garrett, Jiang Hu, and Giovanni Parmigiani
Department of Biostatistics, Sidney Kimmel Comprehensive Cancer Center, and Department of Mathematical Sciences Johns Hopkins University, Baltimore, MD, USA

BioTechniques 34:S22-S29 (March 2003)

ABSTRACT

Current cancer classifications using morphological criteria produce heterogeneous classes with variable prognosis and clinical course. By measuring gene expression for thousands of genes in a single hybridization experiment, microarrays have the potential to contribute to more effective classifications based on molecular information. This gives hope to improve both prognosis and treatment.

Statistical methods for molecular classification have focused on using high dimensional representations of molecular profiles to identify subclasses. These can be noisy, unstable, and highly platform-specific. In this article, we emphasize the notion of molecular profiles based on latent categories signifying under-, over-, and baseline expression. Following this approach, we can generate results that are more easily interpretable, more easily translated into clinical tools, more robust to noise, and less platform-dependent. We illustrate both the methods and the associated software for molecular class discovery on a data set of 244 microarrays comprising six known leukemia classes.

INTRODUCTION

Although cancer classification using morphological criteria has lead to important progress in prognosis and treatment, substantial heterogeneity still exists in current cancer classes because of the difficulty of adequately gauging cancer complexity. Fundamentally, the phenotype of a cancer cell is the sum of its proteins. Variability in cancer cell phenotype can arise from errors in the amino acid sequence of constituent proteins or, more commonly in cancer, differences in the proteins synthesized and their amounts (1). While carcinogenesis is typically the result of accumulated mutations in one or several genes, only those mutations penetrant at the protein level alter the phenotype of a cell. As opposed to a single cell, the phenotype of a tumor is the collective physical manifestation of cancer cells that define clinical course, including response to therapeutics and metastatic potential.

Because of the complex molecular basis of cancer, it is currently hypothesized that heterogeneity of tumor phenotype within tumor classes can be further distinguished by the underlying variability at the molecular level. Molecularly oriented methodologies, which measure abundance of one or several target molecules, have already had an impact on cancer classification. One example is the classification of breast cancers, using the estrogen receptor and the c-erbB-2 markers (2). However, the identification of one or two genes that are differentially expressed may omit genes that better predict cancer phenotype or may be insufficient for differentiating between all the important tumor phenotypes (3).

Screening procedures to identify genes that are differentially expressed across samples or tumors can be developed using high-throughput genomic technologies, such as serial analysis of gene expression (SAGE) (4) or gene expression microarrays (5). Microarrays measure gene expression, a mechanism by which cells can regulate protein synthesis of thousands of genes and expressed sequence tags (ESTs). Validation by more accurate methodologies (i.e., quantitative reverse transcription polymerase chain reaction [QT-RT-PCR]) is then limited to one or several markers identified by microarrays.

Classifications of clinically heterogeneous cancers (e.g., diffuse large B cell lymphomas [DLBCL]) using microarrays is an emerging application of microarray technology. Attempts to derive subclassifications by cell morphology have been unsuccessful (6,7). Patients diagnosed with DLBCL receive the same clinical intervention and prognosis, yet less than half of these patients achieve sustained remission following chemotherapy. Recently, microarray-based subclassifications of DLBCL that segregate chemotherapeutic-responsive tumors from nonresponsive tumors were identified by separate investigators (1,8,9). Classifications using microarrays were carried out in lung adenocarcinoma (10–12), breast cancer (13,14), melanoma (15), colon cancer (16), prostate cancer (17,18), and ovarian cancer (19). More accurate classifications based on molecular phenotype can influence prognostication, individualized treatment, the development of targeted therapeutics, and the identification of biomarkers for earlier diagnosis. The assumption that cancers with similar profiles of gene and EST expression have similar cancer phenotype has been corroborated by numerous investigations.

Despite the accomplishments of microarray-based cancer classifications, the statistical issues for identifying differentially expressed genes in microarray studies remain formidable. Current approaches are reviewed in References 20 and 21. The

prevalent approach for identifying molecular profiles is a combination of clustering and visualization, popularized by Eisen's work (22). This is based on displaying an expression map in which genes and samples are sorted based on hierarchical clustering techniques. Subclasses are defined by proximity of samples in this map, which depends on the set of probes spotted, and the metrics chosen for clustering. These dependencies pose a challenge to biological interpretation, reproducibility of results, and clinical translation.

Also, microarray data exhibit a substantial component of noise, resulting from the numerous concomitant sources of variation affecting expression measurements. Clustering and visualization can be sensitive to this noise. Many visualization approaches use ad hoc nonlinear color scales that effectively reduce the visual impact of expression levels to over-, under-, and equally expressed (e.g., Reference 12). Noise, however, is still affecting the underlying clustering.

