Microarray Data from a Statistician's Point of View

Johanna Hardin

Reports in the news often tell about how genes determine the chances of getting a particular disease or how a genetic mutation can increase susceptibility to certain environmental changes. For example, Familial Adenomatous Polyposis (FAP) is a type of colon cancer that affects one in 8,000 people in the United States. FAP is caused by a mutation in the adenomatous polyposis coli (APC) gene. It has been estimated that a person with FAP has over three times the relative risk of dying than a person without FAP (Nugent et al. 1993).

It is well known that the DNA in a cell's nucleus contains the instructions for building proteins. A gene is a segment of DNA that contains the instructions for building a specific protein. If different genes are active, then different proteins will be produced in a cell. Skin cells are different from muscle cells, for example, because different proteins are present in the two different types of cells. When a gene is active in a cell, we say that the gene is "expressed." Information about genetic activity can give insight into biologic processes and cell behavior—both normal and cancerous.

Measuring genetic activity is the role of molecular biologists. Until recently, scientists analyzed gene activity one gene at a time. Now activity can be measured on tens of thousands of genes simultaneously using a new tool known as a DNA microarray (Eisen and Brown, 1999). Interpreting the gene expression data is the role of statisticians. The huge volume of data from microarray analyses brings new statistical challenges and the need for new analytical techniques.

As statisticians, our role in many scientific fields, particularly in the field of molecular biology, is vital and fascinating. Because microarray data analysis is a new and expanding research area, I cannot hope to cover in this article all of the current research associated with microarray analysis. So my goal is to give an overview of the analysis process and the related statistical issues.

Johanna S. Hardin (jo.hardin@pomona.edu) is an Assistant Professor at Pomona College. She received a BA from Pomona College and an MS and PhD from the University of California, Davis. Her current research interests include analysis of microarray data (normalization, distributional qualities, clustering, and outlier detection) and other high dimensional data sets.

Why Microarrays?

Information (that can be obtained from microarrays) about genes helps us answer a myriad of biological questions:

- What genetic differences are there between healthy people and people with a particular disease?
- Are there genetic subgroups of people with a particular disease who respond positively to a given treatment?
- What kinds of genetic changes happen across time or after frequent doses of a treatment?
- Which genes are co-regulated—have expression levels that increase or decrease concurrently—in a particular biological system?
- What is the likelihood of acquiring a particular disease, given a person's genetic make-up?

What is a Microarray?

DNA microarrays, first introduced commercially in 1996, come in a variety of forms, but they all contain the same basic design. Each microarray consists of thousands of single strands of genetic material tethered to a "chip" the size of a thumbprint. The chips (which are not reusable and should not be confused with computer chips that can store and restore information) are produced at numerous academic and research laboratories, and they are also produced commercially.

Microarray technology uses a fundamental property of DNA called "complementary base pairing." Our DNA gives the blueprint for the functioning of the cell written in sequences of chemical bases: adenine (A), cytosine (C), guanine (G), and thymine (T). These bases bind in a double helix structure to create the DNA molecule (See Figure 1). At each rung along the DNA ladder, A always binds with T, and C always binds with G. Thus A is complementary to T, and C is complementary to G. Each spiral strand is connected to a complementary

strand by the paired bases. A subsequence of the gene characterized by TGAAACT on one strand would have ACTTTGA on the complementary strand of the DNA molecule.

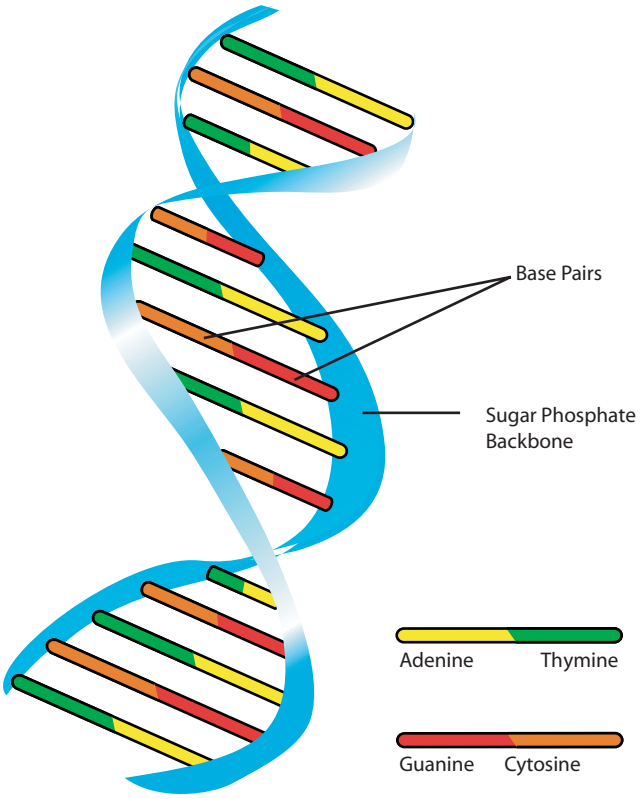

Figure 1. Illustration of the DNA double helix molecule showing the complementary base pairing on the rungs of the DNA ladder.

The DNA code is identical in each cell nucleus through the entire body. However, in order for cells to function appropriately, each different cell type receives a different message from the DNA. A segment of DNA is converted into an intermediary form known as messenger RNA (mRNA) that exits the nucleus and serves as a template for building proteins.

Consequently, we could determine which genes are expressed in a cell by measuring the quantity of mRNA there is in the cell corresponding to that gene. However, free mRNA in a cell is very unstable, so it is treated with an enzyme to convert the mRNA back into DNA. This form of DNA is known as complementary DNA (cDNA).

