Technical Report 005
March 20, 2002

Bioinformatics and Whole Genome Technologies

Richard Simon

Biometric Research Branch

Division of Cancer Treatment and Diagnosis

National Cancer Institute

To Appear In Strategies for Cancer Prevention, GJ Kelloff, ET Hawk, CC Sigman editors. Humana Press Inc.

Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment & Diagnosis
National Cancer Institute
9000 Rockville Pike
MSC 7434
Bethesda MD 20892-7434
Tel: (301) 496-0975
Fax: (301) 402-0560
rsimon@nih.gov

1. Introduction

The last half of the twenty'th century saw some dramatic improvements in cancer treatment, including curative treatments for pediatric acute lymphocytic leukemia, Wilms tumor, osteosarcoma, testicular cancer, Hodgkin's disease, diffuse large B cell lymphoma and other neoplasms. Mortality for breast cancer has been reduced as a result of improvements in chemoprevention, early detection and therapy. For many other solid tumors, however, progress has been more limited.

During this same period, however, biology has undergone the biotechnology revolution. This has provided powerful new reagents, experimental techniques, instruments and assays which have led to whole genome DNA sequencing and genome wide RNA transcript quantification. Proteome wide protein quantification is within view.

The last half of the twentieth century also provided the development of the randomized clinical trial, which has made medicine a science. Unfortunately, the developments in the methodology of clinical trials were in many cases more powerful and effective than the interventions that were brought to clinical trial for evaluation. The randomized clinical trial protected us from the broad introduction of ineffective and toxic treatments, however. It also enabled the identification of improved treatments and chemoprevention of breast cancer that would not have been possible otherwise.

It is important that the interventions brought to clinical trial in the twenty-first century have a stronger scientific basis than has been the case to date. Even today, progress in reducing cancer mortality is limited by an inadequate understanding of the molecular basis of these diseases.

The biotechnology revolution has provided new instrumentation and assays that have facilitated the creation of extensive biological data resources. Biology is in the process of becoming an information and system science, but there are important obstacles in utilizing the data becoming available to create biological knowledge. The following sections will explore some of these obstacles and opportunities in developing genomic-based approaches to cancer prevention research.

2. Bioinformatics

Bioinformatics is an ambiguous term that refers to all aspects of the collection, analysis and integration of biological information. It has components that include software engineering, statistical analysis, and algorithm development. Many biologists and organizations are confused about bioinformatics, don't appreciate it's diversity and are struggling to determine how to select staff or structure a bioinformatics group. A common mis-conception is too narrow a focus on software engineering. It doesn't work to hire software engineers to build analysis tools unless they have access to professional advise about data analysis. That advice should generally come from statisticians involved in data analysis and methodology development. A second common misconception about bioinformatics is that it can be effectively structured as a service component. This is certainly not true for the statistical analysis and algorithm development components of bioinformatics. Viewing bioinformatics as a service component that can be purchased reflects a lack of appreciation for the significance of the changing nature of biology. Taking advantage of the revolution that is taking place in biology requires recognition that the statistical and computational scientists working in bioinformatics must be full members of a research team interacting with wet-lab experimentalists on a collaborative basis and conducting self-initiated research on bioinformatics methodology.

Some biologists view bioinformatics as a service activity for creating analysis tools for use by experimentalists. Whereas analysis tools are important, there are many problems, which require inter-disciplinary collaboration in order to fully understand and utilize the data. This type of collaboration is difficult to achieve in many laboratory environments but has been successfully accomplished between biostatisticians and clinical investigators in clinical trial research. Inter-disciplinary collaboration requires not only mutual respect, but also substantial training of statistical and computational scientists in biology and of biologists in statistical and computational matters. It also requires funding of bioinformatics groups as research groups, interacting with experimentalists and also doing research to extend the methodology of bioinformatics. Understanding the complex interactions among genes and among cells will require a greater use of mathematical methods, not for quantification, but for elucidating the principles involved. This effort will require an environment that encourages the best minds to get involved in bioinformatics activities. Those organizations that holding onto outdated views and old hierarchical organizational structures will not be able to take advantage of the opportunities offered in the genomic era.