In this article, we review and illustrate a statistical approach (23), and the associated software called probability of expression (POE) (24), which emphasizes the notion of molecular profiles based on categories signifying under-, over-, and baseline expression. These molecular profiles are estimated using a statistical mixture modeling technique that reduces noise and allows for a profile definition that is both more comparable across platforms and amenable to biological interpretation, validation, and clinical implementation.

METHODS

Three-Way Categorization of Expression Levels

The underlying assumption of the approach described in this paper is that expression can be usefully described as falling into one of three categories: overexpression, underexpression, or normal expression. These categories are defined by comparing the expression of a gene across samples. Since these qualitative categories of expression are not exactly observable, they are latent. The introduction of latent categories has the following goals: (*i*) to make the analysis robust to outlying observations that are of minor biological significance, but that can strongly affect clustering; (*ii*) to remove the normal component of variation that can be expected in the expression of a gene as a result of both technical and biological variability; (*iii*) to build an expression scale that permits comparison across platforms at the level of the individual gene–sample combinationl and (*iv*) to allow for simple and interpretable definitions of profiles taking the form:

Gene 1 above normal; Gene 2 below normal; ... Gene N normal

Subdivision of gene expression levels in categories can be done at different levels of complexity. For example, simple quantization rules or three-means clustering of expression levels one gene at a time can prove useful. In this article, we take a more systematic inferential approach. In practice, this has the advantages of improving the accuracy and stability of estimated categories and of providing a probabilistic basis for inference about which samples belong to which cancer subtype. Based on prior knowledge of gene expression distributions in microarray experiments, we use a mixture model to capture the gene expression distributions typically observed in microarray experiments. The mixture model permits correction for noise at both the sample and the gene level. Specifically, we use a hierarchical mixture model that borrows strength across genes to aid in estimation of parameters.

The fit produced by the POE model gives probabilities for the three classes. These can be converted to a 1-dimensional scale giving values from -1 to 1. Values approaching -1 provide strong evidence of underexpression, values near 1 provide strong evidence of overexpression, and values near 0 suggest normal (or modal) expression. Visualization of gene expression with POE can proceed in a similar way to standard gene expression visualization, such as hues of red for underexpression, green for overexpression, and black for no evidence of differential expression. This can be thought of as a nonlinear mapping of expression to color. It has two main advantages over traditional approaches: (*i*) it allows for genes that are prevalently noise to be classified as normal in all samples (unlike approaches that normalize genes to have the same range or the same variance); and (*ii*) it allows for a separate signal-to-noise ratio for each gene. POE also provides information about the proportion of samples in which a gene is differentially expressed and summaries of gene similarities.

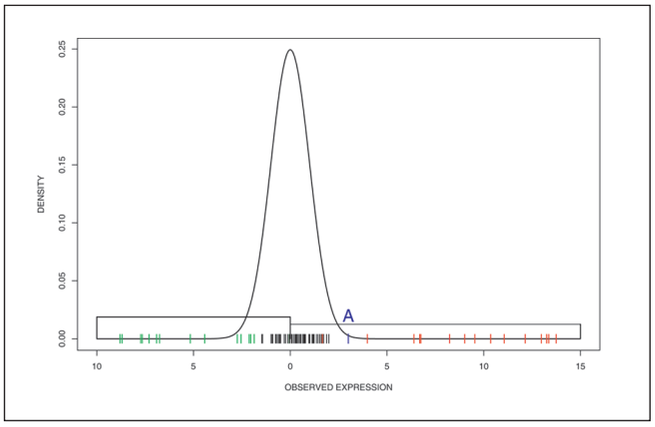

Figure 1. Illustration of the three-way mixture model for gene expression adoptd by POE. The center distribution represents baseline expression measures fit by a normal distribution, and the tails, representing underexpression (left) and overexpression (right), are fit by uniform distribution.

The Mixture Model

Estimation of the probabilities of different expression categories is performed using a mixture model (25,26). The term mixture refers to the fact that the probability distributions of gene expression across samples is made by mixing multiple components, as illustrated in Figure 1. The basic idea is that observations (i.e., gene expression levels) arise from one of the three groups (over-, normal, and underexpression), each having its own distribution. However, in a given set of gene expression data, the component that generated the observation is unknown. We say that the observations are mixed together or that the observations are a "mixture" of the three groups. The goal of the POE model is to untangle the groups in order to approximate the component of the model that best describes a given expression.

