

DNA Microarrays and Bayesian Multiple-Hypothesis Testing

*The University
of Texas
at Austin*

*Undergraduate
Research
Journal*

*Volume III
Spring 2004*

James Scott, *College of Natural Sciences*

Introduction

DNA microarrays are some of the most powerful tools in modern biology. Yet because of that power, they are also some of the most fickle. To put it simply, they give biologists an enormous statistical headache.

What is a DNA microarray? The best way to answer that question is by analogy with a small chessboard, where each square represents a single gene. Except, instead of eight squares by eight like a chessboard, microarrays might have 100 squares by 100, for a total of 10,000 genes—all packed onto a chip of about three square inches.

Scientists use microarrays to monitor the effect of some stimulus upon many of these genes simultaneously. For example, you might want to see how alcohol affects the genetic expression

profile of a rat's brain—in other words, which genes in a brain cell are differentially expressed because of alcohol exposure. So you take two different measurements of the genetic expression profile of rat brain cells (one for sober rats, one for drunk rats). Then you view the results by hybridizing the relevant strands of RNA to bits of the rat genome, which has been partially encoded on microarrays. A few of the genes will have their expression profiles altered with respect to the baseline expression level. In other words, they will be made more (or less) active because of the alcohol. Identify those genes, and you're that much closer to understanding alcohol addiction.

These applications sound great in theory. The problem is that microarray data are very hard to analyze. One factor is the sheer number of genes being tested, which means that random error creeps into the experiment with astonishing frequency. Another problem is the complexity of the underlying statistical model, often called a mixture model. And yet a third problem is the issue of economics. How bad is it if I dismiss a truly active gene as nothing but random noise? And how much will it cost me if I keep doing experiments on a false positive thinking it was real? These costs can be quantified, but traditional statistical methods lack a straightforward way for taking them into account.

Over the summer of 2002, I developed a statistical procedure for dealing with these problems as best as can be expected. It's called Bayesian hierarchical modeling, and I will show you how it performs on a simulated data set. But my goal in this paper isn't to explain the nuts and bolts of that procedure. Instead, I want to explain in more detail why analyzing data from DNA microarrays is so difficult. These lessons, more than any individual procedure, are what I took away from my research experience.

To do that, I need to explain three concepts: random variation, multiple-hypothesis testing, and mixture models. You should be aware that what follows is something of an idealization of how DNA microarray data are collected in the real world. After all, we have to do that to create a statistical model—and this paper is about statistics, not biology. If

you're interested in the process by which scientists transform the raw numbers from microarray experiments into a form usable by statisticians, you should explore Baldi et al.

Random Variation

Almost all physical systems are subject to random variation.

Think, for example, of an experiment that involves flipping a fair coin 10 times. Since the chance of getting heads on each flip is 50%, you'd probably expect to get 5 heads, and this is certainly true on average. But if you performed the 10-coin-flip experiment many times in a row, you'd only get 5 heads about 25% of the time. The other 75% of the time you'd get something else—plenty of 4's and 6's, some 3's and 7's, a smattering of 2's and 8's, and only rarely a 1 or a 9. And if you had nothing to do all day but flip coins, you might even get a run of 10 heads or 10 tails. This is the essence of random variation: everything turns out how you'd expect in the long run, but each trial might be a bit off from the average.

In their own way, genes are no different than coins.

Let's forget about microarrays for moment and talk about just a single gene. That gene's expression profile has a normal numerical range associated with it. In repeated experiments, the gene will stay within this range most of the time, just like a series of 10 coin flips will usually yield between 4 and 6 heads. But every once in awhile, the expression can creep outside that range just by random chance alone—just like you might get 9 out of 10 heads from time to time, even if the coin were fair. This will fool you into thinking that the stimulus you're studying actually caused a change in the gene's expression profile, when the real culprit is just random chance.

Let's say we want to find out whether a single gene is active. If it's inactive, its average, or mean, expression level will be 0 (units don't matter here). Although it won't be *exactly* 0 each time we observe it, it will usually be pretty close—say, for example, between -2.0 and 2.0 about 95% of the time.