Through a denaturing process the double-stranded DNA molecules in the sample are unzipped down the middle into two single-stranded molecules. The microarray chip itself also contains single strands of genes that will attract the single gene strands from the sample. The single strands from the sample will bind with the single strands on the microarray chip to reform the DNA double helix.

In a microarray experiment, the test sample is labeled with a dye and a reference sample is labeled with a dye of a different color. The reference sample serves as a control to which the gene expression in the test sample is compared. For instance, if we wanted to determine which genes are expressed in a tumor sample, we could use a tissue sample from a healthy individual as the reference sample. We would then compare the expression level of each gene in the tumor sample to the expression level of each gene in the reference sample. Suppose the tumor sample had been labeled with a red dye and the reference sample had been labeled with a green dye. Then a red spot on the microarray would indicate that the gene corresponding to that spot is expressed at a higher level in the tumor sample than in the reference sample. Similarly, a green spot would indicate that the gene is expressed at a lower level in the tumor sample.

There are several techniques for constructing DNA microarrays (Schena et al. 1995; Velculescu et al. 1995; Lockhart et al. 1996). Though there are slight differences in the microarray technologies, one basic outline of the microarray procedure can be summarized as follows:

1. Label the sample with a fluorescent dye.
2. Isolate the cDNA from the cells of interest, e.g., tumor cells, plasma cells, etc.
3. Denature the sample so that the cDNA are in single strands instead of the double helix form.
4. Place the sample onto the microarray chip and allow the double helix structure to restore itself.
5. Wash the remaining sample off the chip so only the parts of the sample that have bound to the chip remain.
6. Scan the microarray chip with a laser to quantify the fluorescence of each individual gene. The more of the sample that is stuck to the chip, the higher the fluorescence.

A. Malcolm Campbell at Davidson College has put together an animation of the microarray process which can be seen at website: *www.bio.davidson.edu/courses/genomics/cgip/chip/html*.

In general, the amount of activity of a gene is represented by the number of replicates of that gene in a particular sample of cells. A high fluorescence level indicates that multiple copies of a gene have bound to the chip and that the gene has high activity in the cell. Similarly, a low fluorescence level indicates low activity of the gene in the cell. By quantifying the fluorescence, the gene activity can be compared across different samples, e.g., a group of healthy samples compared to a group of tumor samples.

A sample scan of part of a chip is shown in Figure 2 (see page 6). The image shown in the figure is only part of the chip. Each spot represents a gene and there are thousands of genes on a chip. A red spot indicates that sample 1 (the "red" sample) has high genetic activity for that gene. A green spot indicates that sample 2 (the "green" sample) has high genetic activity for that gene. The yellow spots indicate the genes where the

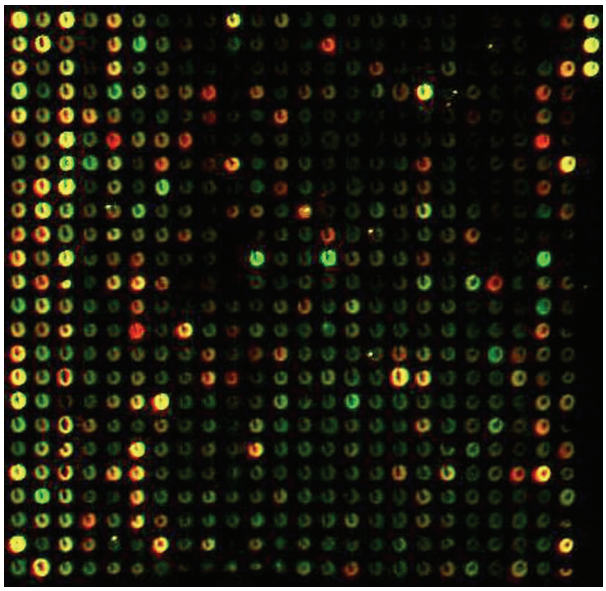

Figure 2. Part of a microarray chip. Each spot represents one gene and the color represents the activity level of the gene in the test sample.

two samples have similar activity, and the black spots indicate where there is no activity.

An example of some microarray data is given in Table 1. The data came from an experiment on aging yeast in Laura Hoopes' lab at Pomona College. The test sample (treatment) contains older yeast cells, while the reference sample (control) contains younger yeast cells. The test sample was dyed red and the reference sample was dyed green. The table only shows the expression level of ten genes as an illustration. In an actual analysis there would be data for many more genes.

From the numerical values we can identify the genes that are highly expressed overall in the experiment and the genes that are just barely expressed. Note genes RPS11b and YBR124W, for example.

Additionally, by taking the ratio of intensities we can identify the genes that are most highly expressed in the treatment sample relative to the control sample and vice versa. Taking the logarithm of the ratio helps to further distinguish the genes with the highest and lowest relative expression levels. Note genes HSP26 and YRO2, for example.

Gene	Red Intensity	Green Intensity	Ration of Intensities	Log $_2$ of Ratio
YBR124W	92	78	1.179	0.238
YBR100W	103	77	1.338	0.420
MRS5	369	357	1.008	0.012
ECM33	3423	2663	1.285	0.362
YBR075W	196	133	1.474	0.559
HSP26	805	175	4.600	2.202
VAP1	158	175	0.903	-0.147
YRO2	118	373	0.316	-1.660
YBR051W	125	135	0.926	-0.111
RPS11B	3855	3739	1.031	0.044

Table 1. Sample microarray data from an experiment on aging yeast cells. Red intensity refers to the test sample and green intensity refers to the reference sample. The ratio of intensities tells us the multiplicative change and the log base-2 ratio gives the difference of the data after a log transformation. Data courtesy of Laura Hoopes of Pomona College.

What is the Statistician's Role?