The biological motivation for a mixture model approach is based on the idea that for each gene there is usually a portion of

the samples with very similar expression values (24). The differences in these expression values for the modal category can be caused by noise in the measurement or by stochastic biological effect (27,28); while the second may be of interest, the two sources are difficult to distinguish in the absence of replicate measurements on the same tumor. Conversely, genes that are differentially expressed may represent genes that are constitutively up-regulated or absent. The range of plausible values for gene expression in aberrantly regulated genes are better approximated by a uniform probability distribution (24). One major advantage of the uniform distribution is that its mathematical properties make its estimation relatively simple, even in the case of a poor model fit. For example, the uniform distribution allows estimation of over- and underexpressed genes with fewer parameters than that required by the normal distributions (23).

In POE, a separate mixture model of the kind illustrated in Figure 1 is fit for each gene. Generally, there are relatively few measurements to estimate gene-specific parameters, especially the spread of the bell-shaped curve representing the modal class. However, technology limitations that affect gene expression in microarray experiments are likely to affect gene expression for all genes in a similar way. Rather than fitting the model to each gene independently, we "borrow strength" across genes by using a hierarchical model that assumes that gene-specific parameters across genes are related (29). The specific forms of the hierarchical distributions used in POE are described elsewhere (24).

Using the estimates from the mixture model, we derive the probability p^{+}_{ji} that gene j in sample i is overexpressed and the probability $p_{\bar{j}i}$ that it is underexpressed. For example, point A in Figure 1 corresponds to an observed expression of 3. For it, $p_{\bar{j}i}$ is 0. The overexpressed uniform component density is 0.0125, while the normal is 0.00443. The ratio of these two values, 2.82, represents the odds of belonging to the uniform component and, therefore, being overexpressed. Using the relation $P = O/(O+1)$, where O are odds and P is probability, the probability p^{+}_{ji} that this observation is overexpressed is 0.74. Conversely, the probability that it is normally expressed is 0.26. Note that in Figure 1, the areas under the three densities together add to one. The area under each density can be interpreted as the fraction of under-, normally, and overexpressed samples. Also, in our model, either $p^{+}_{ji} = 0$, or $p_{\bar{j}i} = 0$ for any i and j.

For each element in the original matrix X, we now have two probability estimates, p^{+}_{ji} and $p_{\bar{j}i}$. Subtracting the probability of underexpression from the probability of overexpression yields a number between -1 and 1, where values between -1 and 0 indicate expressions with positive probability of underexpression, and values between 0 and 1 indicate positive probability of overexpression. Using this quantity ($p^{*}_{ji}=p^{+}_{ji} - p_{\bar{j}i}$), we have a new matrix p^{*} to be used as an alternative to X in clustering and mining.

Gene Mining

The transformation of every element in the X matrix to a qualitative measure of expression augments biological interpretability, is independent of platform, and is more robust to noise. Visualization software can be used at this stage for analysis of all the genes and samples. However, classification algorithms implemented on these molecular profiles will be using genes that are differentially expressed across samples. Approaches that eliminate genes, which contribute essentially noise or are not likely to be differentially expressed across samples, are more likely to provide meaningful classifications. Consequently, the output from the mixture model, p^{*}, can be used to mine for subsets of interesting genes.

The general idea of gene mining is to identify a subset of k distinct genes that show strong evidence of differential expression across samples and to find other genes with similar patterns to these k genes. Similar genes are pooled into groups to generate k subsets of genes. By iteratively choosing a set of "seed" genes, a data set can be mined for interesting classifications. The goals of choosing gene groups are to reduce redundancy in the set of genes, to provide additional context to facilitate interpretations of subtypes (for instance, having genes with functional annotations grouped with ESTs), and to reduce the likelihood of classifications based on artifact (24).

Garrett and Parmigiani and coworkers (24) choose a priori a differential expression pattern of interest and sorted genes by the level of similarity to this pattern. A differential pattern would be, for example, (0.05, 0.20), meaning that 5% of the samples have the gene underexpressed, and 20% of the samples have it overexpressed. A symmetric matrix of gene agreement is calculated based on the similarity of expression across samples. The intersection of row m and column n is a similarity metric for the m^{th} and n^{th} gene. Coherence is defined as the diagonal of the above matrix and represents the probability that the gene expression pattern for a gene would have the same true expression profile in another set of samples with identical expressions. Genes with low coherence are highly noisy or are not fit well by the model. Selection of seed genes is, thus, based on two criteria: similarity to the predefined differential expression pattern and coherence. Genes that show substantial agreement with the seed gene and are above the coherence threshold can be grouped with the seed gene. To find the next seed, the genes in this group are removed from consideration, and the gene with the next highest similarity to expression pattern and coherence is chosen as the seed. While a gene cannot be used twice as a group seed, it can belong to multiple groups. This is a critical property in cancer classification (30), as genes can belong to multiple pathways. We encourage consideration of several types of patterns and varying proportions of under- and overexpression for identifying seeds.