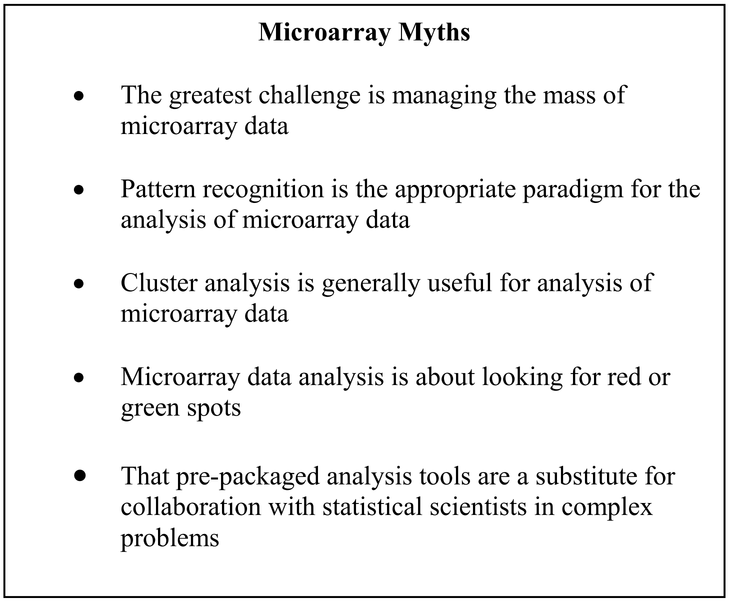
3. DNA Microarrays

DNA microarrays can be used to quantify the abundance of mRNA transcripts for each gene of the human genome in a sample of cells. Although the assay and methods for data analysis are active areas of research, microarrays are currently useful and very powerful. For cancer prevention research, DNA microarrays can be used to elucidate the sequence of changes in gene expression that occur as tumors develop, to identify molecular targets for preventive strategies, and to identify candidate biomarkers for surrogate endpoints in developmental prevention trials.

There are a number of myths prevalent concerning DNA microarrays. Some of these are listed in Table 1. The first is that the greatest challenge in the use of DNA microarrays is management of the large volume of data generated. Whereas effective data management is essential, it can be accomplished using established principals of software engineering and there are an increasing number of commercial and academic database systems available for managing microarray data. The more conceptually challenging problem is determining how to design the experiments, analyze the resulting data so as to obtain reliable information, and to combine sources of information to obtain answers to important biological questions.

A second myth is that pattern recognition is the appropriate paradigm for microarray analysis. Some hold the view that one feeds unstructured specimens into a pattern recognition algorithm to identify unexpected regularities and to provide answers to un-asked questions. This view does not provide a prescription for effective use of microarrays. The microarray is an assay; experiments and analyses must be carefully planned as when using any assay. Microarrays are generally not used to test gene-specific hypotheses. Gene-specific mechanistic hypotheses can often be better addressed using other more sensitive assays. But effective microarray based research has clear objectives, and those objectives drive both the planning of the experiment and the analysis plan. For example, if one wants to identify genes that are dis-regulated early in oncogenesis, then one needs samples of tissue taken at early times during the development of an invasive cancer. The type and number of samples as well as the analysis plan should be determined based on the objectives.

A third myth is that cluster analysis is the generally appropriate method of analysis of microarray data. The microarray is useful for experiments with a wide variety of objectives. Cluster analysis is useful for identifying co-expressed genes and for trying to determine whether a disease is uniform with regard to gene expression, but for many other objectives, cluster analysis is inappropriate. Cluster analysis is not a very powerful approach for comparing expression profiles among pre-defined classes of samples. For example, one may be interested in finding genes that are differentially expressed between tumor and normal samples. The distance metrics used in cluster analysis are generally global distance metrics based on all of the genes or the genes that show variability across the set of samples. This metric may not be sensitive to differences in the relatively few genes that discriminate among the pre-defined classes. Cluster analysis is a non-supervised method, in the sense that it does not utilize information about which sample is in which pre-defined class. Supervised methods for comparing the classes to determine differentially expressed genes and to build multivariate classifiers based on these genes are generally more powerful for such problems.