Although it is preferable for the statistician to have a hand in the experimental design, the statistician often comes into a microarray analysis project once the data have been collected. The statistician's job is to use the numerical fluorescence levels to make claims about the populations of interest. Of course, the methodology will depend on the question at hand. The computations can be broken down into two main parts: data cleaning and data analysis.

Though the microarray construction seems straightforward in theory, in reality there are numerous sources of variation. For example:

- Spots that are not systematically placed on the chip,
- Samples that smear outside of the measurement surface,
- Dyes that fluoresce at different levels (green is "stronger" than red), or
- Arrays with a variable amount of dye.

To address these problems, the data cleaning step involves image processing, normalization, and standardization. Current research on all three cleaning steps is active and growing. In this article I focus on data analysis instead of data cleaning, assuming the data are already "clean." Many software programs designed for microarray analysis give options for cleaning the data.

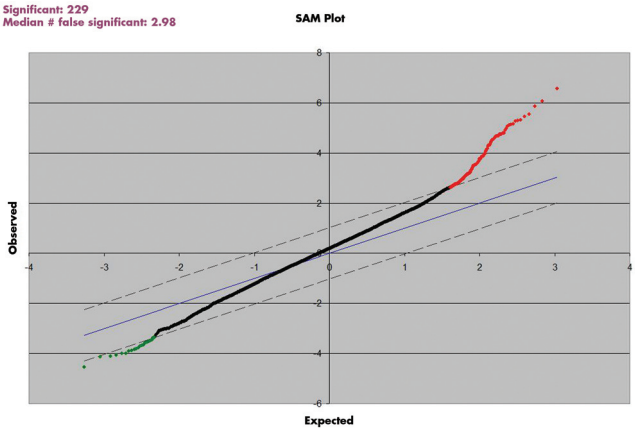

Figure 3. Plot from the SAM analysis for the MM versus MGUS comparison. Each dot represents a particular gene. The *x*-coordinate is the observed value of the test statistic and the *y*-coordinate is the expected value of the test statistics under hundreds of permutations. The dotted boundary is the cutoff for significance given a specified false discovery rate.

What is an Example of Microarray Analysis?

To illustrate some of the typical statistical techniques applied to microarray data, let's examine a real data set from a particular type of commercial chip—Affymetrix (version 5). The samples were taken from three populations: a group with multiple myeloma (a blood cancer abbreviated MM), a group with signs of developing MM (abbreviated MGUS for "monoclonal gammopothy of undetermined significance"), and a healthy group.

In this situation, plasma cells from each of the test subjects were isolated and placed on a microarray chip since multiple myeloma is characterized by plasma cells replicating out of control, which in turn causes organ damage. The Affymetrix chip measures 12,625 genes simultaneously. There were 218 MM samples, 21 MGUS samples, and 45 healthy samples. The data were collected at the Donna D. and Donald M. Lambert Laboratory of Myeloma Genetics, University of Arkansas for Medical Sciences by John Shaughnessy, Jr., and his colleagues (Zhan et al. 2002).

What Statistical Techniques Can We Use?

The tools from basic statistics can be used to address many microarray research questions; however, each research hypothesis requires a different statistical tool.

Comparing Two Groups

Probably the most common research question associated with microarray data is the two group comparison: What differences in genetic activity are there between one group of samples and another group of samples? Usually the first group of samples comes from people with a particular disease and the second group comes from healthy people. We'd like to know what type of genetic activity differentiates the two groups.

For example, we might be interested in comparing the MM group with the MGUS group. The *t*-test from basic statistics can be used to test whether the means of the two populations are the same. By applying the *t*-test separately to each of the 12,625 genes on the chip, we can tell which genes have an average gene expression that is different between the two groups. In the multiple myeloma example the t-test found 422 gene comparisons with *p*-values less than 0.001. Such a *p*-value indicates that the probability is less than one in a thousand that the difference occurred simply by chance.

When the *t*-test is not appropriate (when the data are not normally distributed, for example), we could use the Wilcoxon rank sum method to test whether the median expression levels are the same in the two populations. In the multiple myeloma example the Wilcoxon rank sum test found 341 gene comparisons with *p*-values less than 0.001. In comparing the MGUS and MM samples, the intersection of the genes with *p*-values less than 0.001 for both the *t*-test and the Wilcoxon rank sum test was a set of 269 genes. One analysis approach is to investigate further only those genes that are significant using both types of comparisons.

Other methods have been developed to compare two groups in the context of microarray data. Some researchers have used a modification of the *t*-test (Golub et al. 1999). Researchers at Stanford University have developed a software package (SAM—Significant Analysis of Microarrays) to conduct a permutations test to establish cutoffs for the pairwise comparisons (Tusher et al. 2001; Tibshirani et al. 2002).

The SAM technique applied to the MM versus MGUS data identified 229 genes as showing significant activity. The false discovery rate (FDR) was only about three genes based on hundreds of random permutations of the gene values. Figure 3 (see page 6) shows an output plot from the SAM analysis.

For comparison, Table 2 presents the results of the *t*-test, the Wilcoxon test, and the SAM method.

	Number of Significant Genes ($p < 0.001$)			
Comparison	*t*-Test	Wilcoxon Rank Sum Test	SAM (FDR=3/229)	By Chance
MM versus MGUS	422	341	229	12.63

Table 2. The number of genes judged by each method to have a statistically significant expression level in comparing the MM and MGUS groups. Note that with $\alpha = .001$ there could have been 12.63 genes identified as significant simply by chance even if there was no effect due to the disease.

Comparing Multiple Groups

Since the data actually contain three groups (MM, MGUS, and healthy groups), we could use analysis of variance (ANOVA) to find genes that have an average expression level that is different in at least one group. Just as we used a nonparametric version of the *t*-test (Wilcoxon rank sum test) in the two-group comparison, we could also use a nonparametric version of ANOVA (Kruskal-Wallis test) to analyze non-normal data from more than two groups. Each of these tests produces a *p*-value for the difference across the groups for every single gene. Significant differences can then be identified based on the magnitude of each *p*-value.