After visualization of the k groups of genes, one gene is chosen from each group to be representative of the group. This may be the seed gene, but the gene would generally be selected based on biological plausibility and relevance. For example, a seed gene might be an expressed sequence tag (EST), but another gene in the group may be a gene known to be related to cancer. It would be sensible to choose the cancer gene vs. the EST in this case. Multiple genes per group can be chosen. The marginal profile of a sample is then based on the subset of k genes. Typically three or four subsets of genes is sufficient for classification, as three subsets yields 27 possible profiles ($3^3 = 27$) as follows: {-1,-1,-1},\{-1,-1,0},...,{1,1,1}, where {-1,-1,-1} means all three genes are underexpressed, {-1,-1,0} means that genes 1 and 2 in the pattern are underexpressed, and the third gene is normally expressed, etc. Using p^{*}, we can determine which of these 3^k profiles is most likely for each of the samples. Other classification methods can be used where classification or clusters can be determined using p^{*}.

APPLICATIONS AND SOFTWARE

The software developed for this method is in R (31), which is a free statistical programming language available at (http://cran.r-project.org). The R package with all of the functions for implementing the mixture model described in the methods is called POE. The full library for use on all computer operating systems can be downloaded at (http://astor.som.jhmi.edu/poe). The functions in POE and the visualization tools included as part of POE are described in detail elsewhere (24). As with all complex statistical models, using POE does require some expertise, especially for assessing whether the algorithm reached adequate convergence and whether any major model violations arise.

Next, we illustrate the three-way mixture methodology in the context of leukemia classification. There are six classes of leukemia based on histological and cell morphology criteria. We used gene expression data from microarray experiments to see how well profiles generated by POE can classify the leukemia samples. Leukemia cells were obtained from the bone marrow of 244 pediatric acute lymphoblastic leukemia (ALL) patients (32). Oligonucleotide arrays (Affymetrix, Santa Clara, CA, USA) containing 12 600 genes were used for assaying gene expression. A total of 244 oligonucleotide arrays were used in this analysis. A natural log transform of the expression indices was performed, and data were centered by subtracting the row median from each row. Each row was also divided by its trimmed standard deviation (i.e., the standard deviation of the middle 50% of the samples). The Shapiro-Wilk test for normality was used to identify 1000 genes that had profiles significantly deviating from a normal distribution. The data for these 1000 genes were used for analysis. Profiles closely following a normal distribution are more likely to be reflecting noise alone, which makes the Shapiro-Wilk test preferable to overall variability in preselecting likely classifier genes.

The mixture model was applied to the leukemia data set using the function poe.fit. poe.fit takes a matrix with rows corresponding to genes and columns corresponding to samples as arguments. The following command typed into R would fit the mixture using the leukemia matrix (Y) and store the output in the R object my.output.

```
> my.output ← poe.fit(Y)
```

poe.fit uses a Markov chain Monte Carlo (MCMC) algorithm for determining expression probabilities and other estimates of model parameters. MCMCs are stochastic algorithms; repeated use will not generally provide identical answers, and convergence of the estimated values should be checked on each analysis (33).

The output of poe.fit can then be used by any visualization tools for subsequent analysis, including the function poe.bigpicture in the POE library. Figure 2 displays the red and green images that can be generated by poe.bigpicture on the leukemia data set using the log of the raw data. In contrast, Figure 3 was generated from poe.bigpicture after the raw data had been trans-

Figure 2. Log expression values of leukemia data.