So what happens if we observe an expression level for that gene of 2.1? There are two possible explanations.

Hypothesis A says, “Don’t get excited. The mean expression level is still 0; the extreme result was just a random occurrence.” Hypothesis B, on the other hand, says, “The mean expression level isn’t 0 after all; the extreme result was due to a real effect on the gene.”

Deciding between these two claims is the whole point of a branch of statistics known appropriately as hypothesis testing. Different schools of thought approach hypothesis testing in different ways, but I’m specifically referring to something called Bayesian hypothesis testing. Its goal is to quantify random variation so that, given the data, we know exactly how reasonable each hypothesis can claim to be.

Multiple-Hypothesis Testing

Now that you understand simple hypothesis testing, I should explain multiple-hypothesis testing.

Let’s go back to the coin-flipping example. You know from experience that flipping a coin 10 times and getting 10 heads or 10 tails is very rare. In fact, with a truly fair coin, these each happen about 0.1% of the time, meaning that you’d have to perform 355 trials of the 10-flip experiment to get even a 50% chance at seeing all 10 heads or all 10 tails. Of course, if you did 1,000 trials, your chances of seeing 10 heads or 10 tails shoot to 85%.

These numbers capture an intuition we probably all share. Even very rare events—the lightning strike, the hole-in-one, the royal flush, the 10 heads in a row—will happen eventually if you just wait long enough. It’s like the old proverb: even if you’re one in a million, there are 1,000 people just like you in China.

By now the connection to microarrays should be apparent. A microarray is like stuffing 10,000 single-gene experiments onto a 3-square-inch wafer. We’d fully expect to see some expression levels very far from 0, just by random chance alone, even if—and this is the crucial point—all the genes were inactive.

Let’s plug in some numbers. Imagine we’re doing a microarray test in which each inactive gene has a 5% chance of randomly creeping outside its normal range between -2.0 and 2.0 . If our chip has 10,000 genes on it, then we’d expect around 500 such “false positives.” Now imagine the actual experiment turns up 550 genes

that we observe to be more than 2 units away from 0. Some of them will be false positives—that is, caused by random chance. And some of them will be real alterations of the baseline expression level.

Distinguishing between these two cases is the dilemma of multiple-hypothesis testing. It is the main reason that analyzing DNA microarray data is so hard.

Mixture Models

The problem gets worse. What if a gene is differentially expressed, but only slightly? That gene could be scientifically important, yet you might never pick it out from the sea of random noise.

Again, some numbers will help make the point. An active gene whose mean expression level is, for example, 0.5 looks a lot like an inactive gene, except that its normal numerical range will be offset just a tad. Using the same scale as before, you’d observe it between -1.5 and 2.5 about 95% of the time. Let’s say that on the particular day you ran your microarray experiment, you observed a value for that gene of 0.7. How can you tell whether that number is a small deviation from 0.0, or a slightly smaller deviation from 0.5?

A situation like this is called a mixture model. Most of the genes on a microarray will have mean expression levels of 0, and those that aren’t will have their mean expression levels distributed roughly according to a bell curve that is itself centered around 0. And all of the genes, whether expressed or not, are subject to random variation (which also happens to look like a bell curve). Hence the term “mixture model”: each individual observation can be conceptualized as a mixture of two bell curves. There’s the *population* bell curve, which describes the distribution of differentially expressed, non-zero genes. And there’s also the *observational* bell curve, which describes the distribution of measurements you will get in a series of repeated measurements for a single gene.

The problem of hypothesis testing then becomes the problem of deciding which of these bell curves contributed to each particular data point. Often, however, these bell curves aren’t very different from one another (for the technically inclined, the ratio of their variances isn’t that far from 1). This makes the problem

even harder. Often, the only solution is to run several microarrays, which has the effect of shrinking the observational bell curve so that it becomes sufficiently different from the population bell curve.