Classification

Sometimes the research question has to do with predicting class membership in a set of data for which the classes are already known, that is, classifying a new sample into a known class. In this situation, we could use past data to set up a logistic regression model that can classify a future sample data point. One way to test the accuracy of the model is to classify a subset of points with known class membership that was not used in building the model. These independent data values will give unbiased information about the accuracy of

the classification procedure. When applied to an entire dataset, this procedure is called cross validation. The algorithm for the cross validation procedure is:

1. Partition the data into *k* groups of the same size.
2. Remove the first group from the data and build the model on the remaining *k*-1 groups.
3. Test the removed group of data using the above model and record the predicted class membership.
4. Repeat steps 2 and 3 for each of the *k* groups.
5. Compile the false positive and the false negative rates as a measure of model accuracy.

Using logistic regression with expression level as the explanatory variable and the disease groups (MM vs. healthy, MM vs. MGUS, healthy vs. MGUS) as the response variable, we can create models that predict the classification of future observations into dichotomous categories such as sick or healthy. The accuracy of these three separate models was evaluated using cross validation. The results are displayed in Table 3.

	MM	Healthy	MM	MGUS	MGUS	Normal
Percentage Correctly Classified	96.79%	84.44%	93.58%	38.10%	91.11%	71.43%

Table 3. Results from using logistic regression to predict class membership (MM, MGUS, or Healthy). The effectiveness of the model is evaluated using cross validation. Each entry in the table is the percentage of samples correctly classified.

It is apparent from the results that logistic regression using gene expression values can be used to discriminate between the healthy group and the malignant groups, but it is not useful for discriminating between the two malignant groups (MM and MGUS). This lack of discrimination is seen in the comparison of MM versus MGUS where most of the MGUS samples (62%) were incorrectly predicted to be from MM patients.

This and other discrimination methods for microarray analyses have been compared using cross validation prediction error rates (Hardin et al. 2004).

Clustering

Clustering is a process by which data can be grouped without any preconceived knowledge of the groupings or even of the number of groups. While classification models are referred to as "supervised learning," clustering is sometimes referred to as "unsupervised learning." As with most techniques, there are different clustering algorithms, yet many use some type of metric to establish a distance between two samples or two groups of samples. The concept in clustering is that the closer two items are to each other, the more likely they are related and should therefore be grouped together.

Clustering techniques provide a visual representation of patterns in the data. Groupings or clusters can illustrate relationships that may or may not be known by the researcher. For example, a particular clustering result may demonstrate what gene expressions are useful for characterizing genes with known functions. Or, a clustering result may lead to the discovery of groups of genes that have similar expression patterns. Clustering can also be performed on samples instead of genes. When we cluster samples, we look for similar genetic patterns in groups of individuals.

In hierarchical clustering, the first step is to link the two closest samples. Subsequently, that pair is compared to the remaining samples and either another two samples are linked or the first pair (cluster) is linked to a third sample based on which of these choices represents a shorter distance. This process continues until every sample in the data set is linked to another.

Figure 4 (see page 10) shows the sequential linkages of a sample of patients in our Multiple Mylenoma example. Each vertical line represents one sample. The samples from healthy people are labeled "X" and the samples from the MGUS patients are labeled "MGUS." The MM samples are not labeled.

We can see that the MM samples tend to cluster together to the left and the healthy samples tend to cluster to the right, while the MGUS samples are dispersed throughout. This could indicate that maybe some of the MGUS samples will develop into MM while others of them will remain benign.

To illustrate the clustering process, figure 5 (see page 10) is a magnification of the grouping of predominantly healthy patients on the right side of figure 4. For merges of a pair of samples, the value on the *y*-axis represents the Euclidean distance between the two samples. Where two clusters are merged, the value on the y-axis represents the average of the distances between each of the samples in one cluster and each of the samples in the other cluster. Notice that merges shown at the lower portion of the graph are samples that are the closest to each other (most similar), while merges shown at the upper portion of the graph are samples that are the farthest apart (least similar).

Figures 6 (see page 11) shows clusterings of samples from only MM patients. We notice that there still appear to be some groups of samples even though all of the patients have the disease. This might indicate that some of the samples are genetically related in such a way that those patients would respond similarly to treatment.

Figure 7 (see page 11) shows the results of clustering using a set of 50 completely randomized expression values. Because the data are randomly distributed, we should not expect to see any clustering pattern Interestingly, however, we can see some possible group, even though there should be no structure to the data. But when we compare figures 6 and 7, the groupings of random values in figure 7 are less distinct t han the groups of real values in figure 6. We can also see that the distances between random values in figure 7 are much longer than the distances between real values in figure 6.

Consequently, because the clustering algorithm forces some configuration, we must be careful in deducing that there are significant relationships among the samples. A statistician should use these clustering methods carefully, especially when communicating with nonstatisticians, so as not to overinterpret any apparent structure in the data. Interpreting the groups within a hierarchical cluster is somewhat subjective and does not follow a formal structure of decision-making as in hypothesis testing.

Other classification and clustering techniques commonly used on microarray data include "nearest shrunken centroid" classification (Tibshirani et al. 2002) and "model-based clustering" (Yeung et al. 2001).

Advanced Techniques

Advanced techniques are often applied to microarray data and new methods are constantly being developed to better analyze the data. Some examples of advanced techniques we frequently use include:

• *Time Series Analysis* – With time series analysis we can observe trends over time for organisms like yeast, for example, that change rapidly (Zhao et al. 2001).