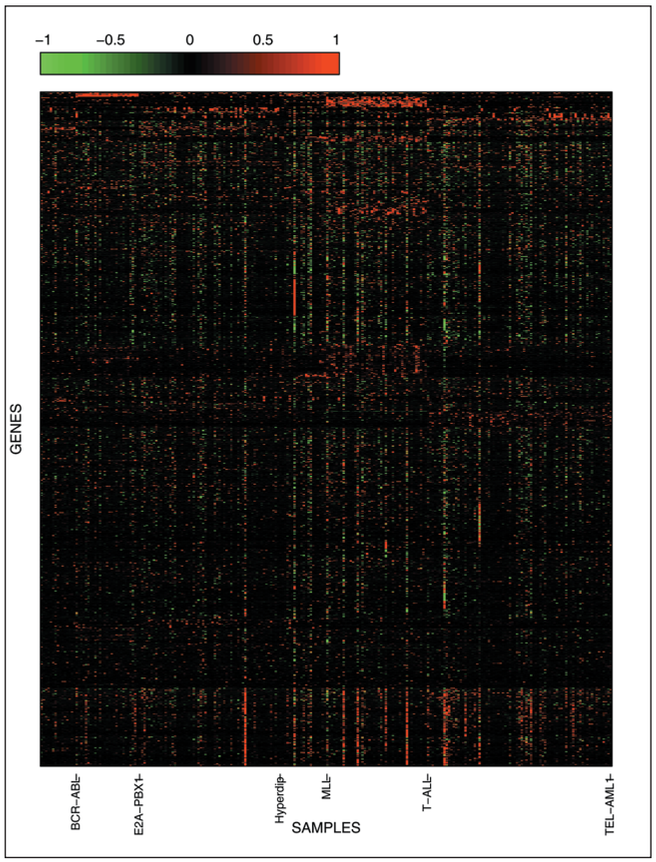

Figure 3. Phat values generated by POE on the leukemia data.

formed to the POE scale of probability scores (22). In both images, red corresponds to overexpression, and green corresponds to underexpression. Comparison of the two figures illustrates how the wide range of intensity values (even when using the log) can mask differential expression patterns, whereas on the POE scale, more red and green hues are evident.

We can examine how POE fits gene expression on a gene-by-gene basis using the poe.onegene command (Figure 4). By default, poe.onegene displays a histogram depicting the mixture model distribution and a quantile-quantile plot (qq-plot) of empirical vs. normal quantiles. This kind of visual diagnostics is critical for assessing the important assumption of normality of the center component. Deviation from normality can be addressed by transforming the expression scale or by developing more general mixture models.

Genes with a distribution, such as that of gene 825 in Figure 4, occur frequently in molecular classification studies. These genes are unlikely to provide any additional power to classification, as the likelihood that the Gaussian behavior reflects only noise is high. On the other hand, these genes will spuriously enter gene clusters produced under traditional clustering and visualization approaches that cluster all genes. POE is conservative in highlighting genes for classification only, when there is strong evidence of separation within a gene.

We searched for seed genes that are likely to be underexpressed in very few (5%) of the samples and overexpressed in 20% of the samples, by using the poe.mine function (Figure 5), with the goal of selecting genes that are good at identifying a

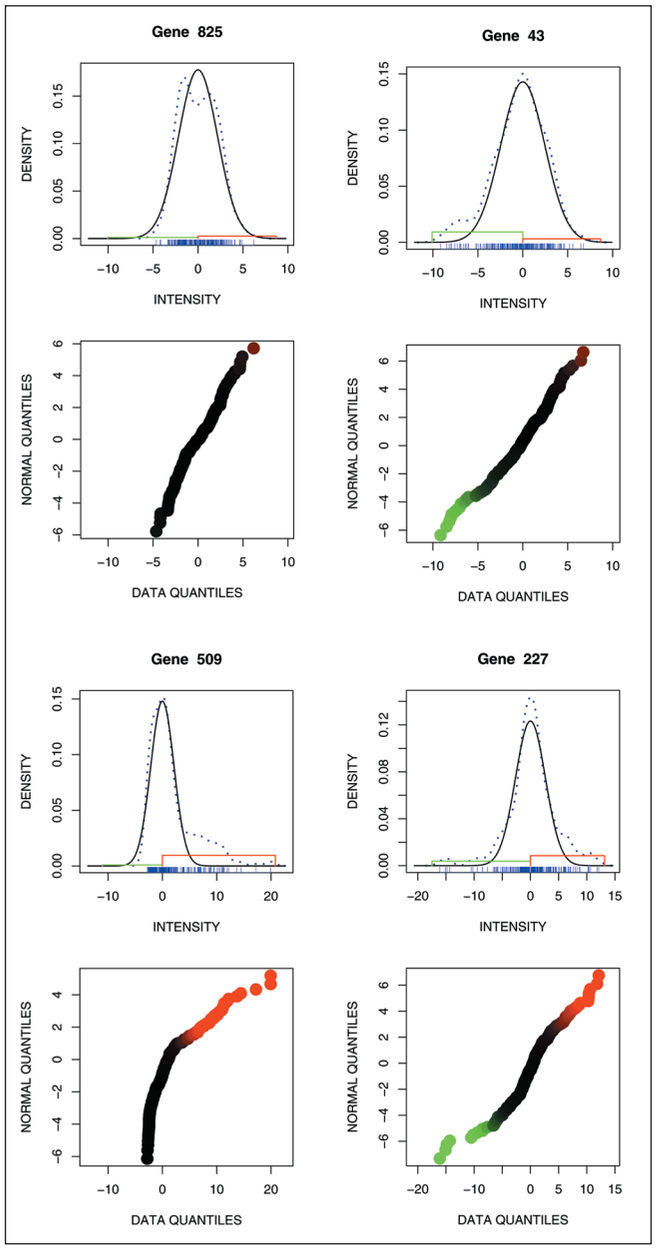

Figure 4. POE is fit to gene 825 (all normal), gene 43 (normal and underexpressed), gene 509 (normal and overexpressed), and gene 227 (all three classes present).

