DOI 10.1002/cyto.10041

The Bioinformatics of Microarray Gene Expression Profiling

John N. Weinstein*, Uwe Scherf, Jae K. Lee, Satoshi Nishizuka, Fuad Gwadry, Ajay, Kim Bussey, S. Kim, Lawrence H. Smith, Lorraine Tanabe, Samuel Richman, Jessie Alexander, Hosein Kouros-Mehr, Alika Maunakea, and William C. Reinhold

Genomics and Bioinformatics Group, Laboratory of Molecular Pharmacology, National Cancer Institute, Bethesda, Maryland

Key terms: gene expression profiling; microarray; biochip; cDNA; oligonucleotide; clustering; clustered image map Cytometry 47:46–49, 2002.

Gene expression profiling will revolutionize biology. That much is universally agreed. But it's harder than it looks. In part, the reasons can be technical—substandard arrays, low signal:noise ratios for rare transcripts, variable backgrounds, cross-hybridizations, the difficulty of processing clinical materials, and so forth. But more often the reasons relate to analysis and interpretation of the data. Inevitably, more time and energy are spent *after* the experiments are finished than before.

We can identify a number of necessary tasks in the analysis of gene expression data, as summarized in Table 1. In the following capsule descriptions, we will focus for concreteness on the two-color fluorescence technologies (1), but analogous steps are pertinent to one-color fluorescence and radioactive detection methods as well. With apologies to the many scientists who have been innovative in this field, we intend, in this short summary, to indicate requirements and options rather than to give a comprehensive review or to apportion credit for the various contributions. The examples will focus primarily on studies from our laboratory.

Task #1: *To establish the computer hardware, software, and personnel infrastructure for handling and analyzing gigabyte or terabyte databases.* There must be somewhere to put the data, and there must be fluent systems for pulling information into the stream of analysis. As data have outgrown Excel (Microsoft, Redmond, WA) spreadsheets, the most common, but by no means only, answers have been database packages like Sybase (Sybase Inc., Emeryville, CA) or Oracle (Oracle Corporation, Redwood Shores, CA). Sometimes, however, flat file formats suffice. For many of the highly multivariate analyses, to be discussed later, hardware speed and memory become significant issues. Most important, however, is the human infrastructure. *Applied bioinformatics*, broadly construed, is practiced by the biologist who is fluent in the use of public and proprietary database resources or who will perform data analyses—preferably under the supervision of a statistically trained individual. Fluency with database resources is something that every biologist should have; microarray data analysis is more specialized. What might be termed *developmental bioinformatics* involves the generation of new algorithms (principally by statisticians or those with expertise in machine learning) and the creation of new software (principally on the basis of expertise in computer science). Experience shows that the best analytical developments arise from close attention to needs arising from actual experimental data sets and biologic questions

Task #2: *To convert images in pixel form to raw expression levels.* Whether one is reading radioisotopically tagged cDNA in a phosphoimager or measuring fluorescent cDNA with a confocal scanner or CCD camera, it is necessary to develop effective image processing algorithms (See 2, 3). The specifics depend on the type of array and detection system used and the quality of the images. As the technologies improve, uncertainties due to such factors as inhomogeneity in the spots, irregular background, scanner artifacts, photobleaching, and lack of spatial registration between channels are diminishing.

Task #3: *To examine the array images for quality control.* This important step is facilitated by software packages that permit surveys of the array image at various levels of resolution and permit individual spots to be examined and compared visually.

Task #4: *To preprocess the expression-level data (i.e., filter, normalize, and/or standardize it).* Generally, the data must be filtered to eliminate flawed spots and genes with insufficient patterns or differences among samples. In the former case, it may be necessary, depending on the nature of the intended analysis, to use statistical or machine learning techniques to impute values for the missing data. The next step is normalization, which usually has been done in the case of two-color studies by tuning a calibration factor, either on the basis of total gene expression in the sample or on the basis of a housekeeping gene

*Correspondence to: John N. Weinstein, National Institutes of Health, Bldg 37, Rm 4E-28, 9000 Rockville Pike, Bethesda, MD 20892
E-mail: weinstein@dtpax2.ncifcrf.gov

Table 1
Tasks in the Analysis of Microarray-Based Gene Expression Data

1. Establish the necessary hardware/software/personnel infrastructure.
2. Convert images in pixels to raw expression levels.
3. Examine the images for quality control.
4. Preprocess the expression level data (filter, normalize, standardize).
5. Analyze and visualize "high-dimensional" data.
6. Search the literature on genes and gene-gene relationships.
7. Integrate the expression data with other types of information.
8. Design the study carefully (replicates, controls)—to be done first.

set. If the samples to be analyzed do not differ from each other dramatically, the former basis for calibration is probably preferable. However, scatter plots of the data with the logarithm of green fluorescence on one axis and that of red fluorescence on the other often show curvature such that a unitary calibration factor will not be the best approach. A number of methods for dealing with this problem have been developed. Ours, included in a computer program called PreProc (L.H. Smith, et al., manuscript in preparation), uses Gaussian-windowed moving averages to fit the curvature (See 4).

