Searching for Evidence of Altered Gene Expression: a Comment on Statistical Analysis of Microarray Data

Janet Wittes, Herman P. Friedman

Hilsenbeck et al. (1) have presented an intriguing experiment that raises important methodologic problems for approaching the assessment of the expression of specific genes. They compared tumors from animals killed at various times to obtain estrogen-stimulated tumors before tamoxifen treatment, tamoxifen-sensitive tumors during tamoxifen treatment but before acquired resistance, and tamoxifen-resistant tumors after resumption of tumor growth. They measured expression values for 588 genes for each of the three tumor types. They argue that a significant deviation in expression between tumor types for a particular gene represents evidence that the specific gene is altered during the development of tamoxifen resistance. In their report, Hilsenbeck et al. define a procedure for identifying "significant" deviations, and they use this method to identify outlier genes. In future studies, they intend to study these outliers to identify "which of the outlier genes are most contributory to the tamoxifen-stimulated phenotype...."

Before embarking on such a search for outlier genes, one needs to know whether different genes need different scales of measurement. Hilsenbeck et al. (1) imply that one can indeed assume the same scale of measurement, i.e., the same variability, for each gene. We believe that replication is essential to determine whether the 588 genes have the same natural scales of measurement. If an individual gene is expressed differentially in tamoxifen-resistant or tamoxifen-sensitive tumors, how do we know that the differential expression reflects the true effect of the tumor rather than underlying differential variability? For example, erk-2 appears as an outlier in the ES/TS and ES/TR plots, where the variables ES, TS, and TR are estrogen-stimulated, tamoxifen-sensitive, and tamoxifen-resistant, respectively; the authors interpret the fact that the expression of erk-2 differs in the estrogen-stimulated tumor as evidence that tamoxifen modifies erk-2 expression. The observation is, however, also consistent with the quite different hypothesis that erk-2 has more variability in expression than most of the other 587 genes. In the case of erk-2, Hilsenbeck et al. have independent verification that the gene is involved in the pathway of interest. What about the genes close to erk-2? How do we know that they are fruitful candidates for further study? Similarly, a gene that does not appear as an outlier and, hence, is not selected as a candidate gene for study may be missed simply because its variability of expression is very low compared with that of other genes. The authors have some data relevant to assessment of within-gene variability, for they measured five tumors of each type; however, they do not describe whether they used those observations to confirm their assumption that the same scale is appropriate for all genes.

If indeed the same scale of measurement is appropriate for all genes, the question then arises whether the authors have employed a useful method for detecting outliers. They have chosen
principal components, a geometric technique for extracting from a set of data an orthogonal coordinate system that defines successive coordinates in order of decreasing variance. Different transformations of the data, however, lead to different principal components. The authors have transformed their data logarithmically. They could have used raw data. They could have normalized their data by subtracting the mean and dividing by the standard deviation. They used the variance-covariance matrix to define their principal components; they could have used the correlation matrix. Each choice leads to different principal components and, hence, potentially to different outliers.

The authors use a robust estimate of variance to define the confidence intervals after extracting the principal components, but they do not use a robust method to extract the components. Look, for example, at their fig. 1. Imagine you were extracting the first principal component from the untransformed trivariate data. Obviously, hsp27 plays an influential role in defining the direction of the first component. Since the deviations of the observations from the first component are the values that determine the outliers, the direction of the first principal component will determine which genes are selected as outliers.

Even more troubling is the fact that the presence of true outliers can determine the direction of the principal components, which in turn, can affect the identification of candidate outliers. Thus, if the purpose of principal components is to find outliers, then we need robust estimates of the components themselves. Otherwise our choice of the outliers depends in part on the outliers themselves. Indeed, an ideal method would be scale invariant and not unduly influenced either by extremely large or small individual observations or by true outliers. We personally would have preferred methods based on ranks, for such techniques would have been invariant to scale and less sensitive to genes with very high or very low expression. The heterogeneity of levels of expression influences the first component, so what does its extraction contribute to the identification of candidate genes? Crucial to the selection of outliers is the assumption that differences between levels of expression for a given gene are independent of the level. It is easy to confirm this assumption. If indeed the differences do not depend on the level, one could simply analyze the differences directly. Box plots of differences would identify univariate outliers in a scale-invariant manner.

Detection of multivariate outliers poses a more difficult problem. Barnett and Lewis (2) present a review of available methods. In addition, software for robust estimation of multivariate location and scale parameters is described by Venables and Ripley (3).

The attempt by Hilsenbeck et al. (1) to verify that their method reliably identifies outlier genes incorporates the same assumptions about the equivalence of scale made by their method. In verifying the utility of a method, one should confirm not only that the approach produces the same results under the original set of assumptions, but also that the same results arise under alternative reasonable assumptions. Ultimately, the only way of knowing whether the method identifies the correct genes is, as the authors point out, to perform experiments that identify whether the outliers are important. Short of that, a simple experimental validation that the method is identifying likely candidate genes would be to study several triplets of arrays. Experimental evidence of the importance of a gene would arise if the same genes popped up in most replications. The authors argue that their method is cost-effective because it does not require replication; if the method identifies too many false-positive genes, it may in fact be expensive.

The explosion of data on gene expression stems in large part from the technical ability to measure expression simultaneously from hundreds or even thousands of genes. The authors wisely recognize the challenge of this mass of data to biology and to statistics. In situations as new and uncharted as this, different people might reasonably use different procedures. New problems need exploration by many methods to gain insight into the data and the relationships of interest. We hope that people will share all data from experiments such as the one described by Hilsenbeck et al. to stimulate methodologic research on this very timely and important problem.

REFERENCES