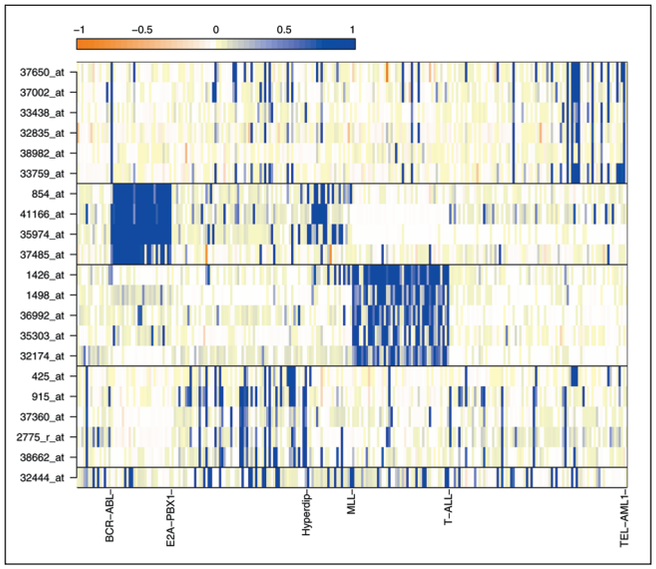

Figure 5. Groups of potential classifier genes extracted by mining for 5% underexpression, 20% overexpression, and 85% coherence.

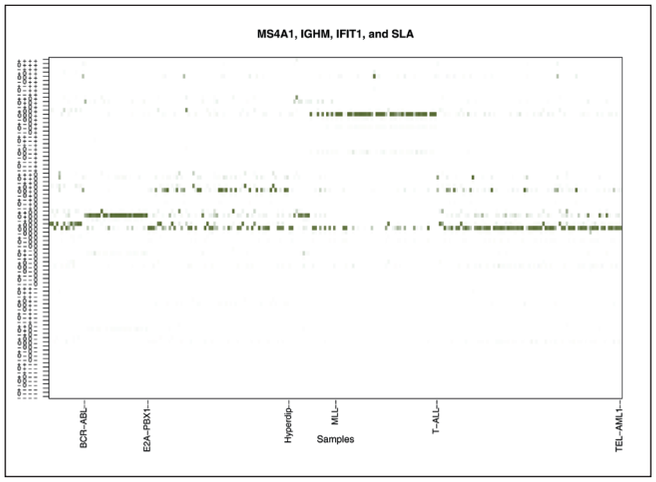

Figure 6. The probabilities of each of all possible molecular profiles generated with the four genes. Samples are sorted by known class.

subclass consisting of about one-fifth of the data. The criteria for gene mining can be relatively arbitrary or can be chosen based on prior hypotheses. Selection of different parameters is an easy method of data exploration.

Based on Figure 5, we selected four candidate classifier genes, one for each of the large groups. The genes are: (*i*) MS4A1, which codes for the CD20 receptor, and is a cell surface antigen on B cells; (*ii*) IGHM (heavy chain immunoglobulin), which is the major immunoglobulin expressed on the surface of B cells; (*iii*) IFIT1 (human interferon inducible protein); and (*iv*) SLA (human src-like adaptor protein), which is involved in signal transduction in T cells (34).

The four genes generate a total of 3^4=81 possible profiles. Figure 6 displays the probabilities with which each of the samples belongs to each of the 81 profiles. Samples are sorted by known leukemia classes. This provides an independent confirmation of the discriminatory ability of our unsupervised analysis. There are three profiles (0000, 0+00, and 000+) receiving substantial probability, and all three identify subgroups of samples that correspond to known subclasses. For example, the majority of the T-ALL samples have a high probability of belonging to the 000+ profile. Probabilities of profiles are also directly usable as a measure of the strength of a proposed classification. POE-based classifications, like any other, should be validated using independent data.

In the POE library, the command poe.pattern displays the different patterns of genes produced by poe.mine, as in Figure 5, while poe.profile produces the image plot of the molecular profile probabilities of Figure 6.