• *Partial Least Squares and Principal Component Analysis*—Both of these methods allow the analyst to reduce the dimensions of the data in a meaningful way. Since many data sets have hundreds of samples with thousands of dimensions, it is important to reduce the dimensions in a way that captures the signal while discounting the noise (Nguyen and Rocke 2002; Yeung and Ruzzo 2001; Bair and Tibshirani 2004).

• *Discriminant Analysis*—This is a way of partitioning the data and can be used for classification problems (Dudoit et al. 2002).

• *Survival Analysis*—This technique is used to evaluate data with censored endpoints that are common in medical studies. "Censoring" occurs when a patient dies or for some other reason does not complete the study. The Cox proportional hazards model—the standard survival model—is not equipped to handle thousands of explanatory variables and so variable reduction techniques must be used to fit survival analysis models (Pauler et al. 2004; Bair and Tibshirani 2004).

What Statistical Issues are Specific to Microarray Analyses?

Many of the techniques used to analyze microarray data are straightforward applications of well-known methodology and some of the established procedures can be modified to handle large data sets. However, some issues cannot be dealt with using standard statistical approaches and research is needed into new techniques to address specific problems.

One problem with microarray data is that the number of genes is almost always bigger than the sample size. This type of sparse data makes inverting covariance matrices impossible, which in turn forces us to pare down the number of variables for methods like regression analysis that use inverted covariance matrices to calculate least squares estimates. Some data reduction techniques have been developed, but there is more work to be done to develop new methods for ascertaining what set of variables would be the most informative.

Because we often are interested in understanding particular genes, we use gene-by-gene techniques like *t*-tests, ANOVA, or regression analysis. Each time a gene is judged to be significant according to one of these tests, there is the risk of producing a Type I error. If we were to set our significance level to $\alpha = 0.05$ and run t-tests on 10,000 genes, we would expect 500 genes to test as significant, even if there is no signal in the data. The problem of controlling for this type of error in general has been studied widely (Benjamini and Hochberg 1995) and is now being researched in the specific context of microarray data (Storey, 2002).

Another problem is that microarray data do not conform to the usual assumptions of many standard statistical tests. The data themselves are in units of fluorescence and are often highly skewed right and can even be negative if a "background adjustment" is needed when the background fluorescence is brighter than the foreground fluorescence. Often log transformations (with some ad hoc adjustment for the negative values) give data that are moderately symmetric. However, log-transformed microarray data may still have highly unequal variances for which many techniques (like ANOVA) are not robust. Transformations and normalizations for microarray data are being researched so that the results from standard statistical analyses, based on the usual requirements, are reliable (Durbin et al. 2002).

What Software is Available for Microarray Analyses?

New software is constantly being developed to perform analyses specifically for microarray data. Because the technology is relatively new, much of the software is being developed in academia and is freely available. Below are summaries of a few of the most commonly used software programs. The synopses are based on my experience and not meant as endorsements or condemnations of any of the software.

• **Bioconductor:** This is a free program that runs in R. It is designed for statisticians who are researching new techniques on microarray data. It is flexible, though it does require basic programming knowledge of R or S-Plus. Bioconductor also has multiple graphs and features designed specifically for extracting information from microarray data.

www.bioconductor.org

• **SAM & PAM:** Significance Analysis of Microarrays (SAM) and Prediction Analysis of Microarrays (PAM) are free software programs that add-in to Microsoft Excel or R. SAM produces pictures and lists of genes that are

