Some Statistical Issues in Microarray Gene Expression Data

Matthew S. Mayo, University of Kansas Medical Center
Byron J. Gajewski, University of Kansas Medical Center
Jeffrey S. Morris
COMMENTARY

Some Statistical Issues in Microarray Gene Expression Data

Matthew S. Mayo, Byron J. Gajewski and Jeffrey S. Morris

Department of Preventive Medicine and Public Health, Center for Biostatistics and Advanced Informatics, Kansas Masonic Cancer Research Institute, and Schools of Allied Health and Nursing, Center for Biostatistics and Advanced Informatics, University of Kansas Medical Center, Kansas City, Kansas; and Department of Biostatistics and Applied Mathematics, The University of Texas M.D. Anderson Cancer Center, Houston, Texas

INTRODUCTION

The use of DNA microarray gene expression data (1–4) in health research has exploded over the last few years. This technology is useful for making inferences about the genomic profile of an individual for use in risk assessment and/or detection of effects. As exemplified by the two papers appearing in this issue (5, 6), researchers are using microarray gene expression data to understand the influences of non-ionizing as well as ionizing radiation on the genotype. Therefore, since there may be a strong future for the use of microarray gene expression data in radiation research, the radiation research community may expect to see more research papers using this technology. It is vital that radiation researchers understand the fundamental statistical issues so that one can decide whether the research follows good, sound, fundamental statistical principles. The purpose of this commentary is to provide very basic guidance (Table 1) as to what should be the minimum standards for such work, statistically, for microarray gene expression data studies. Two of the biggest issues for assessing the validity of an analysis of microarray data studies are preprocessing the data and the analysis of the processed data (7).

PREPROCESSING DATA

The basic analytical data for microarrays involve an n-by-p table of data, where the n rows each correspond to an array and the p columns each correspond to a gene. Given this table, any of a number of different analyses can be done to identify genes differentially expressed across experimental conditions, find clusters of samples or genes with similar expression levels, or build predictive models based on the gene expression levels of sets of genes. To obtain this table, however, a number of preprocessing steps must first be taken to screen out poor-quality arrays, normalize the expression levels across arrays and filter out confounding effects, and quantify the gene expression values from the raw array data. If these steps are not performed properly, the table of expression levels may not be accurate, preventing the possibility of obtaining valid conclusions from the study. The specifics of these steps tend to differ across different platforms, e.g. between Affymetrix oligonucleotide arrays and glass cDNA arrays.

Some comprehensive preprocessing packages have been developed for analyzing Affymetrix data (8, 9). For glass cDNA arrays, various image processing algorithms are necessary to quantify the spots and adjust for background effects. For all platforms, various normalization procedures must be used to filter out systematic biases that can occur within the experiment (10). For example, with cDNA arrays, dye bias is an important factor that must be accounted for (11). Yang et al. (11) describe and compare several types of statistical normalizing techniques including global, intensity-dependent, within-print tip group, scale and composite normalization techniques. Kerr, Martin and Churchill (12) discuss ANOVA-based methods that can simulta-
100 subjects in two groups. Using a two-sample test to compare genes expressed across the specified groups. For example, radiation researchers again could be directed down a class comparison path, once the data are preprocessed (normalized). The first is class comparison, which is used when one is attempting to see which gene expression profiles differ between prespecified groups. Specifically, class comparison involves identifying a list of genes that are differentially expressed across the specified groups. For example, radiation researchers might be interested in seeing how radio-frequency (RF) fields associated with cellular phone use affect the gene expression profile. One might match subjects, animals or cells exposed to RF fields with similar specimens not exposed to RF fields. Then one would compare the gene expression profiles between the two groups.

The key statistical issue in class comparison involves controlling for the multiple comparisons. To appreciate this, consider an example in which there are 1,000 genes and 100 subjects in two groups. Using a two-sample t test with a significance level of 0.05, we would expect 50 of the gene expressions to be significant by chance alone, even if in fact the two groups had the same gene expression profiles. In some settings, classical multivariate statistical techniques such as Hotelling’s $T^2$ can be used to assess differences in the multivariate setting in a way that controls for multiplicity. However, these methods require at least as many samples (n) as variables (p), which clearly does not hold in the setting of microarrays. Thus these approaches cannot be applied here. Another way to alleviate this problem is to adjust for multiple comparisons using a Bonferroni adjustment. But using a Bonferroni adjustment will result in a per comparison significance level of $0.05/1000 = 0.00005$, which might be too conservative because of the difficulty in achieving statistical significance. The conservatism comes from the fact that Bonferroni controls the experiment-wise error rate, the probability of even one false positive, at the alpha level. This criterion is usually not appropriate in the context of microarrays, since we are willing to have some false positive genes, and this stringent criterion leads to a large number of false negative genes whose group effects are not discovered. As a result, researchers have developed a new criterion, the False Discovery Rate (FDR), which is widely considered to be a more appropriate criterion in this context. Rather than controlling the probability of making even one false discovery, the FDR controls the expected proportion of false discoveries.