CONCLUSION

Molecular classification, especially in its unsupervised and partially supervised applications, is difficult. Successful investigations require complex analysis as well as substantive knowledge. Statistical methods cannot generally take an investigator from data to answers. Useful statistical methods are those that can be interfaced easily with biological information. The methodology discussed here attempts the interface between statistics and biology in two ways: (*i*) by defining differential expression in a way that is in tune with the three-way conceptualization underlying much of the biological work in this area; and (*ii*) by promoting a definition of molecular profile that is interpretable, portable across platforms, and independent of the algorithm used to identify a profile in the data.

The POE approach uses qualitative as opposed to quantitative measures of expression. While these may be more inter-

pretable and less sensitive to noise, they are also less powerful, as they may miss alterations in transcription, which are both quantitatively modulated and accurately measureable. The trade-off is, therefore, loss of detailed information vs. the ability to identify and interpret underlying structure. This suggests that POE should be used in conjunction, rather than as an alternative to, quantitative methods, so that this trade-off can be explored in the context of a specific problem.

Combined analyses across platforms are currently critical to the progress of our understanding of the human transcriptome. This is a hard problem. Cross-platform analyses are more likely to work at a qualitative level or using probability-based measures (35). The POE scale can provide the basis for a combined analyses that uses all the data, rather than gene-specific summaries.