Continued on page 12

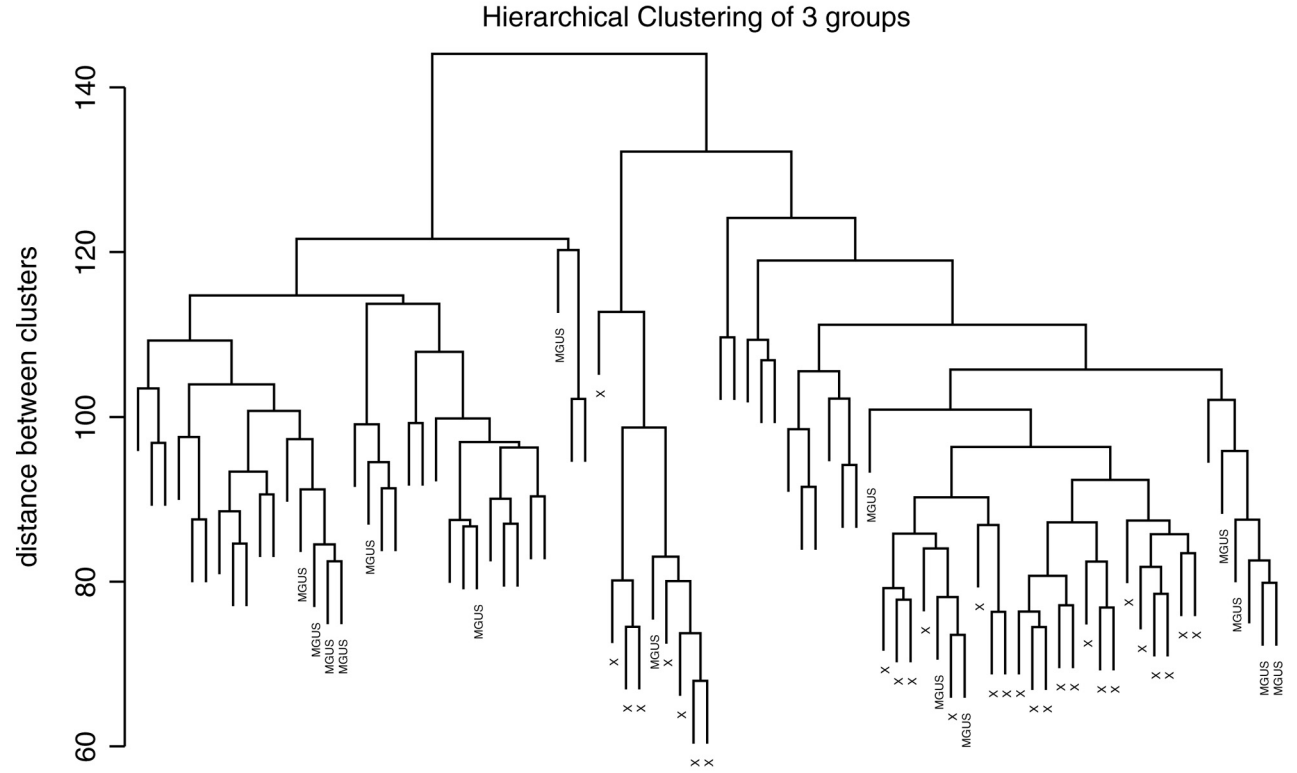

Figure 4. Hierarchical clustering of 85 randomly selected MM, MGUS, and Healthy Samples.

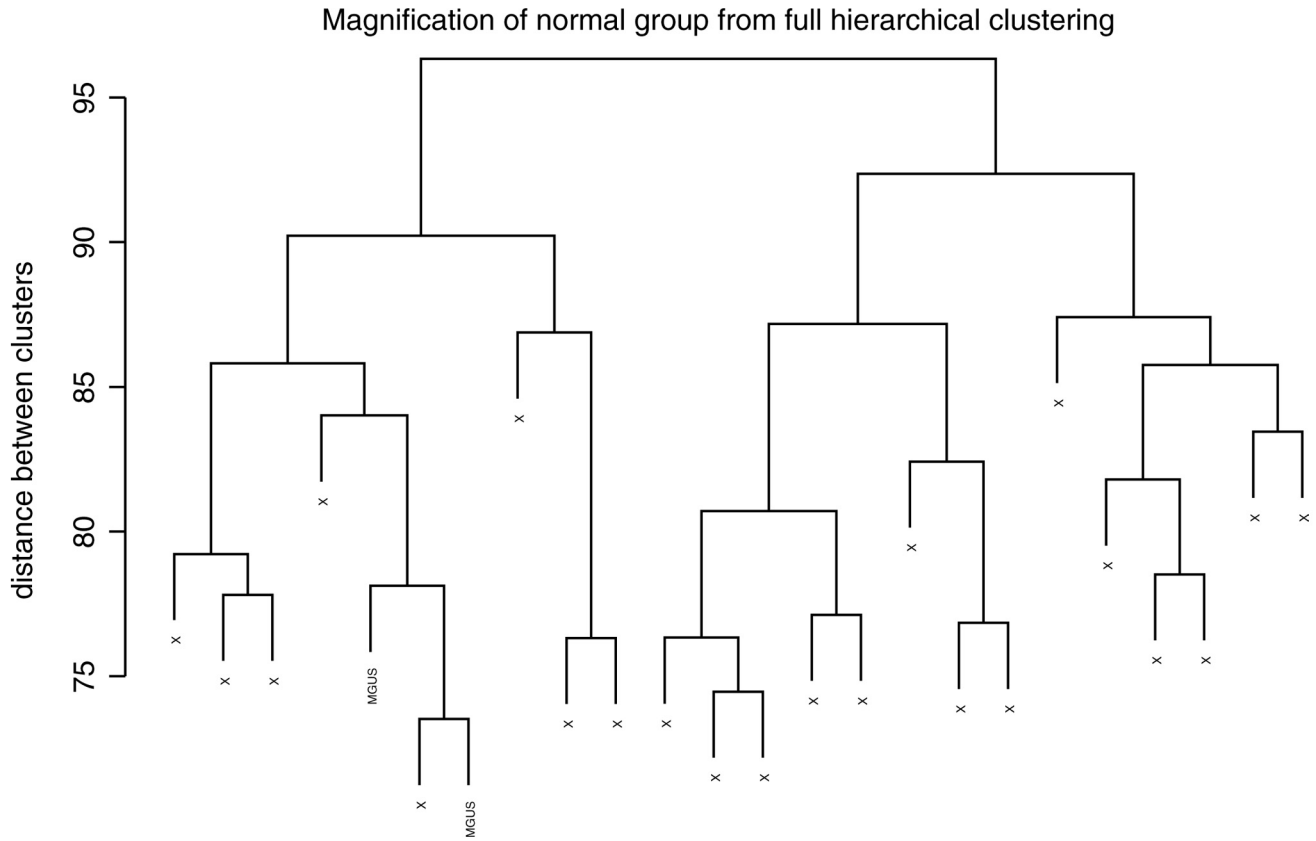

Figure 5. Hierarchical clustering analysis of 24 samples from healthy patients.