The Solution

Bayesian Hierarchical Modeling

Without going into the specifics of Bayesian hierarchical modeling, I want to show you the results this powerful method can give. If you're interested in the gritty details, see Scott and Berger.

Figure 1 shows how the procedure handles the problem of multiple-hypothesis testing for a fake-but-reasonable data set. The column entries across the top are the numerical values of 10 different "signal" observations. These were generated from a bell curve with a mean of 0 and a standard deviation (a measure of how spread-out the distribution is) of 3. I then added different amounts of "noise" genes to the data set, represented by the row entries along the left. These numbers were drawn from a bell curve with mean 0 and standard deviation 1. They were meant to simulate genes whose mean expression level is actually 0 and that conform to the exact assumption I made above—that successive observations of these genes will yield values between -2.0 and 2.0 about 95% of the time.

I then asked the question, "Given all the data, what is the probability that each individual gene is a signal?" The entries in the chart give my procedure's answer to this question for each of the 10 signal genes.

You'll notice a couple of things about the chart. First, even with only 25 noise genes in the mix, the procedure has a hard time identifying the signal genes that are very close to 0 in absolute value. For example, it says that, given the data, there's only a 20% chance that the -0.15 gene is a signal.

In a strict sense, this is a "wrong" answer—we know that the gene is a signal because we created the data set that way, and the procedure tells us that it's probably noise. Yet the figure of 20% says that this wrong answer is an exception. That number means that if you counted up all of the situations with 10 signal genes and 25 noise genes distributed according to these particular

rules, only one in five observations of -0.15 would correspond to a signal gene. The other four out of five such observations would be noise genes. So, while the procedure gets the answer wrong in this instance, it's playing the odds. Classifying this observation as a noise gene will be the right thing to do 80% of the time, and that's what the procedure does.

Second, you'll notice that as the number of noise genes increases, the probability that the marginal observations are signal genes goes way down. This is exactly what a good multiple-hypothesis testing procedure ought to do. As the total number of observations (both signal and noise) goes up, so does the number of opportunities for a rare event to happen by random chance alone. So, with more observations, the threshold for declaring something a signal ought to get harder and harder to meet—that is, further and further from 0.

The price you pay, of course, is ignoring some signal genes that are themselves close to 0. But you really don't have much choice in the matter. If you don't adjust the acceptance threshold for the total number of observations, your experiment will rapidly be overwhelmed by false positives. You'll be stuck in the lab forever trying to sift through them.

Now look at Figure 2 for a closer look at four individual observations. Here, we don't know beforehand which ones are signals and which are noise genes—we only know what the procedure tells us. The number at the top of each frame is the actual observation. The bar at 0 represents the probability, given all the data, that the observation is just noise. And the curve represents the distribution of likely values of the gene's mean expression level, given that the observation is actually a signal. At a glance, you can determine two things from one of these figures. First: how likely is it that the gene is just random noise? Second: if it's not random noise, is it likely to be far enough from 0 to be worth the time it would take to pursue it further?

Conclusion

I hope you now understand why analyzing data from DNA microarrays is hard. The combination of multiple-hypothesis testing and mixture models is a tough nut to

crack. Even an optimal procedure like Bayesian hierarchical modeling often gets the answers wrong for genes that live close to 0.

The lesson here is one we'd all do well to learn. Sometimes the answer to a scientific problem isn't a fancy statistical procedure. It's to get back into the lab to take more data.

References

- Baldi, Pierre and Wesley Hatfield. *DNA Microarrays and Gene Expression: From Experiments to Data Analysis and Modeling*. New York: Cambridge University Press, 2002.
- Berger, James O. *Statistical Decision Theory and Bayesian Analysis*. New York: Springer, 1980.
- Robert, Christian P. *The Bayesian Choice: A Decision-Theoretic Motivation*. New York: Springer, 1994.
- Scott, James G. and James O. Berger. *An Exploration of Aspects of Bayesian Multiple Testing*. Duke University. 27 June 2003. <www.stat.duke.edu/~berger/papers/multcomp.html>.