Additional problems in analysis arise from asymmetries between the two labeled species in their hybridization properties. Better results are obtained with reciprocal

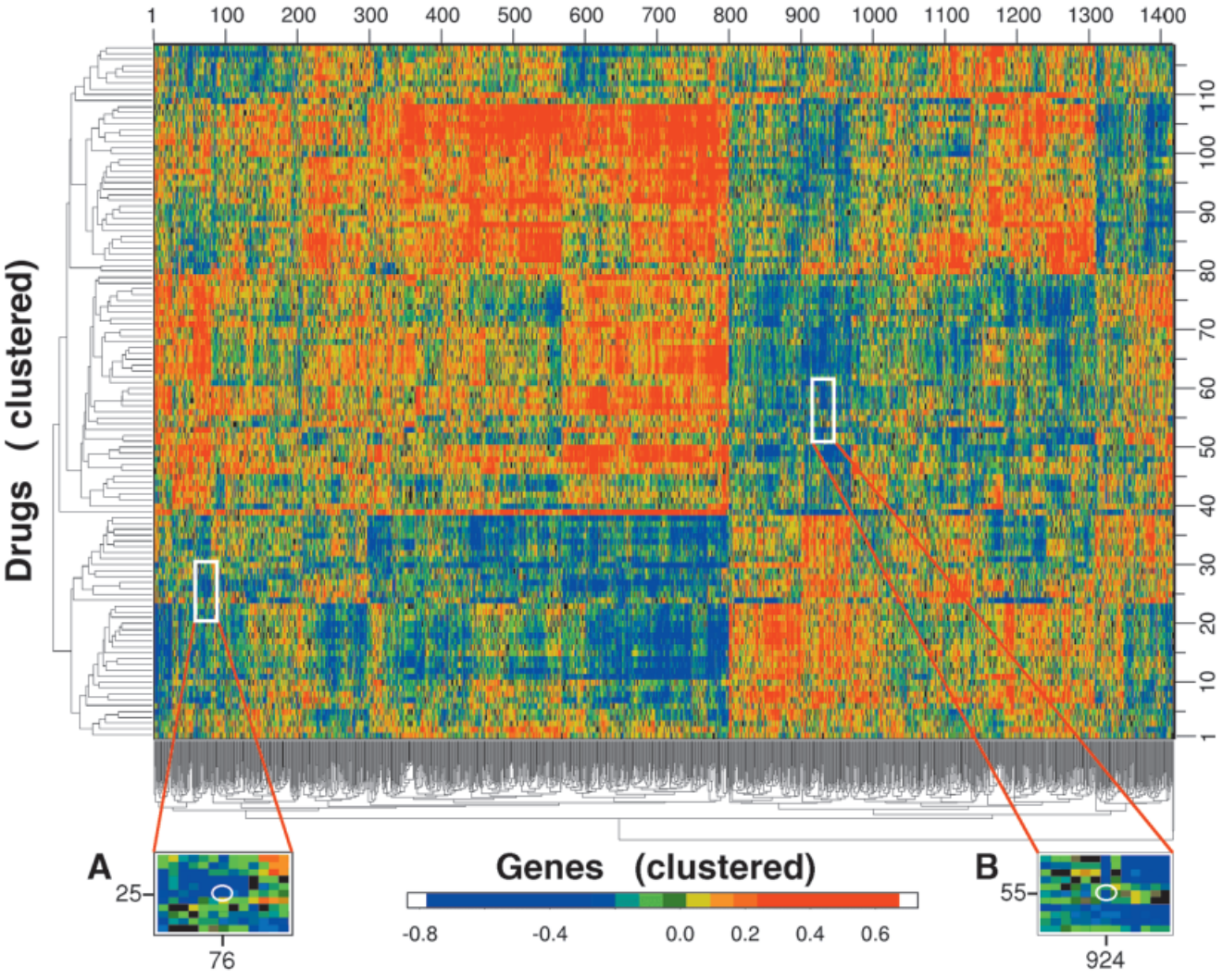

FIG. 1. Clustered image map (CIM) relating activity patterns of 118 tested compounds to the expression patterns of 1376 genes in the 60 cell lines of the National Cancer Institute's Drug Discovery Program. A red point (high positive Pearson correlation coefficient) indicates that the agent tends to be more active against cell lines that express more of the gene; a blue point (high negative correlation) indicates the opposite tendency. Genes were cluster-ordered on the basis of their correlations with drugs (mean-subtracted, average-linkage clustered with correlation metric); drugs were clustered on the basis of their correlations with genes (mean-subtracted, average-linkage clustered with correlation metric). Insert **A** shows a magnified view of the region around the point (white circle) representing the correlation between the dihydropyrimidine dehydrogenase gene (#76) and 5-fluorouracil (#25). Insert **B** is an analogous magnified view for the asparagine synthetase gene (#924) and the drug L-asparaginase (#55). These two correlations have led to insights of potential medical significance (10). Modified from (10).

averaging, that is, by running replicate arrays in which the colors are reversed (4). After normalization, a number of choices must be made before higher level analysis. Should the data be log transformed to obtain more nearly normal distribution and heteroskedasticity? Must it then be thresholded? Should the mean over samples for a given gene be subtracted from the expression levels? The mean over genes for a given sample? Should the levels be divided by a measure of the dispersion such as the standard deviation in one or (by an iterative process) both directions? Should the continuous values be binned, binarized, or turned into ranks for analysis? The answers to these and other such questions often depend on characteristics of the data or the nature of the question being asked. For example, if the important information resides in relative, rather than absolute, gene expression values across a database, one will likely subtract the mean or median across samples.

Task #5: *To analyze and visualize high-dimensional data.* The simplest experimental design is binary—for example, comparison of cancer with normal cells or malignant with non-malignant. More complicated is the time course, for example before and during a treatment. More demanding still is the large database of samples to be analyzed for patterns. The latter two types of data are often presented in the form of what we term clustered image maps, and others have called heat maps. We introduced clustered image maps (CIMs) for pharmacological, genomic, and proteomic studies in the mid-1990s (5–7). Our collaborators later developed a red-black-green color scheme for CIMs (8, 9). Figure 1 shows a slightly more complex CIM (10) that relates patterns of gene expression to patterns of pharmacologic potency in the 60 human cancer cell lines used in the National Cancer Institute's Drug Discovery Program (11, 12). A flexible program for creating CIMs is available at our web site, http://discover.nci.nih.gov