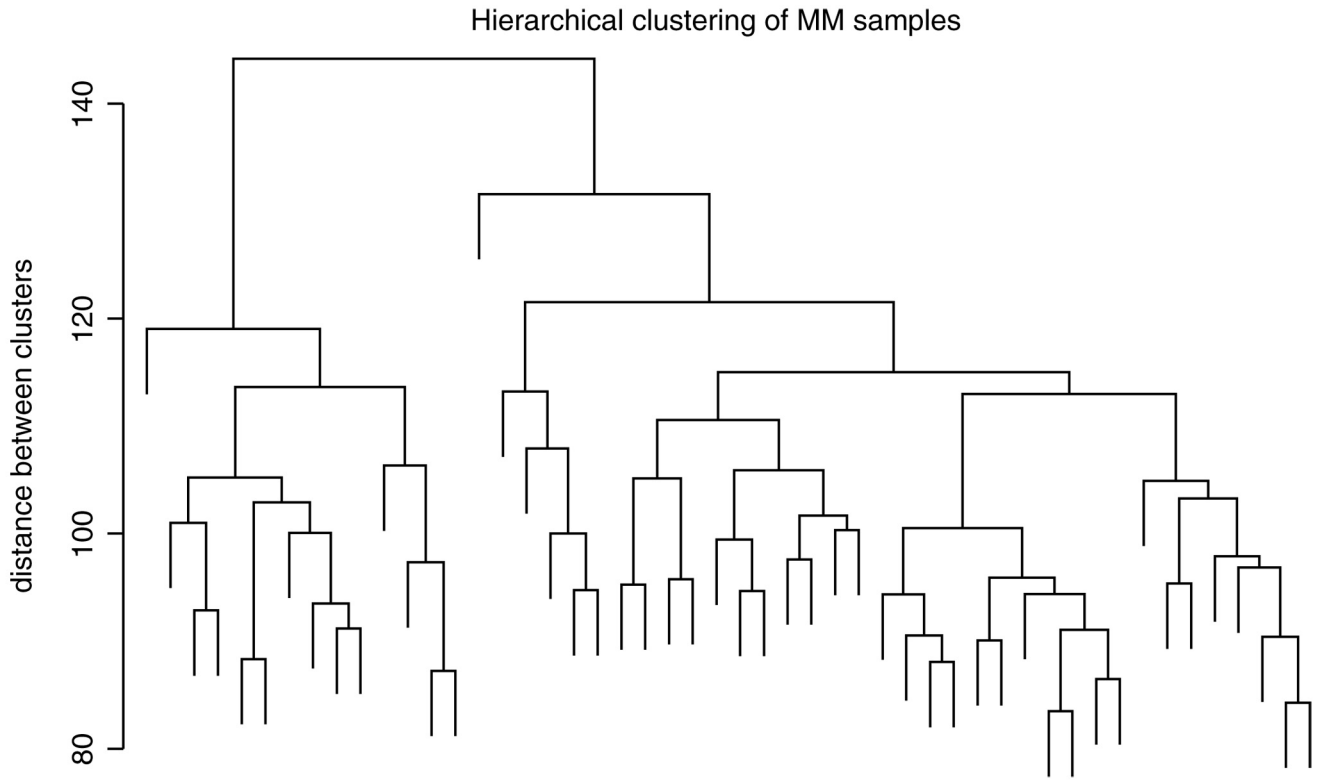

Figure 6. Hierarchical clustering analysis of 50 samples from MM patients.

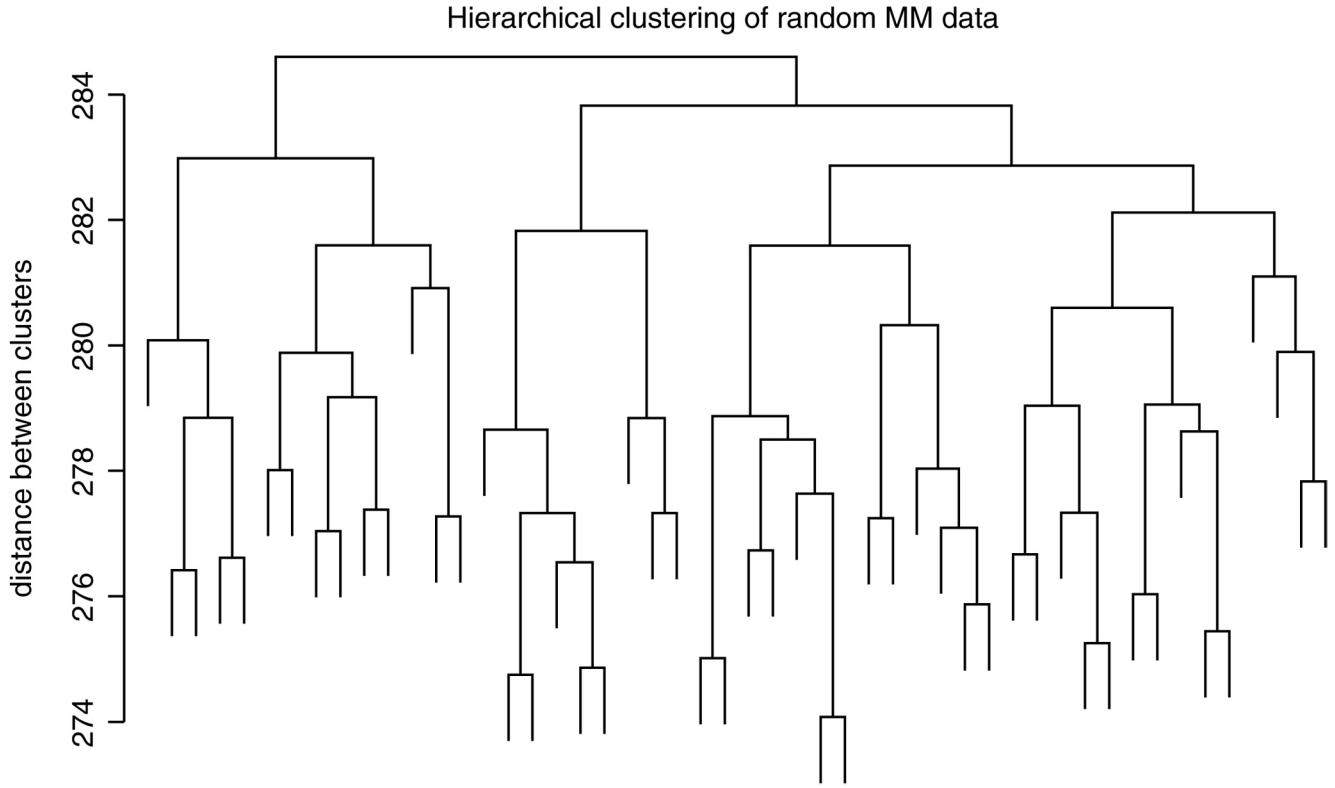

Figure 7. Hierarchical clustering analysis of 50 random expression values.