Microarray Myths

- The greatest challenge is managing the mass of microarray data
- Pattern recognition is the appropriate paradigm for the analysis of microarray data
- Cluster analysis is generally useful for analysis of microarray data
- Microarray data analysis is about looking for red or green spots
- That pre-packaged analysis tools are a substitute for collaboration with statistical scientists in complex problems

The fourth microarray myth listed in Table 1 is that microarray data analysis is about looking for red and green spots. Early use of cDNA microarrays were based on single arrays in which RNA from a collection of wild type cells was labeled with one florescent dye (e.g. Cy3) and hybridized against RNA from a mutated cell type, labeled with a different florescent dye (eg. Cy5). Computer software that performs the image analysis of the pixel level data computes two numbers at each pixel location on the slide. One number is the intensity of the fluorescence when illuminated with laser light of the intensity that causes the Cy3 dye to fluoresce and the other number is the intensity when the slide is illuminated with light that causes the Cy5 dye to fluoresce. The relative magnitude of these two numbers can be color coded and displayed. Usually the color-coding used ranges from green to red. The spots in which there is more mRNA from one sample relative to the other will appear either reddish or greenish.

There are many problems with analysis by color. One problem is that it assumes that you can draw conclusions based on analysis of a single microarray. There are so many sources of variation that are not represented by looking at a single microarray that this is not usually the case. This confusion was enhanced by the publication of formulae and “error models” that claimed to represent confidence intervals for the red to green ratio on a single array. Unfortunately, these confidence intervals do not incorporate many important sources of variability. RNA is easily degradable and differences in handling the cells or tissues compared on an array can greatly influence the results. The labeling reaction is also a major source of variability and it cannot be properly controlled when analyzing a single array. Not only is there substantial variability involved in labeling, there is also bias. The two labels commonly used, Cy3 and Cy5, have different affinities for DNA and different sensitivities for fluorescence. The relative affinities are gene dependent and the fluorescence bias varies based on the level of expression. One may attempt to control some of these biases by “normalization” of the data, but normalizations are imperfect and obtaining unbiased results generally requires multiple arrays and in some cases dye-swap replication.

Generally the greatest source of variation in microarray studies is biological variation. If you wish to compare tissue of tumors from a specified tissue to normal tissue of the same type, you need to study multiple tumors and multiple normal tissues. If you do multiple arrays with sub-aliquots of the same specimen of tumor and normal tissue, then you may learn something about relative gene expression in those two RNA samples. You may be able to control for the labeling bias and variation and the hybridization variation, but you won't know whether the differential expression found is the result of differential tissue handling, or whether it applies more generally to tumors and normal tissues of that type. Even for experiments involving cell lines instead of tissues, gene expression can seriously vary with the confluence state at which the cells are harvested. As the cells start to crowd and compete for nutrients, various pathways get turned on and others get turned off. Hence, comparing expression in an RNA sample from one cell line to expression in an RNA sample from a different cell line, unless the experiment is replicated at the biological level of repeating growth and harvest of the cells, we will not know whether the results are more than experimental artifacts. The amount of replication appropriate depends on the degree of variability from all sources and is discussed somewhat by Simon et al. (1) along with other aspects of the design of microarray experiments. There is value in doing some replication of arrays at a lower level, duplicate arrays for the same RNA samples independently labeled, in order to assure that your experimental technique, instrumentation and reagents are working properly.

The final myth listed in Table 1 is that software analysis tools are a substitute for collaboration with professional statistical scientists on major studies using microarrays. Many biologists perform a small number of microarrays in order to get a view of gene expression in order to plan more definitive experiments using either microarrays or other technologies. It is important that good software analysis tools be available for such use. We have developed BRB-Array Tools (2) as a DNA microarray analysis package for use by biologists. BRB-Array Tools is also intended as a tool for helping to educated biologists in good statistical practices in analysis of microarray data. There are too few available statisticians experienced in the analysis of microarray data and often if falls to biologists to analyze their own data. For many experiments, however, collaboration with professional statisticians experienced in the design and analysis of microarray data is very important. Studies involving DNA microarrays are in many ways more complicated than clinical trials or cohort studies. There are many more opportunities to miss-analyze the data and publish erroneous conclusions. There are many signal processing analysis steps which require careful examination of the raw data. These include image analysis of pixels, background adjustment, quantification of signal, normalization, combining probe signals on Affymetrix arrays, identification of artifacts and quality assessment. There are many types of experimental artifacts. There are also many competing methods of analysis, not all of which are equally good. The available software packages cannot be relied upon to produce good data automatically for all sets of arrays. There are equally many complex issues of data analysis after the signal processing stage. These include issues such as what analysis methods to use, how to control for multiple comparisons, how to evaluate a multivariate classifier and how to perform and validate cluster analysis. Some available software packages do not handle these issues effectively or even validly. Many packages over-emphasize cluster analysis for problems where it is inappropriate. The Affymetrix software currently available does not even provide for the comparison of expression levels of genes in two sets of samples by a standard statistical test. It provides only for comparison of individual samples.