Depending on the questions to be asked, high-dimensional data sets may be analyzed by supervised or unsupervised methods. The former include, for example, techniques based on regression, discrimination, or prediction; the latter on techniques such as clustering (5, 6, 9, 10, 13), principal components analysis, or multidimensional scaling. There is no right method of analysis. Demands of the data and the scientific questions asked will condition the choice.

Task #6: *To search the biomedical literature and public databases for information on genes or gene-gene relationships.* Most gene expression microarray experiments produce long lists of genes with possible significance, and the problem is to distinguish causally interesting relationships from epiphenomenal ones and from statistical coincidence. For that purpose, outside information is generally necessary. Microarray studies are a form of omic research (14–16), but interpretation of the data from them generally requires synergy with classical hypothesis-driven studies of one gene, one gene product, or one process at a time. To facilitate searches of the literature in this context, we developed the program MedMiner (17), which is freely available (along with databases and other analytical tools) at http://discover.nci.nih.gov. It uses a combination of GeneCards from the Weizmann Institute, PubMed from the National Library of Medicine, semantic analysis, syntactic analysis, and keywords to find and organize key sentences from abstracts on genes, gene-gene relationships, and gene-drug relationships. MedMiner can speed up by 5- to 10-fold the rate at which the voluminous literature on important genes is organized and interpreted. A version called EDGAR (Extraction of Data on Genes And Relations) based on deeper semantic analysis is under development (18).

Task #7: *To integrate the expression data with other types of information.* Very often, the gene expression data are most richly understood and most valuable when related to other types of information at the protein, DNA, functional, or pharmacologic level. Figure 1 provides an example of the pharmacologic connection (10).

Task #8: *To design the study carefully (in terms of controls, replicates, internal standards, and design points).* This step should come first, of course. In microarray studies, it is often not feasible to go back afterward and fill in the gaps in an imperfectly designed or executed experimental series. Because arrays are expensive, the tendency is to skimp on replicates and controls, but that is almost always a mistake. Some of the best and most often-used databases placed in the public domain to date suffer from these insufficiencies. Even if it is not practical to use sufficient replicates for all samples, selected replicates (and replicated genes on each array) pay major dividends.