A landmark paper by Storey and Tibshirani (14) described the basics of FDR. Let $F$ be the number of false positives and $T$ be the number of true positives. The FDR for an experiment is the average $F/(F + T)$, or the ratio of the false positives divided by all positives. There are a number of methods available for controlling FDR, including significance analysis of microarrays (SAM) (15), Empirical Bayes methods (16), beta-uniform mixture (BUM) (18), spacing LOESS histogram (SPLOSH) (19, 20), and other Bayesian methods (21–23). Most of these methods assume independence among the genes. Some recent research relaxes this assumption and can handle dependent genes (24). Some of these methods work by first computing a suitable test statistic, e.g. a t statistic, for each gene. Then they identify a cut point on the P values below which the gene is considered significant, while controlling the FDR at some level alpha. Another alternative to the FDR approach is to do permutation tests (25, 26). Without suitably addressing the multiple comparisons issue with these high-dimensional data, radiation researchers again could be directed down a dubious research path.

One debate in all statistical research is whether to use parametric statistical methods or non-parametric statistical methods for making class comparisons. For example, some parametric methods rely on assigning a normal distribution to the data and using, for example, a t test to compare genes across groups. Other approaches do not make distributional assumptions and instead use the rank sum test. Obviously one would like to use the technique that minimizes the FDR. Shedden et al. (27) compared seven methods for pro-

### TYPES OF MICROARRAY STUDIES

Once the data are preprocessed (normalized), the researcher moves to the analysis of the n-by-p table. To understand the analysis issues, one must understand the goals of the particular research project. There are three basic types of research goals when one is using microarray data. The first is class comparison, which is used when one is attempting to see which gene expression profiles differ between prespecified groups. Specifically, class comparison involves identifying a list of genes that are differentially expressed across the specified groups. For example, radiation researchers might be interested in seeing how radio-frequency (RF) fields associated with cellular phone use affect the gene expression profile. One might match subjects, animals or cells exposed to RF fields with similar specimens not exposed to RF fields. Then one would compare the gene expression profiles between the two groups.

The key statistical issue in class comparison involves controlling for the multiple comparisons. To appreciate this, consider an example in which there are 1,000 genes and 100 subjects in two groups. Using a two-sample t test with a significance level of 0.05, we would expect 50 of the gene expressions to be significant by chance alone, even if in fact the two groups had the same gene expression profiles. In some settings, classical multivariate statistical techniques such as Hotelling’s $T^2$ can be used to assess differences in the multivariate setting in a way that controls for multiplicity. However, these methods require at least as many samples (n) as variables (p), which clearly does not hold in the setting of microarrays. Thus these approaches cannot be applied here. Another way to alleviate this problem is to adjust for multiple comparisons using a Bonferroni adjustment. But using a Bonferroni adjustment will result in a per comparison significance level of $0.05/1000 = 0.00005$, which might be too conservative because of the difficulty in achieving statistical significance. The conservatism comes from the fact that Bonferroni controls the experiment-wise error rate, the probability of even one false positive, at the alpha level. This criterion is usually not appropriate in the context of microarrays, since we are willing to have some false positive genes, and this stringent criterion leads to a large number of false negative genes whose group effects are not discovered. As a result, researchers have developed a new criterion, the False Discovery Rate (FDR), which is widely considered to be a more appropriate criterion in this context. Rather than controlling the probability of making even one false discovery, the FDR controls the expected proportion of false discoveries.

A landmark paper by Storey and Tibshirani (14) described the basics of FDR. Let $F$ be the number of false positives and $T$ be the number of true positives. The FDR for an experiment is the average $F/(F + T)$, or the ratio of the false positives divided by all positives. There are a number of methods available for controlling FDR, including significance analysis of microarrays (SAM) (15), Empirical Bayes methods (16), beta-uniform mixture (BUM) (18), spacing LOESS histogram (SPLOSH) (19, 20), and other Bayesian methods (21–23). Most of these methods assume independence among the genes. Some recent research relaxes this assumption and can handle dependent genes (24). Some of these methods work by first computing a suitable test statistic, e.g. a t statistic, for each gene. Then they identify a cut point on the P values below which the gene is considered significant, while controlling the FDR at some level alpha. Another alternative to the FDR approach is to do permutation tests (25, 26). Without suitably addressing the multiple comparisons issue with these high-dimensional data, radiation researchers again could be directed down a dubious research path.

One debate in all statistical research is whether to use parametric statistical methods or non-parametric statistical methods for making class comparisons. For example, some parametric methods rely on assigning a normal distribution to the data and using, for example, a t test to compare genes across groups. Other approaches do not make distributional assumptions and instead use the rank sum test. Obviously one would like to use the technique that minimizes the FDR. Shedden et al. (27) compared seven methods for pro-
producing Affymetrix expression scores. They found that the
data processing method has a much bigger impact on FDR
than the choice between using a parametric or a non-para-
metric technique. This may be because the t statistic is ro-
 bust to distributional assumption.