In comparing expression levels of genes in two sets of samples, cognizance must be taken of the multiple comparison issue. If one compares expression levels between two sets of samples for 10,000 genes, then one expects 500 false positives statistically significant at the 0.05 level. This is the expected number even-though the expression levels are correlated among sets of some of the genes. The correlation effects the variance of the distribution of the number of false positives, but not the expected number. Hence, the conventional 0.05 significance level in comparing expression levels of individual genes is not appropriate. Some investigators select genes differentially expressed between the two classes at the 0.05 level, and then cluster the samples with regard to that gene set. This is an erroneous way of evaluating whether the classes are different with regard to expression profile. Even if two classes do not truly differ with regard to expression profile, there will be on average 500 (out of 10,000) genes significant at the 0.05 level and the classes will cluster separately with regard to this set of false positive genes.

There are many methods for controlling the number of false positives. One is to use a stringent 0.001 threshold for declaring significance. There are other methods that specifically control the "false discovery rate" which is the proportion of genes claimed to be differentially expressed between the classes, which are false positives. Some methods also take into account the correlation structure of expression level among genes and thereby gain statistical power.

There is substantial interest in developing multivariate classifiers of two or more pre-defined classes based on gene expression levels. There is a substantial literature on different types of mathematical functions that can be used as classifiers ranging from linear discriminant functions to neural networks. These methods were not developed, however, for problems where the number of candidate predictors vastly exceeds the number of cases (samples). Many of the methods do not work well in that setting. The key principles in developing an effective multivariate classifier are selection of informative features and avoiding over-fitting the data. For contexts where the number of candidate predictors (genes) is orders of magnitude greater than the number of cases, complex methods that have many parameters to be determined from the data, such as neural networks, often perform very poorly (3).

It is essential to obtain an unbiased estimate of the misclassification rate of multivariate classifiers in high dimensional situations with relatively few cases. Applying the classifier to the same set of cases from which it was developed results in a severely biased estimate of misclassification rate unless a cross-validation (or other bias reduction) procedure is properly used. One simple type of cross-validation is to separate the data into a training set and a validation set before *any* analysis is performed and to not look at the validation set until a fully specified model is developed on the training set. The fully specified classifier is then applied to the cases of the validation set, without any additional variable selection, fitting of parameter values or estimating cutoff values. An unbiased estimate of misclassification rate can be obtained in this way.

An alternative approach that can be used in some circumstances is algorithmic cross-validation. For example, with algorithmic leave-one-out cross validation (4) one sample is set aside as a singleton validation set. The classification model is *developed from scratch* on the training set defined by the remaining samples. The classification model developed on the training set is used to classify the validation sample that was excluded from the training set. One then records whether that classification was correct or not. This process is repeated n times, where n is the total number of samples. Each time a different sample is excluded and a classification model *developed from scratch* using the same algorithm on the training set consisting of the remaining samples. Leave-one-out cross-validation is often performed incorrectly. "Developed from scratch," means that all variable selection and other steps must be re-performed on each training set. The variables (genes) in the model will change for each training set. No pre-analysis that uses the class labels can be performed using the entire dataset. Inexperienced data analysts sometimes select the genes using the entire data, or select the principal components to be used using the entire data. Then they cross-validate the parameters of the model. This generally gives a quite biased estimate of the misclassification rate.

4. Conclusion

Biotechnology has given rise to genomic and important new tools, approaches and opportunities for understanding the nature of cancers. Such understanding will lead us to major improvements in reducing cancer mortality through prevention, early detection and molecularly targeted treatment. Taking advantage of the opportunities available to us will require, however, a greater appreciation of the changes that have taken place and the need for closely interacting with statistical and computational scientists in a setting of collaboration among equal research scientists.

References

1.Simon R, Radmacher MD, Dobbin K. Design of studies using DNA microarrays. Genetic Epidemiology, 2002 (In Press).

2.Simon R, Peng A. BRB-ArrayTools Users Guide. http://linus.nci.nih.gov/BRB-ArrayTools

3.Dudoit S, Fridlyand J, Speed T. Comparison of discrimination methods for the classification of tumors using gene expression data. Journal of the American Statistical Association, 2002 (In Press).

4.Radmacher MD, McShane LM, Simon R. A paradigm for class prediction using gene expression profiles. Journal of Computational Biology, 2002 (In Press).