REFERENCES

1. **Alizadeh, A.A., M.B. Eisen, R.E. Davis, C. Ma, I.S. Lossos, A. Rosenwald, J.C. Boldrick, H. Sabet, et al.** 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature *403*:503-511.
2. **Hayes, D.F. and A.D. Thor.** 2002. c-erbB-2 in breast cancer: development of a clinically useful marker. Semin. Oncol. *29*:231-245.
3. **Mohr, S., G.D. Leikauf, G. Keith, and B.H. Rihn.** 2002. Microarrays as cancer keys: an array of possibilities. J. Clin. Oncol. *20*:3165-3175.
4. **Velculescu, V., L. Zhang, B. Vogelstein, and K. Kinzler.** 1995. Serial analysis of gene expression. Science *270*:484-487.
5. **Schena, M., D. Shalon, R. Davis, and P. Brown.** 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science *270*:467-470.
6. **Harris, N.L., H. Stein, S.E. Coupland, M. Hummel, R.D. Favera, L. Pasqualucci, and W.C. Chan.** 2001. New approaches to lymphoma diagnosis. Hematology 194-218.
7. **The Non-Hodgkins's Lymphoma Prognostic Factors Project.** 1993. A predictive model for aggressive non-Hodgkin's lymphoma. N. Engl. J. Med. *329*:987-994.
8. **Huang, J.Z., W.G. Sanger, T.C. Greiner, L.M. Staudt, D.D. Weisenburger, D.L. Pickering, J.C. Lynch, J.O. Armitage, et al.** 2002. The t(14;18) defines a unique subset of diffuse large B-cell lymphoma with a germinal center B-cell gene expression profile. Blood *99*:2285-2290.
9. **Shipp, M., K. Ross, P. Tamayo, A.P. Weng, J.L. Kutok, R.C. Aguiar, M. Gaasenbeek, M. Angelo, et al.** 2002. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nat. Med. *8*:68-74.
10. **Bhattacharjee, A., W.G. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, C. Ladd, J. Beheshti, et al.** 2001. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc. Natl. Acad. Sci. USA *98*:13790-13795.
11. **Beer, D.G., S.L. Kardia, C.C. Huang, T.J. Giordano, A.M. Levin, D.E. Misek, L. Lin, G. Chen, et al.** 2002. Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nat. Med. *8*:816-824.
12. **Garber, M.E., O.G. Troyanskaya, K. Schluens, S. Petersen, Z. Thaesler, M. Pacyna-Gengelbach, M. van de Rijn, G.D. Rosen, et al.** 2001. Diversity of gene expression in adenocarcinoma of the lung. Proc. Natl. Acad. Sci. USA *98*:13784-13789.
13. **Perou, C.M., S.S. Jeffrey, M. van de Rijn, C.A. Rees, M.B. Eisen, D.T. Ross, A. Pergamenschikov, C.F. Williams, et al.** 1999. Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. Proc. Natl. Acad. Sci. USA *96*:9212-9217.
14. **Sorlie, T., C.M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M.B. Eisen, et al.** 2001. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc. Natl. Acad. Sci. USA *98*:10869-10874.
15. **Bittner, M., P. Meltzer, Y. Chen, Y. Jiang, E. Seftor, M. Hendrix, M. Radmacher, R. Simon, et al.** 2000. Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature *406*:536-540.
16. **Alon, U., N. Barkai, D. Notterman, K. Gish, S. Ybarra, D. Mack, and A. Levine.** 1999. 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.
17. **Dhanasekaran, S., T. Barrette, D. Ghosh, R. Shah, S. Varambally, K. Kurachi, K.J. Pienta, M.A. Rubin, and A.M. Chinnaiyan.** 2001. Delineation of prognostic biomarkers in prostate cancer. Nature *412*:822-826.
18. **Miura, K., E.D. Bowman, R. Simon, A.C. Peng, A.I. Robles, R.T. Jones, T. Katagiri, P. He, et al.** 2002. Laser capture microdissection and microarray expression analysis of lung adenocarcinoma reveals tobacco smoking- and prognosis-related molecular profiles. Cancer Res. *62*:3244-3250.
19. **Schwartz, D.R., S.L.R. Kardia, K.A. Shedden, R. Kuick, G. Michailidis, J.M. Taylor, D.E. Misek, R. Wu, et al.** 2002. Gene expression in ovarian cancer reflects both morphology and biological behavior, distinguishing clear cell from other poor-prognosis ovarian carcinomas. Cancer Res. *62*:4722-4729.
20. **Kohane, I.S. A. Kho, and A.J. Butte.** 2002. Microarrays for an Integrative Genomics. MIT Press, Cambridge, MA.
21. **Garrett, E.S. and G. Parmigiani.** 2003. POE: statistical tools for molecular profiling. *In* The Analysis of Gene Expression Data: Methods and Software. Springer, New York.
22. **Eisen, M.B., P.T. Spellman, P.O. Brown, and D. Botstein.** 1998. Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA *95*:14863-14868.
23. **Parmigiani, G., E.S. Garrett, R. Anbazhagan, and E. Gabrielson.** 2002. A statistical frame-work for expression-based molecular classification in cancer. J. R. Statistical Soc. Series B *64*:717-736.
24. **Parmigiani, G., E.S. Garrett, R.A. Irizarry, and S.L. Zeger.** 2003. The analysis of gene expression data: an overview of methods and software. *In* The Analysis of Gene Expression Data: Methods and Software. Springer, New York.
25. **Titterington, D.M., A.F.M. Smith, and U.E. Makov.** 1985. Statistical Analysis of Finite Mixture Distributions. John Wiley & Sons.
26. **McLachlan, G.J. and D. Peel.** 2000. Finite Mixture Models. Wiley, New York.
27. **Oleksiak, M.F., G.A. Churchill, and D.L. Crawford.** 2002. Variation in gene expression within and among natural populations. Nat. Genet. *32*:261-266.
28. **Enard, W., P. Khaitovich, J. Klose, S. Zollner, F. Heissig, P. Giavalisco, K. Nieselt-Struwe, E. Muchmore, et al.** 2002. Intra- and interspecific variation in primate gene expression patterns. Science *296*:340-343.
29. **Carlin, B.P. and T.A. Louis.** Bayes and Empirical Bayes Methods for Data Analysis. Chapman and Hall, Boca Raton, FL, 2000.
30. **Hastie, T., R. Tibshirani, M.B. Eisen, A. Alizadeh, R. Levy, L. Staudt, W.C. Chan, D. Botstein, and P. Brown.** 2000. "Gene shaving" as a method for identifying distinct sets of genes with similar expression patterns. Genome Biol. *1*:RESEARCH0003.1-RESEARCH0003.21.
31. **Ihaka, R. amd R.R. Gentleman.** 1996. A language for data analysis and graphics. J. Computational Graphical Stat. *5*:299-314.
32. **Yeoh, E.J., M.E. Ross, S.A. Shurtle, W.K. Williams, D. Patel, R. Mahfouz, F.G. Behm, S.C. Raimondi, et al.** 2002. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. Cancer Cell. *1*:133-143.
33. **Gilks, W.R., S. Richardson, and D.J. Spiegelhalter.** 1996. Markov Chain Monte Carlo in Practice. Chapman and Hall, London.
34. **Sosinowski, T., A. Pandey, V. Dixit, and A. Weiss.** 2000. Src-like adaptor protein (slap) is a negative regulator of T cell receptor signaling. J. Exp. Med. *191*:463.474.
35. **Rhodes, D.R., T.R. Barrette, M.A. Rubin, D. Ghosh, and A.M. Chinnaiyan.** 2002. Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. Cancer Res. *62*:4427-4433.

Address correspondence to:

Dr. Giovanni Parmigiani
The Sidney Kimmel Comprehensive Cancer Center at
Johns Hopkins University
550 N. Broadway, Suite 1103
Johns Hopkins University
Baltimore, MD 21205-2011, USA
e-mail: gp@jhu.edu