The second type of research study, called class predic-
tion, has a similar setup for comparison in that there is a
prespecified set of groups. However, now the researcher is
interested in using the microarray data to predict the group
that they belong to. For example, does one’s genotype pre-
dict the group of subjects exposed to radiation compared to
the group of subjects that are unexposed?

The statistical problems associated with class prediction
are classic in data mining. Data mining techniques can have
a problem with creating models that over-fit the data, mak-
ing it difficult to predict a future dataset using the same
decision rule. Researchers deal with this problem by split-
ting the data into two pieces, a training data set and a val-
 idation data set. The training data set is used to select the
decision rules for prediction. Then the validation data set
is used to test whether the model defined from the training
data set is reasonable. One problem with such an approach
is that the analyst must be very careful how training and
validation data sets are selected. That is, a random sample
should be selected—not a sample that works well for the
answer desired. Another problem can occur if the analysts
improperly select their model using the validation set. That
is, they use the training set to get a model and test the
model. The testing fails, so the analyst goes back to the
training set and fits a new model then tests the new model.
They repeat this process until they get a good fit on their
training set. This is invalid because they are using their
validation data to fit their model and also assess their pre-
diction errors. The prediction errors obtained by such a pro-
cedure are not valid.

One way to fix this problem is to use cross-validation.
For example, suppose that the researcher has 50 arrays; a
fourfold cross-validation would be to select 40 arrays, fit a
model, and then use the 10 other arrays to validate the
model. The analyst would repeat this five times, getting a
fourfold cross-validation of their model. The key to this
methodology is that the user must perform the model se-
lection at each of the steps to account for the uncertainty
in the model selection process.

The third type of research study, called class discov-
ery (28), involves no prespecified classes in the data. That is,
the researchers are interested in finding some sort of a struc-
ture to the genome of a particular population. One way to
do this is using clustering techniques. There are two ways
to classify gene expression. One way is to cluster the sam-
ples. In this case one might be interested in seeing how
radiation-exposed subjects and unexposed subjects cluster
without specifying their classes. A second example of clus-
tering would be to cluster the genes. In this case the re-
searcher might be interested in identifying sets of genes
with correlated expression levels. Clustering is useful for
identifying relevant biological structure in the data and also
for identifying structure caused by systematic biases in the
data. For example, if the samples cluster strongly based on
the day on which the arrays were run, that may indicate a
strong day effect that must be accounted for in the analysis.

EXAMPLES IN THIS ISSUE OF RADIATION RESEARCH

In this issue of Radiation Research, there are two papers
using microarray gene expression data. The study of effects
of exposure to RF radiation on gene expression by Qutob
et al. (5) is an example of class comparison. They used
James-Stein shrinkage F tests to compare gene expressions
in an exposed group to those in a control group. This tech-
nique is attractive since they used it to control the FDR.
They normalized their expression values by using a tech-
nique in the library of a freeware statistical package called
R. The second paper in this issue, by Whitehead et al. (6),
is also a class comparison study of how exposures to dif-
terent types of RF fields affect the genotype in mice. They
normalized their array data using a scaling technique. The
authors state that the GeneChip data were analyzed using
a two-tailed t test and that the expected number of false
positives was estimated from t tests on 20 permutations
of the six sham RF-field-exposed samples. The use of the
sham-sham experiments to estimate the number of false
positives and control the overall type I error rate is an in-
novative and well-thought-out approach to estimate and
control the FDR. One note of caution should be made; the
lack of statistical significance should not be used to con-
clude a lack of effect. The lack of significance could be a
function of the small sample sizes and the inability to detect
differences, especially when using stringent methods to
control the FDR. Thus a slightly more tempered title and
conclusions might have been warranted for this manuscript.
Neither manuscript details a priori power calculations for
readers to determine if either study had a chance to detect
clinically meaningful differences; this would have bolstered
the conclusions that could have been drawn from each. The
data from each of these studies could be used to aid the
design of future studies to detect clinically meaningful dif-
f erences in gene expression.

CONCLUSION

In this commentary we have outlined some of the basic
statistical issues in microarray gene expression data. This
technology is dynamic. The preprocessing and statistical
analysis techniques are evolving daily. Many of the prepro-
cessing and analysis methods for microarray data are de-
tailed, and open-source software is made available as part
of the Bioconductor project (29), which is continually up-
dated as new methods are developed. This is an excellent
source for methods for analyzing data from microarray ex-
periments.

Keeping up with this technology can be daunting, but
the fundamentals are clear. By spending the necessary time properly designing your study (fundamental to all studies), preprocessing the data, and using a statistical analysis technique that corresponds to the particular goal of the study, radiation scientists can feel confident that their research findings are sound and will likely pass the scrutiny of peer review. For both papers in this issue (5, 6), the authors used appropriate methodologies to preprocess the data and analyze the data and used methods to control the FDR. They can conclude there is insufficient evidence to say that RF-field exposure alters gene expression.

ACKNOWLEDGMENT

Dr. Mayo’s work was supported by the following grant: NCI R23 CA095835.

Received: February 24, 2006; accepted: March 7, 2006

REFERENCES