This whirlwind summary of the tasks involved in analysis of microarray gene expression data has by no means touched on all of the important ingredients of the problem, let alone presented them in satisfactory detail. More important than the details of method, however, are common sense and an appreciation of basic statistical principles. Artificial intelligence may one day produce software that can substitute for the human judgment and expertise currently required for gene expression analysis. But such software would look nothing like what is now available.

LITERATURE CITED

1. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 1995;270:467–470.
2. Chen Y, Dougherty ER, Bittner ML. Ratio-based decisions and the quantitative analysis of cDNA microarray images. J Biomed Optics 1997;2:364–374.
3. Ermolaeva O, Rastogi M, Pruitt KD, Schuler GD, Bittner ML, Chen Y, Simon R, Meltzer P, Trent JM, Boguski MS. Data management and analysis for gene expression arrays. Nat Genet 1998;20:19–23.
4. Zhou Y, Gwadry FG, Reinhold WC, Miller L, Smith LH, Scherf U, Liu E, Kohn KW, Pommier Y, Weinstein. Transcriptional regulation of mitotic genes by camptothecin-induced DNA damage: Microarray analysis of dose- and time-dependent effects. Cancer Res Submitted.
5. Weinstein JN, Myers TG, Buolamwini J, Raghavan K, van Osdol W, Licht J, Viswanadhan VN, Kohn KW, Rubinstein LV, Koutsoukos AD, Zaharevitz D, Grever MR, Monks A, Scudiero DA Chabner BA Anderson NL, Paull KD. Predictive statistics and artificial intelligence in the U.S. National Cancer Institute's Drug Discovery Program for Cancer and AIDS. Stem Cells 1994;12:13–22.
6. Weinstein JN, Myers TG, O'Connor PM, Friend SH, Fornace AJ, Kohn KW, Fojo T, Bates SE, Rubinstein LV, Anderson NL, Buolamwini JK,

van Osdol WW, Monks AP, Scudiero DA, Sausville EA, Zaharevitz DW, Bunow B, Viswanadhan VN, Johnson GS, Wittes RE,Paull KD. An information-intensive approach to the molecular pharmacology of cancer. Science 1997;275:343–349.
7. Myers TG, Anderson NL, Waltham M, Li G, Buolamwini JK, Scudiero DA, Paull KD, Sausville EA, Weinstein JN. A protein expression database for the molecular pharmacology of cancer. Electrophoresis 1997;18:647–653.
8. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 1998;95:14863–14868.
9. Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, Iyer V, Jeffrey SS, Van de Rijn M, Waltham M, Pergamenschikov A, Lee JC, Lashkari D, Shalon D, Myers TG, Weinstein JN, Botstein D, Brown PO. Systematic variation in gene expression patterns in human cancer cell lines. Nat Genet 2000;24:227–235.
10. Scherf U, Ross DT, Waltham M, Smith LH, Lee JK, Tanabe L, Kohn KW, Reinhold WC, Myers TG, Andrews DT, Scudiero DA, Eisen MB, Sausville EA, Pommier Y, Botstein D, Brown PO, Weinstein JN. A gene expression database for the molecular pharmacology of cancer. Nat Genet 2000;24:236–244.
11. Boyd MR. Status of the NCI preclinical antitumor drug discovery screen. In: DeVita VT, Hellman S, Rosenberg SA, editors. Cancer: principles and practice of oncology update. Philadelphia: Lippincott; 1989. p 1–12.
12. Paull KD, Shoemaker RH, Hodes L, Monks A, Scudiero DA, Rubinstein L, Plowman J, Boyd MR. Display and analysis of patterns of differential activity of drugs against human tumor cell lines: Development of mean graph and COMPARE algorithm. J Natl Cancer Inst 1989;81: 1088–1092.
13. Michaels GS, Carr DB, Askenazi M, Fuhrman S, Wen X, Somogyi R. Cluster analysis and data visualization of large-scale gene expression data. Pac Symp Biocomput 1998;:42–53.
14. Weinstein JN. Fishing expeditions. Science 1998;282:628–629.
15. Weinstein JN, Buolamwini JK. Molecular targets in cancer drug discovery: Cell-based profiling. Curr Pharm Design 2000;6:473–483.
16. Weinstein JN. Pharmacogenomics: Teaching old drugs new tricks. N Engl J Med 2000;343:1408–1409.
17. Tanabe L, Scherf U, Smith LH, Lee JK, Hunter L, Weinstein JN. MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling. Biotechniques 1999; 27(6):1210–4, 1216–7.
18. Rindflesch TC, Tanabe L, Weinstein JN, Hunter L. EDGAR: Extraction of drugs, genes and relations from the biomedical literature. Pac Symp Biocomput 2000;:517–528.