significant across groups, while controlling for the false discovery rate. It also correlates gene expression to clinical parameters. PAM performs classification of microarray data using nearest shrunken centroid methods.

www-stat.stanford.edu/~tibs

• **Cluster & Treeview:** These programs perform hierarchical clustering across both samples and genes. The results are displayed in tree-based images with label information and colors representing expression levels. Cluster and Treeview are both freely available software programs.

http://rana.lbl.gov/EisenSoftware.htm

• **BRB ArrayTools:** This software is designed as a free add-in to Microsoft Excel for visualization and statistical analysis of microarray data. It contains various methods including class comparison, class prediction, and permutation tests for significance levels.

http://linus.nci.nih.gov/BRB-ArrayTools.html

• **GeneSpring:** This package is widely used by biologists and geneticists. It is user friendly and has many good statistical techniques, including adjustments for multiple comparisons. However, it is not free and not as flexible for statistical research as other programs.

www.sigenetics.com/GeneSpring/GeneSpring.html

Summary

For biologists, microarray technology has opened new avenues to access a new world of knowledge quickly and inexpensively. Never before has it been possible to study so many genes simultaneously on so many samples. However, any technology is limited by its ability to extract information.

As statisticians, it is our role to ensure that the information obtained from microarray experiments is valid and interpreted appropriately. Many of the statistical concepts from the last century are applicable to microarray analysis, but we must also open our minds to new techniques and methodologies that will be better suited for this new generation of data. In this century, our contribution to science will be to develop the analyticaltools that can handle future generations of data yet to come.

References

Bair, E., and Tibshirani, R. 2004. Semi-supervised methods to predict patient survival from gene expression data, *PLOS Biology*, 2:511-522.

Benjamini, Y., and Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society*, B, 57:289–300.

Dudoit, S., Fridlyand, J., and Speed, T. 2002. Comparison of discrimination methods for the classification of tumors using gene expression data, *Journal of the American Statistical Association*, 97:77–87.

Durbin, B., Hardin, J., Hawkins, D., and Rocke, D. 2002. A variance-stabilizing transformation for gene-expression microarray data, *Bioinformatics*, 18:105S–110S.

Eisen, M., and Brown, P. 1999. DNA arrays for analysis of gene expression, *Methods in Enzymology*, 303:179-205.

Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C., and Lander, E. 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science*, 286:531–537.

Hardin, J., Waddell, M., Page, C., Zhan, F., Barlogie, B., Shaughnessy, Jr., J., and Crowley, J. 2004. Evaluation of multiple models to distinguish closely related forms of disease using DNA microarray data: an application to multiple myeloma, *Statistical Applications in Genetics and Molecular Biology*, 3, article 10.

Lockhart, D., Dong, H., Bryne, M., Follettie, M., Gallo, M., Chee, M., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., and Brown, E. 1996. Expression monitoring by hybridization to high-density oligonucleotide arrays, *Nature Biotechnology*, 14:1675-1680.

Nguyen, D., and Rocke, D. 2002. Tumor classification by partial least squares using microarray gene expression data, *Bioinformatics*, 18:39–50.

Nugent, K., Spigelman, A., and Phillips, R. 1993. Life expectancy after colectomy and ileorectal anastomosis for familial adenomatous polyposis, *Diseases of the Colon and Rectum*, 36:1059–1062.

Pauler, D., Hardin, J., Faulkner, J., LeBlanc, M., and Crowley, J. 2004. Survival analysis with gene expression, Balakrishnan, N., and Rao, C., editors, *Handbook of Statistics 23: Advances in Survival Analysis*. Elsevier.

Schena, M., Shalon, D., Davis, R., and Brown, P. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray, *Science*, 270:467-470.

Storey, J. 2002. A direct approach to false discovery rates, *Journal of the Royal Statistical Society*, B, 64:479–498.

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. 2002. Diagnosis of multiple cancer types by shrunken centroids of gene expression, *PNAS*, 99:6567–6572.

Tusher, V., Tibshirani, R., and Chu, G. 2001. Significance analysis of microarrays applied to the ionizing radiation response, *PNAS*, 98:5116–5121.

Velculescu, V., Zhang, L., Vogelstein, B., and Kinzler, K. 1995. Serial analysis of gene expression, *Science*, 270:484-487.

Yeung, K., Fraley, C., Murua, A., Raftery, A., and Ruzzo, W. 2001. Model-based clustering and data transformations for gene expression data, *Bioinformatics*, 17:977–987.

Yeung, K., and Ruzzo, W. 2001. Principal component analysis for clustering gene expression data, *Bioinformatics*, 17:763–774.

Zhan, F., Hardin, J., Kordsmeier, B., Bumm, K., Zheng, M., Tian, E., Sanderson, R., Yang, Y., Wilson, C., Zangari, M., Anaissie, E., Morris, C., Muwalla, F., vanRhee, F., Fassas, A., Crowley, J., Tricot, G., Barlogie, B., and Shaughnessy, Jr., J. 2002. Global gene expression profiling of multiple myeloma, monoclonal gammopathy of undetermined significance and normal bone marrow plasma cells, *Blood*, pages 1745–1757.

Zhao, L., Prentice, R., and Breeden, L. 2001. Statistical modeling of large microarray data sets to identify stimulus-response profiles, *PNAS*, 98:5631–5636. ■